Journal of Risk and Uncertainty, 19:1–3; 203–235 (1999)

Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues

DANIEL KAHNEMAN
Woodrow Wilson School of Public Policy, Princeton University, Princeton, NJ 08544

ILANA RITOV
Hebrew University, Jerusalem

DAVID SCHKADE
University of Texas, Austin

Abstract

Participants in contingent valuation surveys and jurors setting punitive damages in civil trials provide answers denominated in dollars. These answers are better understood as expressions of attitudes than as indications of economic preferences. Well-established characteristics of attitudes and of the core process of affective valuation explain several robust features of dollar responses: high correlations with other measures of attractiveness or aversiveness, insensitivity to scope, preference reversals, and the high variability of dollar responses relative to other measures of the same attitude.

Key words: preferences, attitudes, contingent valuation, psychology and economics, utility assessment

JEL Classification: D00, H00

Introduction

Economics and psychology offer contrasting perspectives on the question of how people value things. The economic model of choice is concerned with a rational agent whose preferences obey a tight web of logical rules, formalized in consumer theory and in models of decision making under risk. The tradition of psychology, in contrast, is not congenial to the idea that a logic of rational choice can serve double duty as a model of actual decision behavior. Much behavioral research has been devoted to illustrations of choices that violate the logic of the economic model. The implied claim is that people do not have preferences, in the sense in which that term is used in economic theory (Fischhoff, 1991; Slovic, 1995; Payne, Bettman and Johnson, 1992). It is therefore fair to ask: if people do not have economic preferences, what do they have instead? Does psychology provide theoretical notions that can account, at least in some contexts, both for apparent violations of the rational model of preference and for the regularities of observed

choices? Behavioral research has documented several psychological processes that provide partial answers to this question, including concepts such as mental accounting, loss aversion and hyperbolic discounting. To this set of conceptual tools the present treatment adds the concept of *attitude*, which we borrow from social psychology, and the core process—we label it *affective valuation*—which determines the sign and the intensity of the emotional response to objects.

The main topic that we discuss in this paper—the valuation of environmental public goods—is far from the core of economic discourse. It is an unusual case in which some economists have proposed to use responses to hypothetical questions as a measure of economic preference. In the contingent valuation method (CVM), survey respondents are asked to indicate a stated willingness to pay (SWTP) for public goods, including goods from which they derive no personal benefit, such as the continued existence of obscure species and the maintenance of pristine lakes in inaccessible areas. The proponents of CVM have argued that properly elicited statements of WTP reveal genuine economic preferences, to which consumer theory applies (Mitchell and Carson, 1989; Hoehn and Randall, 1987; Smith, 1992).

We develop here an argument made earlier (Kahneman and Ritov, 1994) that statements of WTP are better viewed as expressions of attitudes than as indications of economic preferences. The conflicting views of the nature of SWTP lead to different interpretations of apparently anomalous features of CVM results, such as the low sensitivity to variations of scope and the discrepancy between the estimates of SWTP derived from open-ended and from referendum questions. The supporters of CVM have sometimes dismissed these anomalies as artifacts of poor technique (Carson and Mitchell, 1993; Smith, 1992), or explained them in terms of standard economic concepts, such as incentive compatibility and substitution and income effects (Hanemann, 1994; Randall and Hoehn, 1996; Smith, 1992). In contrast, the thesis of the present paper is that the anomalies of CV are inevitable manifestations of known characteristics of attitudes and attitude expressions.

To demonstrate the generality of the analysis of SWTP in terms of attitudes, we draw on an experimental study of the setting of punitive damages in product liability cases (Kahneman, Schkade and Sunstein, 1998). The tasks faced by a respondent to a CV survey and by a juror have little in common in the context of an economic analysis; consumer theory may apply to the former but surely not to the latter. In the framework that we propose, however, the two tasks are very similar. Both require the individual to express an attitude—to an environmental problem or to a defendant's actions—by using a dollar scale. The striking parallels between the findings in the two situations strongly support the attitude model.

The evidence that we present is drawn exclusively from studies of verbal answers to hypothetical questions about public issues. It is perhaps not surprising that, on this favorable terrain, the concepts of attitude and affective valuation provide a useful account of the data. It is early to say whether these concepts will prove equally useful in other domains to which the theory of economic preference is usually applied. On current evidence, it is possible to accept an attitude model for hypothetical CV responses while retaining the idea that the standard model of

rational choice applies to more consequential decisions. This appears to be the position of economists who have criticized CVM (e.g., Diamond and Hausman, 1994). We believe, however, that the idea that actions are often interpretable as relatively direct expressions of an affective valuation is likely to prove useful in the analysis of many economically significant behaviors.

The paper is organized in two parts. The first part, which includes sections 1–4, introduces the concepts of attitude and affective valuation and explores some contrasts between attitudes and economic preferences, with examples from studies of contingent valuation and of punitive damages. Sections 5 and 6 apply a psychophysical analysis of dollar responses to explain both the unpredictability of jury awards and some important results of CV research. Section 7 discusses implications and Section 8 concludes.

This article covers much ground and asserts many claims with relatively little documentation. To facilitate a separate assessment of the claims and of their associated evidence, we present our argument in the form of a series of propositions, with brief discussion of each proposition in turn.

1. Introduction to valuation

1-1) *The concept of* ***attitude*** *has been defined as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor"* (*Eagly and Chaiken*, 1996). *The core of an attitude is a* ***valuation***, *which assigns to the entity an* ***affective value*** *that can range from extremely positive to extremely negative*.[1]

1-2) *Affective values vary in sign* (*positive or negative*) *and in intensity*. *The intensity of valuation is relative*: *an attitude object considered on its own is implicitly compared to a set of objects of the same general kind*. (see section 4).

1-3) *The concept of attitude has a considerably broader range of application than the standard concept of economic preferences*. In contrast to economic preferences, which are about commodity bundles (Varian, 1984), objects of attitudes include anything that people can like or dislike, wish to protect or to harm, want to acquire or to reject. People have attitudes toward abstract concepts, individual persons and social groups, events in their personal past and historical figures. Expressions of attitude are also diverse: they include smiles and frowns, verbal statements of approval or abuse, physical assault, charitable contributions, answers to survey questions, and many others. The valuation component of attitudes is assumed to be automatic and to facilitate a broad range of responses that express positive or negative affect (Fazio, Sanbonmatsu, Powell, and Kardes, 1986; Pratto, 1994; Tesser and Martin, 1996).

1-4) *People's attitudes to objects and to activities that affect these objects are usually consistent*. For example, a positive affective response to dolphins is likely to be associated with a positive valuation of actions that protect members of this species.

The link between attitudes and actions is often far from perfect, however (Eagly and Chaiken, 1993).

1-5) *The objects of attitudes and valuations are mental representations, not objective states of affairs. Valuations are therefore subject to framing effects and violate the logic of extensionality.* In an example much discussed by philosophy students, an individual may have different attitudes to the evening star and to the morning star, although they are the same star. People can also have different attitudes to the same packaged meat depending on whether it is described as containing 5% fat or as being 95% fat-free. The latter example is a *framing effect*, in which two descriptions evoke different valuations although they are transparently co-extensional—they refer to the same state of the world. Many large and robust framing effects have been identified by students of individual decision making (e.g., Tversky and Kahneman, 1986) and students of political attitudes (Bartels, 1998; Quattrone and Tversky, 1984; Zaller, 1992). Framing effects violate a condition of extensionality (Arrow, 1982) or invariance (Tversky and Kahneman, 1986), which is commonly taken for granted in economic analyses of preference. The psychological analysis of attitudes and valuations explicitly rejects the extensionality assumption.

1-6) *The following is a partial list of the properties of attitudes and of the ways they differ from preferences.* (*i*) *Attitudes are defined by the affective value of objects considered one at a time, not by choices.* (See sections 2 and 4.) (*ii*) *Attitudes violate extensionality. The same object may evoke different valuations depending on its description and on the context in which it is evaluated.* (*iii*) *The separate attitudes to two objects do not necessarily predict the outcome of a choice or direct comparison between them: reversals can occur when the comparison alters the relative salience of some attributes* (*Hsee*, 1996), *or when the objects belong to different categories.* (See section 4.) (*iv*) *The attitude to a set of similar objects is often determined by the affective valuation of a prototypical member of that set. The size of the set is neglected in this mode of valuation, which violates the logic of preferences.* (See section 3.) (*v*) *Alternative measures of attitudes differ in their precision, statistical efficiency and susceptibility to biasing influences. Dollar measures are inferior on all three counts.* (See sections 5 and 6.)

2. The evaluation factor

A central claim of the present treatment is that diverse responses to an object often express the same affective valuation. Consequently, the answers to ostensibly different questions are expected to yield similar rankings of attitude objects. The present section provides some evidence for this hypothesis. The data that we consider for each object are *averages* of attitude measures obtained from different samples of respondents. The correlations that we discuss in this section answer the following question: do different ways of probing average attitudes to a set of objects yield similar attitude orders?

2-1) *The affective value of an object is the major determinant of many responses to it, which are called* ***attitude expressions****. A correlational analysis of the responses to a set of objects normally yields a strong* ***evaluation factor****, which captures the commonality among diverse expressions of the same attitude*. The classic set of studies that introduced the semantic differential technique (Osgood, Suci and Tannenbaum, 1957) still provides the best illustration of this proposition. Participants in SD studies are presented with a series of objects or concepts. Their task is to rate each object in turn on a set of seven-point scales defined by bipolar adjectives, such as GOOD-BAD, KIND-CRUEL, BEAUTIFUL-UGLY, LARGE-SMALL, STRONG-WEAK, MASCULINE-FEMININE, IMPORTANT-UNIMPORTANT, and others. The range of objects to which this technique can be applied is hardly constrained: it includes particular objects, events, abstract ideas, activities, and nonsense figures. The participants are instructed to work quickly and to rate each object on every scale, regardless of whether or not it applies literally. Thus, 'wisdom' and 'Paris' could both be rated on the scales LARGE-SMALL and HOT-COLD—most people will rate wisdom as larger and colder than Paris.

For our purposes here, the most important conclusion of studies of the semantic differential is that the factorial structure of SD data is surprisingly simple. The same structure has been confirmed in many studies. The largest factor to emerge is invariably an ***evaluation factor***, so labeled because the highest loadings are on scales such as GOOD-BAD, KIND-CRUEL and BEAUTIFUL-UGLY. The evaluation factor typically accounts for about 50% of the variance in scale responses. The scales that define the evaluation factor are not perfectly correlated, of course, and the differences among them are meaningful. For example, 'justice' is likely to be rated higher on the GOOD-BAD scale than on the KIND-CRUEL scale. Large discrepancies are rare, however, and the different evaluation scales generally yield similar orderings of the objects of judgment.

2-2) *Attitudes can be expressed on a scale of dollars, as well as on rating scales. Valuations expressed in dollars are highly correlated with those expressed on rating scales*. Willingness to pay for environmental goods—e.g., the maintenance of species—is one possible expression of attitudes to these goods, and to interventions that affect them. Similarly, attitudes to defendants in civil trials can be expressed by an amount of punitive damages. Studies in both domains have examined the following two hypotheses: (i) different measures of the valuation of issues are highly correlated, as in the semantic differential; (ii) dollar measures belong to the cluster of attitude measures.

Kahneman and Ritov (1994) studied the valuation of 37 topics, including a wide array of environmental problems and other public issues. The issues were presented as headlines, in which a brief description of a problem was followed by a single sentence describing a proposed intervention. An example was "THE PEREGRINE FALCON IS THREATENED BY POLLUTION. Intervention: Support special program to protect the Peregrine falcon." Several measures were used: SWTP for the proposed intervention, degree of political support for the interven-

tion, personal satisfaction expected from making a voluntary contribution (both on a 0–4 rating scale), and a rating of the importance of the problem as a public issue, on a 0–6 rating scale. The participants in the study were visitors at the San Francisco Exploratorium. Each participant used only one of these four response scales to evaluate anywhere from 9 to 19 assigned of the problems. The total sample was 1441, and the number of respondents to any particular version of a problem was 50–115.

The 37 problems were ranked by the sample means for each of the response measures. Rank correlations between these means are shown in Table 1. The numbers on the diagonal represent measures of reliability, obtained by a bootstrapping procedure. Table 1 indicates that the rankings of the issues by the different measures were quite similar. Indeed, the correlations between orders derived from different measures were not substantially lower than the reliabilities of the individual measures.

What do ratings of importance, predictions of moral satisfaction, statements of political support and indications of willingness to pay have in common? Our answer is that these expressions share a common affective core, which is so prominent that it allows the public attitude order over objects to be measured almost interchangeably by ostensibly diverse responses.

Payne et al. (1999) observed a similar result in a study of 190 citizens who responded to five CV surveys of realistic length and detail. The topics were air quality in the Grand Canyon, oil spill prevention, and preservation of wolves, salmon, and migratory waterfowl. Each respondent expressed an evaluation of each commodity in SWTP and on four 0–10 rating scales—importance compared to other problems in society, seriousness compared to other environmental problems, use value and existence value. Respondents came for two separate two-hour sessions, scheduled two weeks apart. In the first session a given respondent responded to all five commodities on either SWTP or the four rating scales. In the second, they again responded to all five surveys, but using the response mode(s) they did not use in the first session. The results showed rank correlation levels between response modes similar to those of Table 1 (ranging from .67 to 1.00), despite the many differences in stimuli and procedure from the Kahneman and Ritov study.

Table 1. Rank correlations between mean evaluations of 37 issues

	SWTP	Support	Importance	Satisfaction
SWTP	(.87)			
Support	.84	(.85)		
Importance	.76	.84	(.88)	
Satisfaction	.84	.87	.85	(.90)

From Kahneman and Ritov, 1994.

Our next example is drawn from a study which employed a similar design to study the psychology of punitive damages. Kahneman, Schkade, and Sunstein (1998) constructed 28 vignettes of cases in which a firm was found liable for compensatory damages in a product liability case. Each participant responded to a subset of 10 of these cases. Separate groups of respondents were asked to answer one of three questions about each scenario: "how outrageous was the defendant's behavior?" (on a 7-point scale), "how severely should the defendant be punished?" (on a 7-point scale), or "how much should the defendant be required to pay in punitive damages?" (in dollars). The respondents were 899 jury-eligible adults. An average of 107 respondents responded to each different case-question combination. The 28 cases were ranked by the mean ratings of outrage and punitive intent, and by the median dollar award. The correlations between these rankings are shown in Table 2.

Here again, we may ask what the three responses have in common that results in such high correlations. The outrage rating appears to be a rather direct measure of the affect evoked by cases of personal injury. The high correlations indicate that the same affective valuation also dominates ratings of punitive intent and judgments of punitive damages in dollars. The hypothesis that expressions of attitude are dominated by a shared affective reaction—in this case, by a degree of outrage—is again strongly supported.

The results shown in Tables 1 and 2 are correlations between averages of large samples, computed over objects. It is important to note that these correlations are not necessarily representative of the results that would be obtained within the data of individual respondents (Nickerson, 1995). As in the case of other summary statistics, it is possible for group results to be dominated by a few individuals who (i) produce more variance than others, and (ii) have an atypical pattern of responses. These hypotheses are readily testable (e.g., by examining the effects of standardizing the data of each individual), and we are satisfied that they did not apply to the data reported in this section.[2]

2-3) *Each expression of attitude also has its specific and distinctive determinants, but these account for less variance than the core affective value*. The example of justice being GOOD but not necessarily KIND was used earlier to show that different expressions of the evaluation factor in the semantic differential are not interchangeable. The same conclusion applies to the factor of affective valuation that

Table 2. Rank correlations between mean evaluations of 28 cases

	$ Awards	Outrage	Punishment
$ Awards (median)	(.89)		
Outrage	.80	(.96)	
Punishment	.92	.86	(.98)

From Kahneman, Schkade and Sunstein, 1998.

could be extracted from diverse responses in the data of Tables 1 and 2. It is convenient to analyze an expression of affective valuation as the sum of three separable components:

$$X = A + S + e \tag{1}$$

where A is the shared affective valuation, S is a response-specific component, and e is an error term. The high correlations shown in the previous section indicate that the first of these components accounts for much more variance than the second. The shared affective value dominates the diverse expressions of attitudes. As the following examples illustrate, however, the specific content associated with different responses is both interesting and important.

Kahneman, Schkade and Sunstein (1998) offered an ***outrage model*** to account for both the similarities and the differences between the measures of outrage, punitive intent and punitive awards. They examined the differences in two experiments. The first experiment demonstrated that rated outrage was the same regardless of whether harm was severe or mild. This result is intuitively plausible: a behavior can be judged as more or less outrageous without knowing its consequences. In contrast, ratings of punitive intent and assessments of punitive damages were both sensitive to the severity of harm. Punishment involves a retributive intent, which depends on the consequences of the act that is to be punished; this is the intuition that justifies treating murder and attempted murder as distinct crimes. A second experiment showed that the size of the defendant firm had a large effect on the amount awarded in punitive damages, but no effect whatsoever on either outrage or punitive intent. This result is also plausible: a payment that constitutes 'very severe' punishment for a small firm may be quite insignificant for a larger one. As in the early studies of the semantic differential, we observe a pattern of meaningful differences among highly correlated expressions of the same affective valuation. Detailed examinations of responses to public goods also reveal systematic discrepancies between highly correlated measures (Kahneman and Knetsch, 1992). As the high correlations in these studies suggest, however, the discrepancies between measures are small in magnitude, relative to the large common influence of the underlying affective valuation.

3. Valuation by prototype and the scope problem

The evidence reviewed in the preceding section confirmed the similarity between the rankings of objects by different measures of attitude, and provided suggestive evidence that the core of attitude is an affective valuation. In this section we argue that the affective valuation of a prototypical exemplar often determines the global attitude to sets of objects. We show that this process can explain an important finding of contingent valuation research: the inadequate sensitivity of SWTP to the quantitative aspects of problems and solutions.

3-1) *People hold stored prototypes of many categories. They also form prototypes or representative exemplars of new categories and sets that they encounter*. The prototypes of tables, of birds and of Harvard MBA's are widely shared among members of the relevant culture. People also form ad hoc representations of a typical day of a seaside vacation, or of a typical resident of a city they visit. These representations of prototypes are evoked in the service of thinking about concepts and classes (Barsalou, 1992).

3-2) *In **judgment by prototype**, a global judgment of a category or set is determined primarily by the relevant properties of its prototype*. The principle of judgment by prototype extends the older idea that a representativeness heuristic is involved in many intuitive judgments about uncertain events (Kahneman and Tversky, 1972, 1973; Tversky and Kahneman, 1971, 1983).

3-3) *When the size of the set is logically relevant to its valuation, judgment by prototype leads to a bias of **extension neglect**: Unless attention is specifically directed to it, the size of the set has little or no influence on its valuation. This pattern has been observed in different contexts, in which extension neglect takes different forms* (*Kahneman*, 1995). To illustrate the generality of the phenomenon of extension neglect, we briefly describe three examples:

(i) Intuitive statistical inferences are often made by assessing the similarity between the statistic of a sample and the parameter of a population. The sample and the population are both ensembles, but the judgment about them is based mainly on the relation between the prototypes that represent them. Intuitive inferences based on such reasoning are characterized by extreme lack of sensitivity to sample size, which is the form that extension neglect takes in this task (Griffin and Tversky, 1992; Kahneman and Tversky, 1972; Tversky and Kahneman, 1971).

(ii) In a familiar paradigm for the study of intuitive prediction, subjects judge the probability that an individual is a member of a specified social category (defined by a profession or an avocation) on the basis of a personality sketch (Kahneman and Tversky, 1973; Tversky and Kahneman, 1982). Probability is judged by the similarity of the individual's personality to the stereotype of the target category. For example, an individual described as "argumentative, flashy, self-confident and competitive" will be judged more likely to be a lawyer than to be an engineer, because the description resembles the stereotype of the former profession more than that of the latter. In this paradigm, extension neglect takes the form of inadequate sensitivity to the base rates of outcomes (Kahneman and Tversky, 1973; see also Koehler, 1996; Novemsky and Kronzon, 1999).

(iii) Extension neglect has also been observed in a paradigm in which participants are exposed for some time to an unpleasant experience. The participants provide a continuous report of current discomfort, using an 'affect meter.' Later they provide a global judgment of the entire episode. Various experiences have been studied, including unpleasant films (e.g., of an amputation), immersion of the hand in cold water, exposure to loud noise, and painful medical procedures (see Kahneman, Wakker and Sarin (1997) for a review). For our purposes, an episode of

discomfort can be construed as a set of unpleasant moments. The duration of the episode is the measure of extension. Valuation by prototype implies that participants will construct or remember a typical moment of the episode, and evaluate the episode as a whole by the level of unpleasantness associated with the prototypical moment—the duration of the episode will be neglected. The hypothesis of duration neglect has been confirmed in several experiments, with both ratings and choices as dependent variables (Kahneman, Wakker and Sarin, 1997).

In all three situations, judgment by prototype and extension neglect can cause violations of monotonicity. People commonly underestimate the strength of evidence provided by 'weak' results in a large sample, compared to stronger results in a small sample (Tversky and Kahneman, 1971). They assign a higher probability to the statement 'Linda is a bank teller and a feminist' than to the statement 'Linda is a bank teller,' if the description of Linda resembles the stereotype of a feminist but not the stereotype of a bank teller (Tversky and Kahneman, 1982). Because the prototypical moment of an episode of discomfort is strongly influenced by how the episode ends, adding a period of diminishing pain to an episode makes it less aversive, in violation of dominance (Kahneman et al., 1993).

3-4) *In some applications of contingent valuation, a problem or a solution is specified by the quantity of a homogeneous good. In such cases, extension neglect takes the form of* ***insensitivity to scope****: the quantitative attribute has little weight in the valuation, which is determined mainly by the affective response to a prototypical instance of the good.* Economic theory imposes stringent constraints on the response to variations in the quantities of a good. Diamond and his colleagues (Diamond et al., 1993; Diamond, 1996) have formulated these constraints as a simple add-up test for SWTP in CV surveys: after allowing for an income effect, SWTP for the conjunction of two parts should equal the sum of SWTP for one part, plus SWTP for the second part conditional on already having the first part. It is generally agreed that adequate sensitivity to scope is essential to the acceptability of CVM (NOAA panel on Contingent Valuation, 1993).

Sensitivity to scope has been studied in several research paradigms (see section 3-6). We are concerned here with a particular variant, the ***quantity design***, in which participants indicate their willingness to pay for a specified amount of a relatively homogeneous good.[3] The amount of the good is varied across groups of respondents. A well known example of this experimental design is due to Desvousges et al. (1992). The question these authors put to their respondents can be paraphrased as follows: "(2,000, or 20,000, or 200,000) migrating birds die each year by drowning in uncovered oil ponds, which the birds mistake for bodies of water. These deaths could be prevented by covering the oil ponds with nets. How much money would you be willing to pay to provide the needed nets?"

The principle of valuation by prototype applies in straightforward fashion to this example. The story constructed by Desvousges et al. probably evokes for many readers a mental representation of a prototypical incident, perhaps an image of an exhausted bird, its feathers soaked in black oil, unable to escape. The hypothesis of

valuation by prototype asserts that the affective value of this image will dominate expressions of the attitude to the problem—including the willingness to pay for a solution. Valuation by prototype implies extension neglect. Although the number of birds that die in oil ponds is surely a relevant consideration, we would expect that—unless the respondents' attention is specifically directed to it—the number of bird deaths will have little effect on SWTP or on other measures of attitudes. Indeed, mean SWTP was $80, $78 and $88, respectively, for saving 2,000 birds, 20,000 birds or 200,000 birds annually (Desvousges et al., 1992).

Similar results have been obtained in other applications of the quantity design. In an early study using this design, Kahneman and Knetsch (see Kahneman, 1986) found that Toronto residents were willing to pay only a little more to clean up all the polluted lakes in Ontario than to clean up polluted lakes in a particular region of Ontario. McFadden and Leonard (1993) reported that residents in four western states were willing to pay only 28% more to protect all 57 wilderness areas in those states than to protect a single area. Jones-Lee et al. (1995) found that the SWTP of UK respondents for a program to reduce the risk of non-fatal road injuries increased by only 29% when the number of prevented injuries was increased by a factor of three. Laboratory studies show similar insensitivity to the quantity of the good. Baron and Greene (1996, experiment 8), for instance, found no effect on SWTP of varying the number of lives saved by a factor of 10.

There is research in which the effects of quantitative variations appear to be larger, though certainly not enough to satisfy economic theory. For example, Carson and Mitchell (1995) describe an unpublished study of the value of reducing the risk associated with chlorination of drinking water. They report that an increase of risk from .004 to 2.43 annual deaths per 1,000 (a factor of 600) yielded an increase of SWTP from $3.78 to $15.23 (a factor of 4). This result does not contradict the general conclusion of other research in this area: the response to variations of scope is so slight that it is not explicable in the standard terms of economic analysis.

Explanations of insensitivity to scope in terms of an income effect are implausible, because the amounts are so small. Explanations in terms of substitution effects are equally unattractive. Several studies have shown that reminding subjects of substitutes or of their prior endowment does not substantially change their response (Loomis et al., 1994; Neill, 1995; Ritov, Baron and Hershey, 1993). An interpretation in terms of substitution effects, if it were taken seriously, would be potentially disastrous for the environment. It would indeed be good news for polluters if the public's demand for clean lakes in Ontario could be satisfied by cleaning up a small subset of its lakes.

Our aim in this section was not to deal with the details of the heated controversy concerning sensitivity to scope (see, for example, Carson and Mitchell, 1995; Frederick and Fischhoff, 1998). Our goal is both simpler and more ambitious: we hope to have shown that inadequate sensitivity to scope in CV surveys that employ the quantity design is *inevitable*, because this phenomenon is an instance of a broad

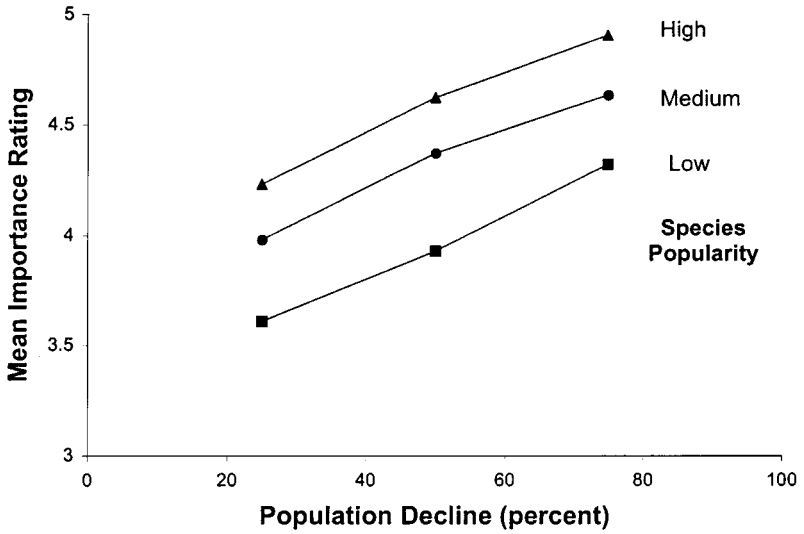
class of similar effects that have been observed in diverse contexts and are explained by a single psychological principle.

3-5) *Extension neglect is neither universal nor absolute. When extension information is both salient and readily interpretable an* ***additive extension effect*** *is observed: the effects of the valuation of the prototype and of the size of the relevant set are additive. This pattern violates normative rules that require non-linear combination of the two types of information.* In the situations we have discussed, the relevance of extension may be obvious if the quantity mentioned in the problem is readily classified as high or low. Under such circumstances, responses will show some sensitivity to extension. For example, even naive respondents will appreciate that an annual death rate of .0004% from chlorinated water is very low, because of the impressively large number of leading zeros. However, there are situations in which the quantitative information is less easily interpreted: unless the two numbers are seen together, for example, the subjective difference between two large quantities such as 20,000 or 200,000 birds dying in oil ponds is not very impressive (Hsee, 1996). These are the conditions under which complete neglect of scope may be observed.

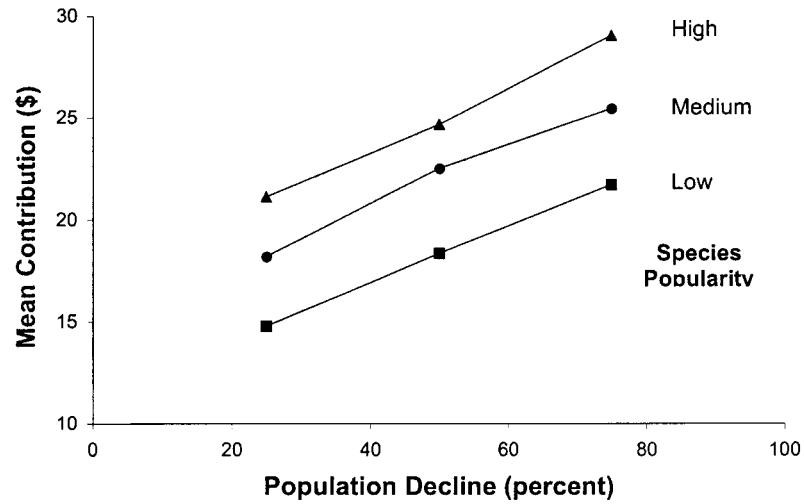
Studies of extension neglect in other domains have shown that multi-trial experiments in which extension varies from trial to trial have two effects: they draw attention to extension as a relevant feature, and they provide a standard that helps the subject assess values of the extensional attribute as high or low. Extension is not completely neglected under these conditions. Indeed, significant effects of extension have been found in within-S experiments in all the domains we have mentioned. When the base-rate of outcomes is varied from trial to trial, people pay attention to it (Novemsky and Kronzon, 1999). When the duration of episodes that are to be evaluated varies from trial to trial, duration neglect is imperfect (Schreiber and Kahneman, 2000; Varey and Kahneman, 1992). Sample size also affects judgments in within-subject experiments (Griffin and Tversky, 1992).

A remarkable regularity appears in these experiments: the valuation of the prototype and the extension of the set (base-rate or duration) contribute in strictly additive fashion to the global judgment (see also Anderson, 1996, p. 253). The participants in these experiments appear to reason as follows: "this medical procedure is quite painful, but it is short" or "this medical procedure is quite painful, and it is also long." In contrast to the logic of global evaluation, which requires multiplicative or quasi-multiplicative effects of extension, the size of the set is used as an extra feature in this reasoning.

The additive extension effect is also found in the valuation of environmental goods. Kahneman and Ritov (unpublished research) presented several groups of respondents messages such as the following: "the population of Dolphins in a coastal preserve has declined by 50%." The species mentioned ranged widely in emotional appeal, and the population decline was also varied. Some respondents rated the importance of the problem. Others indicated, for each species, how much of a contribution of $40 to a general environmental fund they would divert to restore the population of the species in the nature preserve. Figures 1a and 1b



(a)



(b)

Figure 1. a. Mean importance ratings, by species popularity and degree of population decline. b. Mean contributions, by species popularity and degree of population decline.

present the results for both response measures. The striking feature of these data is that both the dollar measure and the rating of importance exhibit nearly perfect additivity of the effects of species popularity and size of population decline. Precisely the same pattern of results has been observed in studies of individual prediction (Novemsky and Kronzon, 1999), and of the global evaluation of episodes

(Schreiber and Kahneman, 1999). A related result was obtained by DeKay and McClelland (1996), who found that the species attributes and the probability of survival were combined additively in people's ranking of programs to preserve endangered species.

We draw several conclusions from this research. First, some effect of extension can be obtained by a procedure, such as the within-subject experimental design, which simultaneously draws attention to the quantitative variable and provides a frame of reference for responding to it. Second, a demonstration that people can be responsive to extension and scope under some conditions is not sufficient to support the conclusion that they always use extension in accordance with the relevant logic. Third, and most important, we again find that the anomalies observed in studies of the value of public goods do not remain either puzzling or unique when they are viewed in the context of similar phenomena in other domains.

3-6) *Several different designs have been used to test sensitivity to scope. The designs are psychologically different, but the normative pattern defined by the add-up test (Diamond, 1996) is unlikely to be satisfied in any of them.* Sensitivity to scope has been examined in two designs other than the quantity design that was discussed in previous sections. (i) In the ***explicit list*** design, respondents in different groups value nested lists of heterogeneous goods. For example, one group may assess the value of saving both the birds and the fish in a region, while other groups value the birds or the fish in isolation. (ii) In the ***embedding*** design, SWTP for a good (e.g., saving dolphins) is obtained in two ways: (a) by a direct question (b) by a sequence of questions, first eliciting SWTP for an inclusive good, then the fraction of that amount that should be allocated to a specified good (e.g., SWTP for saving marine mammals, then an allocation to dolphins).

The various tests of scope are equivalent in an economic analysis, and Diamond's add-up test is applicable to all three. In a psychological analysis, however, the designs differ in important ways. The quantity design involves a set or category of elements that are similar in essential respects (e.g., polluted lakes, or different members of the same species). In contrast, the two other designs involve heterogeneous elements, which are not readily represented by a single prototype (Rosch and Lloyd, 1978). There is some evidence that a process of ***judgment by maximum*** operates in the valuation of heterogeneous categories and lists (Levav, 1996). A related result was reported by Rottenstreich and Tversky (1997) in a study of judgments of frequency for explicit lists (e.g., "How many Stanford students major in either English or Geography?" Judgment of the total frequency of an explicit disjunction were barely higher than judgments of its maximum. Judgment by maximum, of course, violates the add-up rule.

Carson et al. (1994) reported a study using an explicit list design, which they described as a demonstration of sensitivity to scope. Unfortunately, a basic flaw of their study invalidates their conclusions. The study was concerned with the valuation of the damage that deposits of DDT in the soil of LA Harbor has caused to

the reproductive ability of two salient species of birds (Peregrine Falcon, American Bald Eagle) and two relatively obscure species of fish (White Croaker, Kelp Bass). The authors observed higher SWTP when the description of the problem involved all four species than when it involved only the fish. Of course, the results are equally consistent with the hypothesis that WTP to restore two important species of endangered birds is higher than WTP to restore two relatively obscure species of fish. The hypothesis of judgment by maximum suggests that the value attached to the four species would not be much higher than the value attached to the most important of these species. The results of an informal replication of the LA Harbor study, using ratings of importance, were generally consistent with this hypothesis (Levav, 1996). There is no reason to expect that the results of CV using explicit lists will satisfy the add-up test (see also Frederick and Fischhoff, 1998).

The findings obtained in the embedding design unequivocally violate the add-up rule. For example, Kemp and Maxwell (1993) found that SWTP for protection from oil spills off the coast of Alaska was $85 when the good was considered on its own, but only $0.29 when it was derived as an allocation of SWTP for a more inclusive category (environmental protection programs). Kahneman and Knetsch (1992) reported similar findings.

The central point of this section has been that inadequate sensitivity to scope is not a surprise. On the contrary, it would be a major surprise to observe measures of SWTP that reliably conform to the add-up test. This conclusion is relevant to the frequently expressed hope that the scope problem might be overcome by improved instructions, exhortation or added information. Insensitivity to scope is the inevitable result of general rules that govern human judgment. It is naive to expect broad psychological laws to be overcome by minor methodological adjustments.

4. Context-dependence and valuation reversals

A preference reversal is said to exist when two strategically equivalent methods for probing the preference between objects yield conflicting results (see, e.g., Hsee, 1996; Tversky and Thaler, 1992). Preference reversals simultaneously challenge two basic tenets of the standard economic analysis of choice: the existence of a preference order, and the assumption of extensionality. One of the crucial differences between the concepts of economic preference and attitude is that preference reversals are anomalous only for the former, not for the latter. In this section we discuss preference reversals that arise from the context dependence of attitudes and affective valuations. Norm theory (Kahneman and Miller, 1986) provides the theoretical background for this discussion.

4-1) *An object that is considered in isolation evokes a comparison set of similar objects. The valuation of the object is relative to the set that it evoked. Features that are common to the evoked set play no role in relative judgments and valuations*. For an illustration of the relativity of judgment to an evoked set, consider the following

two questions: "Is a subcompact car BIG or SMALL?", "Is a bald eagle BIG or SMALL?" The plausible answers are that a subcompact is small and a bald eagle is big. The categories of cars and birds are spontaneously evoked by the mere mention of their members, and these categories provide the norm for a relative judgment of size. The conventions of language allow the entire range of size adjectives, from 'tiny' to 'enormous' to be applied to cars and to birds, to countries and to bacteria.

As we show later, expressions of attitudes show a similar relativity. Furthermore, casual observation indicates that affective values—not only the words used to describe them—are themselves relative. Thus, a guest's rude behavior at a party can arouse intense outrage and anger. Murder is much worse than rudeness, of course, but murder is not part of the evoked context that determines the emotional response to a rude remark. The relativity of affective value explains why people often seem to care intensely about matters that they can also view as trivial when the context changes.

4-2) *Explicit comparison of several objects imposes a shared context for their judgment and valuation. When the objects belong to different categories, comparisons and isolated valuations can yield discrepant results. Differences between the modes of valuation are found both in dollar measures and in ratings.* Table 3 presents preliminary tests of this hypothesis, drawn from two different studies. The same pair of issues was used in both studies: damage to coral reefs caused by cyanide fishing in Asia, and increased incidence of multiple myeloma among the elderly. We surmised that the latter issue would be perceived as a fairly minor public health problem, whereas a threat to coral reefs would appear significant in an ecological context. We also surmised that public health problems would be assigned a higher general priority than ecological problems, but that this priority would only become relevant in a direct comparison.

Table 3. Responses to an ecological and a public health problem, by presentation order

Study 1				
	Moral satisfaction		Importance	
	First	Second	First	Second
Coral Reefs	3.54	3.24	3.78	3.62
Myeloma	2.84	4.18	3.24	4.26
Study 2				
	Moral satisfaction		SWTP[a]	
	First	Second	First	Second
Coral Reefs	3.47	3.05	$45	$59
Myeloma	2.98	3.76	$69	$109

[a] All values of WTP in excess of $500 were adjusted to $500.

The procedure of the two studies was similar. The participants were first asked to evaluate one problem; they were then shown the other problem and were asked to respond to it, with an explicit instruction to consider both problems before responding. The independent variable was the order of presentation of the two problems. The participants in Study 1 were 100 visitors at the San Francisco Exploratorium. They were asked to rate the importance of each problem and the satisfaction they would expect to get from making a contribution to its solution (both on a 0–6 scale). The data for Study 2 are drawn from a larger study, in which participants were jury-eligible residents of Austin. Some participants ($N = 130$) provided ratings of satisfaction as in Study 1. Others ($N = 261$) indicated their WTP to contribute to a solution; when they encountered the second problem they were instructed to treat it as the only cause to which they would be asked to contribute.

Our hypothesis about the role of context in judgments predicts a statistical interaction effect in each of the panels of Table 3: the difference between the valuations of the myeloma and coral reefs problems is expected to be larger when these items appear in the second position than in the first. The rationale for this prediction is that the difference between the categories of ecological and human problems is only salient when the issues are directly compared, not when they are valued in isolation. The predicted interaction is highly significant ($p < .001$) in each of the four panels of Table 3.

The context effect observed in SWTP is especially noteworthy, because the linguistic convention that allows words such as 'important' or 'satisfying' to be understood in a relative sense does not apply to the dollar scale. To appreciate the difference between scales that allow relativity and scales that do not, consider the questions: "What is the size of an eagle, in meters?", "What is the size of a subcompact, in meters?" Of course, there is no reason to expect any effect of category on answers to this question. A context effect on a size judgment expressed in absolute units indicates a visual illusion—a change in the underlying perception, not in the language used to describe it. By the same logic, the finding of a context effect on a dollar measure implies that the evaluation itself, not only the expression of it, is altered by the comparison.

Kahneman, Schkade and Sunstein (unpublished data) investigated the effects of a comparison context on punitive damage awards. The study was motivated by the observation that the highest punitive awards are commonly found in cases involving large financial harm, probably because the size of the compensatory damages provides a high anchoring value. Punitive damages are generally lower in cases of personal injury, where compensatory damages are also lower. We surmised, however, that cases that result in personal injury are, as a class, more outrageous than cases in which the only losses involve money. Of course, no jury ever considers cases of the two types at the same time, but we predicted that forcing jurors to do so in an experiment would alter or reverse the usual pattern of punitive awards. A sample of 114 jury-eligible citizens provided a punitive damage assessment for either a personal injury case adapted from Kahneman, Schkade and Sunstein

(1998): (a child seriously hurt because of a flawed child-safety cap), a financial harm case (business fraud), or both. Participants were told that compensatory damages had already been awarded, in the amount of $500,000 for the personal injury and $10 million for the financial harm. As predicted, respondents who judged only one case assessed greater punitive damages in the financial case (median = $5 million) than in the personal injury case (median = $2 million). However, a strong majority (75%) of respondents who judged the two cases together assessed larger awards in the personal injury case, resulting in a striking reversal of median awards ($2.5 million for the personal injury; $0.5 million for the financial harm). More recent data indicate similar effects in ratings of outrage and punitive intent. The same interpretation applies to these results and to the findings summarized in Table 3. Cases of personal injury and of financial harm, when considered in isolation, are apparently compared to very different evoked contexts. Here again, the conclusion that the context alters the underlying emotion is justified by the finding of an effect on a dollar measure. A financial transgression that appears outrageous on its own apparently arouses much less outrage when directly compared to an action that causes a child to suffer a severe injury (Kahneman, Schkade, Ritov and Sunstein, 1999).

4-3) *Choice is a special case of comparative valuation, whereas pricing (or the setting of WTP) is normally done by considering a problem in isolation. The different contexts of choice and pricing explain some preference reversals between the two tasks.* The analysis of context effects in the preceding section helps explains preference reversals between choice and SWTP that were reported by Kahneman and Ritov (1994). Seven critical pairs of items were constructed, each including one ecological issue and one public health problem. The responses of two groups of respondents were compared. One group encountered both items in a questionnaire that elicited statements of WTP for interventions to alleviate each of several (12–14) problems, which the respondents were instructed to consider independently. Other respondents were asked to make a choice between two items from the same list. They were told that "It sometimes happens that budget constraints force a choice between two desirable projects. One has to be given up, at least for now, so that the other can go forward." The respondents were then asked which of the two interventions they would retain, if they had to make this choice.

Robust reversals of preference were obtained. On average, only 41% of the respondents who stated different WTP for the two items indicated greater willingness to pay for the public health problem.[4] However, 66% of responses favored the public health issues in the choice condition. The difference between the two conditions was statistically significant separately for each of the seven pairs of items. A different pattern was observed in five other pairs of issues, in which the two issues were drawn from the same category. In these control pairs, the proportions favoring one issue over another were quite similar in choice and in SWTP.

4-4) *The context dependence of valuations suggests three observations*: (*i*) *the hope of measuring preferences by SWTP is unrealistic*; (*ii*) *a suitable choice of context may help improve the rationality of elicited preferences*; (*iii*) *there is no general attitude order, but modeling context-dependent affective valuations is feasible in principle*. The finding of preference reversals between SWTP and choice implies that willingness to pay does not provide a stable measure of the position of an object in a preference order—in our view, because there is no stable preference order to be measured. Like the scope problem that was discussed in the preceding section, the context dependence of SWTP is an unavoidable consequence of basic cognitive and evaluative processes. It is not a result of defective procedures, and it will not be eliminated by improved survey methods.

The reversals of valuation that we have observed in both SWTP and punitive damages raise a significant prescriptive question: When different methods for eliciting attitudes yield conflicting results, which method should be used? In general, of course, decisions that are based on a richer set of considerations and on a broader context are more likely to be stable and to satisfy standards of rationality. This principle suggests that asking people for choices may be better than asking them to consider issues in isolation. We have seen, for example, that the priority of public health over ecological concerns is effectively suppressed in the SWTP measure, and only becomes evident when respondents must compare items from the two categories. Similarly, the difference in the outrageousness of actions that cause physical or financial damage was suppressed when cases were considered in isolation, and only revealed by a direct comparison. The benefits of improved rationality are more likely to be achieved if the context of comparison is truly broad, and if it has been selected impartially. Mere exhortations to consider many possibilities (NOAA panel, 1993) are not likely to be effective.

Our findings provided further evidence for a simple negative conclusion: there is no comprehensive and coherent 'attitude order.' This is not a message of despair. The phrase "Individual I likes/dislikes to extent X the description D of object O, considered in context C" is, at least in principle, subject to measurement, verification and modeling. We already know, for example, that different measures of liking will yield similar estimates of X, and that if two objects spontaneously evoke the same context C, measurements of their relative preference by liking and by choice will probably be consistent. Attitudes do not lack structure, but their structure is vastly more complex than the structure that economic analysis attributes to human preferences.

5. The psychophysics of valuation

The results of section 2 demonstrated that average dollar responses for large groups yield much the same ranking of attitude objects as do other measures of attitudes. To the proponents of contingent valuation or to the supporters of the jury system this is faint praise, because they need much more than a ranking of

objects. The goal of asking survey respondents to assess a public good or of asking jurors to assess punitive damages is to obtain a dollar value that is meaningful in absolute terms, not only in relation to other objects. Can this goal of absolute measurement be realized? In this section we draw on psychophysical research to examine the measurement properties of the dollar scale, and to compare it to other measures of affective valuation.

5-1) *The attitude expressions elicited in surveys can be classified as* ***category scales*** or ***magnitude scales***. These terms are borrowed from the field of psychophysics, the study of the functions that relate quantitative expressions of subjective reactions to physical variables. For example, the perceived loudness of tones that vary in amplitude can be measured on a bounded category scale (e.g., from 'not loud at all' to 'very very loud'). Loudness can also be measured on a magnitude scale by presenting the subject with a series of tones, with the instruction to assign a given number (known as the ***modulus***) to a specified standard tone, and to assign numbers to other tones relative to this common modulus. The defining characteristics of a magnitude scale are that it is unbounded, has a meaningful zero, and expresses the ratios of the relevant underlying variable.

In terms of this classification of scales, the normal practice of survey research is to use category scales, such as numerical ratings on a bounded scale. However, attitudes can also be measured using magnitude scales (Lodge, 1981; Stevens, 1975). For example, Stevens (1975) reported judgments of the severity of crimes, and also of the severity of different legal punishments, using an unbounded magnitude scale.

5-2) *Studies of magnitude scaling in the context of psychophysical measurement have yielded several generalizations, which apply as well to the domain of attitude measurement* (*Stevens*, 1975). (i) There is a fair degree of agreement among observers on the *ratios* of the magnitudes that they assign to the sensations evoked by particular stimuli. (ii) In the absence of a designated common modulus, there are large individual differences in the mean values of judgments: some observers assign generally high numbers to all stimuli, others assign low numbers. (iii) The distribution of responses to any stimulus is positively skewed; a log-normal distribution often provides an adequate fit. (iv) The standard deviation of the judgments of different stimuli is approximately proportional to their means; this relationship holds both when the same individual judges each stimulus several times and when the judgments are contributed by different observers. In contrast, category scales are characterized by a negligible correlation between the mean and the standard deviation of judgments. (v) In general, magnitude judgments of sensory intensity are a power function of the relevant physical variable: for example, brightness is a power function of luminance and loudness is a power function of sound amplitude (both with an exponent of approximately $1/3$). (vi) Magnitude scales are generally related by a power function to category scales of the same stimuli.

5-3) *The elicitation of dollar responses is a special case of magnitude scaling without a modulus*. The scale of dollars is unbounded and its zero is a meaningful response; the respondents (participants in CV surveys or jurors in civil cases) are not provided with a standard problem to which a specified dollar amount must be assigned (i.e., a modulus). The defining characteristics of scaling without a modulus are therefore satisfied. The results obtained with dollar scales are similar to the results that are observed with magnitude scales in psychophysical studies. In particular, the distribution of dollar responses is positively skewed, both within the responses of each individual and within the responses to any given problem. The distribution of the mean dollar judgments of individual respondents is also highly skewed. Finally, the high correlation between the mean and the standard deviation of individuals, which is expected for magnitude scales, was observed both by Kahneman and Ritov (1994; r = .93) and by Kahneman, Schkade and Sunstein (1998; r = .90).

5-4) *As expected for an application of magnitude scaling without a common modulus, dollar responses are statistically less efficient than category scale measures of the same attitudes*. We have seen that the averages of different attitude expressions in large samples yield similar rankings of objects (see Tables 1 and 2). However, dollar responses produce much lower signal-to-noise ratios than do rating scales. Tables 4 and 5 present results from separate analyses of variance for each of the response measures used in the two studies. The analysis partitions the variance of responses into three components: (i) *Object* (*signal*): the variance associated with differences among objects of judgments (e.g., public goods that differ in value, personal injury cases that vary in the outrageousness of the defendant's actions). (ii) *Respondents*: the variance associated with individual differences in the mean level of responses, over objects (e.g., some respondents state generally higher WTP than others, some experimental jurors are generally more severe than others). (iii) *Noise*: the residual variance, which combines the effects of individual differences in variability, idiosyncratic responses of some respondents to some objects or topics, and various sources of measurement error.

Tables 4 and 5 document a striking discrepancy in the strength of the signal (as indicated by the proportion of variance explained) between dollar measures and

Table 4. Proportion of variance explained by problems

	Raw	Ranks
Support	.08	.26
Importance	.16	.28
Satisfaction	.12	.26
SWTP	.04	.23

From Kahneman and Ritov, 1994.

Table 5. Proportion of variance explained by scenarios

	Raw	Ranks
Outrage	.29	.42
Punishment	.49	.58
$ Awards	.06	.51

From Kahneman, Schkade and Sunstein, 1998.

attitude expressions measured on standard bounded scales. The proportion of Object variance (i.e., signal) was 2 to 4 times larger for rating scales than for SWTP in Kahneman and Ritov (1994). The advantage of the rating scales was even more pronounced in responses to product liability cases, where the amount of Object variance was 5 to 8 times higher for ratings than for dollar responses (Kahneman, Schkade and Sunstein, 1998).

The low signal/noise ratio of dollar awards implies poor agreement among individuals, and even among juries. Kahneman, Schkade and Sunstein (1998) used Monte Carlo techniques to assess the average rank-correlation between dollar awards across cases for simulated "juries" of size 12: the estimated reliability (.42) appears unacceptably low.[5]

5-5) *Some transformations of dollar responses improve statistical efficiency, by reducing the effects of the skewness of magnitude scales and of the large individual differences in moduli*. For example, logarithmic and rank transformations of each individual's dollar responses both yield substantial improvements of signal/noise ratio. Transforming SWTP responses to a logarithmic scale doubled the percentage of Object variance (from 4% to 8%), to a level comparable to the other measures. Logarithmic transformation of punitive awards yielded even more dramatic improvement (Object variance increased from 6% to 42%). As shown in Tables 4 and 5, a ranking transformation also yielded a substantial increase in the relative amount of Object variance in both studies. The success of these transformations is due to the fact that the effect of individual differences in the use of the dollar scale is reduced by the logarithmic transformation and eliminated by the ranking transformation. The good performance of the transformed measures also demonstrates that the dollar response contains useful information about respondents' attitudes. If the objective of research is to rank order a set of objects, the dollar response—suitably transformed, and with a sufficiently large sample—provides as much information as other expressions of affective evaluation. Of course, the proponents of CV and of the current jury system hope for much more, since their goal is to obtain an exact dollar amount.

5-6) *Individual differences in the use of the dollar scale are large, and may be arbitrary to a substantial extent*. In psychophysical research, magnitude scaling without a common modulus yields large individual differences in the responses to stimuli, because subjects spontaneously adopt quite different moduli. If two subjects who

share the same underlying psychophysical function adopt different moduli, their responses to all stimuli will differ by a constant of proportionality, which is the ratio of their individual moduli.

In the psychophysical laboratory, differences in moduli are usually considered to be entirely arbitrary, a mere source of statistical noise. Except for very unusual circumstances (e.g., deafness), there is little reason to believe that an individual who consistently assigns low numbers to the loudness of tones actually experiences less loudness than an individual who assigns higher numbers. Are the moduli that CV respondents and jurors apply in assigning dollar responses also arbitrary? A positive answer to this question would remove the rationale for any procedure in which the absolute values that people state are taken seriously, including contingent valuation and the setting of monetary punishments by juries.

There are several ways of testing whether individual differences in the use of the dollar scale are meaningful or arbitrary. (i) *Prediction of behavior*. Several studies have examined the correlation between hypothetical responses to WTP questions and actual behavior (e.g., Cummings, Harrison and Rutstrom, 1995; Foster, Bateman and Harley, 1997; Seip and Strand, 1992). The data indicate a substantial upward bias in hypothetical responses. (ii) *Search for correlated variables*. If the difference between high-SWTP and low-SWTP respondents is real, it should be correlated with other characteristics of these individuals, such as income, or other indications of involvement in environmental issues. These correlations have been examined in some studies, and are usually low or nonexistent. Kahneman, Schkade and Sunstein (1998) also failed to find significant correlations between the average size of the awards set by individual respondents and several relevant predictors, including demographic attributes and individuals' ratings of the importance that they attached to different features of the cases, such as the degree of malice or the amount of harm suffered by the plaintiff. (iii) *Susceptibility to anchoring*. The large anchoring effects that we discuss in the next section indicate that dollar responses are very labile, both in CV surveys and in punitive awards. Arbitrary numbers that are mentioned in a question have considerable influence on responses—much as arbitrary moduli do.

We do not yet have the data needed to evaluate the relative size of the arbitrary and of the meaningful components in the variability of dollar responses. The available evidence, however, hardly justifies reliance on the absolute values of judgments denominated in dollars. There is at present no reason to believe that dollar responses contain useful information that cannot be obtained more simply and accurately by using other expressions of attitudes.

6. Anchoring effects

The procedure of asking people to state their maximal WTP for a good has been largely supplanted in CV practice by a protocol in which respondents are asked how they would vote in a hypothetical referendum that would guarantee the

provision of public good at a specified cost to the household. Different groups of respondents face different proposed payments, and the cumulative frequency distribution of positive responses is used to estimate the parameters of the underlying distribution of WTP. The estimates of WTP that are generated by this estimation technique are substantially higher than the estimates obtained by an open-ended question, such as "what is the maximum amount of payment for which you would still support the proposition?" (Desvousges et al., 1992; McFadden, 1994). The referendum format has been defended on grounds of its supposedly superior incentive compatibility (Hanemann, 1994; Hoehn and Randall, 1987). We do not directly debate this claim here (see Green et al., 1998). Following the broad strategy of this article, we show instead that the discrepancy between the two types of WTP questions can be parsimoniously explained by a well-understood process of anchoring, which produces similar effects in contexts to which the incentive compatibility idea does not apply.

6-1) *Tasks in which respondents indicate a judgment or an attitude by producing a number are susceptible to an* ***anchoring effect****: the response is strongly biased toward any value, even if it is arbitrary, that the respondent is induced to consider as a candidate answer*. Anchoring effects are among the most robust observations in the psychological literature. In a striking demonstration Wilson and his collaborators induced an anchoring effect by the following procedure: they required subjects to write the last four digits of their SSN, then to state whether they thought that the number of physicians and surgeons listed in the local yellow pages was higher or lower than that number. Finally, the subjects provided an open-ended estimate of the number of physicians and surgeons. The estimates that different subjects offered were strongly correlated with their social security number (Wilson et al., 1996). The necessary and apparently sufficient conditions for the emergence of anchoring effects are (i) the presence of some uncertainty about the correct or appropriate response, and (ii) a procedure that causes the individual to consider a number as a candidate answer. A vast literature has documented anchoring effects in estimation tasks (see, e.g., Strack and Mussweiler, 1997; Wilson et al., 1996), as well as in other settings, including negotiations (Ritov, 1996), and the setting of both compensatory (Chapman and Bornstein, 1996) and punitive awards (Hastie, Schkade and Payne, 1999).

Jacowitz and Kahneman (1995) proposed an index of the size of anchoring effects, which they applied to estimation tasks. They first obtained a distribution of answers to open-ended questions about quantities such as the length of the Amazon or the height of the tallest redwood, and observed the 15th and 85th percentiles of the estimates for each quantity. These values were used as anchors for two additional groups. Respondents in these anchored groups first answered a binary question such as "is the height of the tallest redwood more or less than X?", where the value of X was either the high or the low anchor for that problem. The anchoring index was defined as a ratio. The numerator is the difference between the median estimates of the anchored groups; the denominator is the

difference between the high and low anchors. By this measure, the anchoring effects were very large: the median anchoring index in a set of 20 problems was .49.

6-2) *Anchors have a suggestive effect on the answers to binary questions. With scales bounded on one side (such as the dollar scale) this effect causes an upward bias in binary answers, relative to corresponding open-ended responses. In the context of CV surveys, this bias explains the discrepancy previously observed between estimates of WTP from referendum questions and from open-ended questions.* The design employed by Jacowitz and Kahneman (1995) allows a comparison between two proportions: (i) the proportion of respondents in the original group (unanchored) who spontaneously offered an estimate higher than an anchor; (ii) the proportion of respondents in the anchored group who stated that the same anchor was lower than the true value of the quantity. In the absence of bias, the two proportions should be the same. However, the results showed a pronounced bias: on average, respondents in the anchored group judged the high anchor to be lower than the true value on 27% of occasions, very significantly more than the 15% expected from the responses of the unanchored group. Furthermore, there was a pronounced asymmetry in the bias: the low anchors were judged to be too high on only 14% of occasions. The asymmetry was due to the prevalence of estimation problems in which the range of possible answers is bounded by zero, e.g., the height of the tallest redwood. The result of this bias, of course, is that the estimates inferred from the binary question were generally much higher than the estimates obtained directly from open-ended questions. The discrepancy between the two response modes is similar to the discrepancy observed in CV research between estimates of WTP derived from open-ended and from referendum questions (Desvousges et al., 1992; McFadden, 1994).

The similarity between the effects of anchors on estimates of uncertain quantities and on SWTP were explored in a study reported by Green et al. (1998). Visitors at the San Francisco Exploratorium were recruited to answer five questions, including estimates of three quantities (height of the tallest redwood in California, average monthly gasoline used by car owners, annual rainfall in wettest spot on earth) and two WTP questions (save 50,000 off-shore seabirds each year from dying in oil spills, reduce auto accidents in California by 20%). The first and the last questions in each questionnaire were WTP questions. As in the Jacowitz-Kahneman study, a calibration group provided open-ended answers to all five questions. Five anchored groups answered a binary question about each quantity before estimating it. The anchors used in the binary question were chosen to be at the percentiles 25, 50, 75, 90 and 95 of the distribution of open-ended responses.

As expected, comparison of the anchored open-ended responses to the responses of the unanchored groups revealed a large anchoring effect, in both estimation and WTP questions. For example, the mean estimate of the height of a tallest redwood ranged from 282 feet (with 180 ft as an anchor) to 844 ft (with an anchor of 1,200 ft). Similarly, mean SWTP to save 50,000 birds annually ranged from $20.30 (with a $5 anchor) to $143.12 (with a $400 anchor).

An anchoring effect was also observed in answers to binary questions, for both estimates and SWTP. On average, there were 4.3% of answers exceeding the highest anchor in the calibration group, but 21.6% of respondents in the anchoring condition judged the same anchor to be too low. The pattern for low anchors was quite different: 21.5% of unanchored answers were lower than the low anchor, but the same anchor was judged to be too high on only 15.8% of occasions. As in the earlier study, high anchors induced a much larger bias. As a consequence of this asymmetric anchoring effect, the cumulative distribution derived from binary questions stochastically dominated the distribution of open ended answers. Over the five questions, the average ratio of the mean of the distribution inferred from binary questions to the unanchored mean was 3.43 (2.97 for the three estimation questions, 4.13 for the two WTP questions).[6]

This study again illustrates the benefits of searching for parallel phenomena across domains. The psychological analysis reveals that the tasks of estimating positive quantities and of determining a willingness to pay are deeply similar to each other, in both their open-ended and binary versions. They yield similarly skewed distributions of responses, are susceptible to similarly asymmetric anchoring effects, and therefore produce the same discrepancy between the parameters estimated from open-ended and from binary questions. In light of these observations, an explanation of the discrepancy in estimates of WTP in terms of incentive compatibility has little appeal, because it cannot be applied to the identical finding in another task.

7. Applications

The central claim of this paper has been that people are better described as having attitudes than preferences—perhaps in every domain, but certainly in the domain of public concerns. In contrast, CVM is rooted in the assumption that conventional consumer theory applies to public goods, including non-use goods such as the continued existence of the whooping crane. At least in principle, the dollar value of such a good could be read off an individual's preference order. The assumption of an inclusive preference order appears to be widely shared among economists, including critics of CVM (e.g., Diamond and Hausman, 1994) and among rational-agent theorists in political science (see Bartels (1998) for a discussion). In this theoretical framework, the main question to be asked about contingent valuation is the accuracy of measurement that it provides.

The problem with CVM, in our view, is not imprecise measurement but an incorrect theory. If consumer theory does not capture the nature of people's value for environmental goods, there can be no more hope of measuring the economic value of the whooping crane than there is of measuring the physical properties of the ether. Of course, many people do value the whooping crane and will even pay to preserve it. We have described these people as having a positive affective

valuation of whooping cranes, which induces a positive attitude to interventions that will preserve this species. These valuations can be expressed in many ways, including statements of WTP, actual payments, and votes in both simulated and real referenda. Attitude objects can be ordered reliably by sample averages of diverse expressions of valuation, including SWTP. As we have seen, however, these valuations lack some of the essential properties that economic theory requires of preferences. In particular, expressions of affective valuation are susceptible to framing effects (Bartels, 1998; Zaller, 1992), inadequately sensitive to scope and severely context dependent. Moreover, dollar measures of valuation are especially susceptible to the effects of anchors and of arbitrary moduli.

The extreme context-dependence of attitudes undermines the most compelling rationale that has been offered for the contingent valuation method. As Hanemann (1994) pointed out, the referendum question presents the respondent with a realistic task of formulating a voting intention, and answers to such survey questions have often been found to predict voting outcomes with fair accuracy. However, the only permissible inference from this argument is that CVM results predict the outcome of a real referendum that precisely mimics the context and framing of the survey question (Payne, Bettman and Schkade, 1999). The results do not provide reliable information about the voting outcomes that would be obtained with different wording of the question, or if the target proposition were embedded in a particular list of propositions. The evidence that SWTP diminishes steadily when several causes are considered in sequence (Carson and Mitchell, 1995; Payne et al., 1999) is another illustration of context dependence and another demonstration that CVM results are not sufficiently robust to provide a basis for policy.

Our pessimism about the validity of CVM does not imply despair about the possibility of using public attitudes as an aid to policy making. The affective value that people attach to issues probably conveys useful information about their possible reactions to policy proposals or to actual outcomes. More formal approaches to the elicitation of priorities are also possible, if they are developed with adequate respect for the psychology of valuation. For example, a scale of value for environmental damage could be developed by constructing a small set of hypothetical benchmark scenarios, covering a broad range of damage magnitude and commodity importance. Two criteria for including scenarios in the scale would be: (i) high consensus in the attitudes of the public to the scenario; and (ii) a hope of achieving professional and political consensus on appropriate dollar values. Public attitudes would be one input into this process, but probably not the only one. We expect that experts would bring in relevant considerations that lay judgment is prone to neglect, such as the scope and duration of the damage. The objective of the scaling effort would be to provide a mapping from attitudes and other relevant factors to dollar values for a particular class of environmental commodities.

Once a scale is established, a real issue that arises could be valued by survey in which respondents would explicitly compare the current problem to the benchmark scenarios. The measures of attitude used in this comparison would be chosen by

psychometric criteria: measures of judged importance and political support would probably be preferred to SWTP. A dollar value would be assigned based on the rank of the target issue among the benchmark scenarios of the standard scale. One advantage of this proposal is that the difficult conceptual problems of anchoring the dollar value of public goods in the preferences and opinions of the citizenry would be addressed just once, in the process of constructing the initial scale linking monetary value to attitude. Clearly, professional and political consensus is more likely to be achieved in dealing with hypothetical questions constructed for this purpose than in evaluating real goods in the context of litigation. Rutherford, Knetsch and Brown (1998) make a similar argument and propose that damage schedules be developed to replace ad hoc valuation based on SWTP responses.

The other domain that we have discussed, the setting of punitive damages, is a descendant of an old tradition which requires a small group of citizens to express their attitudes in dollars. It is remarkable that the jury system appears designed to enhance rather than minimize the deficiencies of human judgment: juries are instructed to consider cases one at a time, using a dollar measure without a modulus. Not surprisingly, dollar awards are erratic, in spite of a high level of agreement on ratings of outrage and punitive intent. Sunstein, Kahneman and Schkade (1998) provide a detailed analysis of possible reforms of the jury's task, which would require jurors to do what they can do well, not what they can do poorly. The determination of what jurors can do well combines normative evaluations with empirical facts. For example, if a normative analysis concludes that juror's intuitions about appropriate severity of punishment are valid, but their ability to translate these intuitions into dollars is weak—a plausible conclusion in view of the data reported here—the system could be reformed by requiring jurors to provide graded verbal statements of the severity of punishment that they consider just, leaving to the judge the task of translating this intent into a dollar amount.

Taken together, the examples of CV and punitive damages show that the debate about the nature of preferences and about the rationality of agents is not merely theoretical. The procedures that lead to some significant societal decisions may take different forms, depending on whether the decisions of individual citizens are best understood as a reflection of attitudes or of standard economic preferences.

8. Concluding remarks

The stereotyped role of the psychologist in the inter-disciplinary conversation about the nature of human choice is that of a critic, engaged in the construction of counter-examples to the economist's rational models. We have attempted to expand this role here, by focusing on the power and generality of psychological principles, rather than on the limitations of rational choice theory. Our theme has been that phenomena that appear anomalous from the perspective of standard preference models are in fact predictable—indeed, inevitable—consequences of well-established rules of judgment and valuation, which apply in domains that are

beyond the reach of choice theory. The alternative to rational choice as a descriptive model is neither chaos nor an endless list of ad hoc claims. It is a manageable set of concepts and testable propositions, which often predict surprising parallels between ostensibly different behaviors in different domains.

The evidence that we have discussed in this article was restricted to hypothetical questions. However, the progression of ideas from the explanation of hypothetical questions to the understanding of economically consequential behavior has an encouraging history, albeit a brief one (much of it is collected in Thaler, 1992). An example is the notion of loss aversion (Tversky and Kahneman, 1991), which was originally formulated in the context of hypothetical choices between gambles, further developed in market experiments with real stakes, and eventually extended to significant economic phenomena. The idea that some actions are expressions of affective valuations is, in our view, a candidate for a similar trajectory.

Acknowledgments

The U.S.-Israel Binational Science Foundation, the National Science Foundation and the Environmental Protection Agency provided support for the preparation of this article. Shelley Chaiken, Jack Knetsch, Kristine Kuhn and Barbara Mellers provided valuable comments.

Notes

1. The terms ‘valuation’ and ‘affective value’ are not standard in the attitude literature, but the position we take is widely shared.
2. Within-subject correlations were computed in the study of Payne et al. (1999) and they were quite high: the median correlation between rating scales was .69, and the median correlation between rating scales and individual SWTP was .51. The lower value of the correlations with SWTP is due to the high degree of noise in dollar responses (see section 5).
3. In the currently most popular variant of CVM, known as the referendum format, respondents are not required to state their maximal SWTP, but only to answer a yes-no question about their willingness to pay a specified amount. The distribution of SWTP is then inferred from the responses to various amounts. We discuss the referendum method in section 6.
4. SWTP was the same for the two issues in about 40% of the cases—most often because both responses were zero.
5. The higher value shown in Table 2 (.89) was obtained with “juries” of 107 members.
6. These results are based on a parametric estimation procedure described in detail by Green et al. (1998). A non-parametric estimation procedure yielded similar ratios: 2.14 for uncertain quantities, 2.22 for SWTP.

References

Anderson, Norman. (1996). *A Functional Theory of Cognition*. Mahwah, NJ: Erlbaum.
Arrow, Kenneth. (1982). “Risk Perception in Psychology and Economics,” *Economic Inquiry* 20, 1–9.

Baron, Jonathan and Joshua Greene. (1996). "Determinants of Insensitivity to Quantity in Valuation of Public Goods: Contribution, Warm Glow, Budget Constraints, Availability, and Prominence," *Journal of Experimental Psychology*: *Applied* 2, 107–125.

Barsalou, Lawrence. (1992). *Cognitive Psychology*: *An Overview for Cognitive Scientists*. NJ: Erlbaum.

Bartels, Larry M. (1998). "Democracy With Attitudes," Paper presented at the annual meeting of American Political Science Association, Boston.

Carson, Richard, Michael Hanemann, Raymond Kopp, John Krosnick, Robert Mitchell, Stanley Presser, Paul Ruud, and V. Kerry Smith. (1994). *Prospective Interim Lost Use Value Due to DDT and PCB Contamination in the Southern California Bight*. La Jolla, CA: Natural Resource Damage Assessment.

Carson, Richard and Robert Mitchell. (1993). "The Issue of Scope in Contingent Valuation Studies," *American Journal of Agricultural Economics* 75, 1263–1267.

Carson, Richard and Robert Mitchell. (1995). "Sequencing and Nesting in Contingent Valuation Surveys," *Journal of Environmental Economics and Management* 28, 155–173.

Chapman, Gretchen, and Brian Bornstein. (1996). "The More You Ask for the More You Get: Anchoring in Personal Injury Verdicts," *Applied Cognitive Psychology* 10, 519–540.

Cummings, Ronald, Glenn Harrison, and Elizabeth Rutstrom. (1995). "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible?" *American Economic Review* 85, 260–266.

DeKay, Michael and Gary McClelland. (1996). "Probability and Utility Components of Endangered Species Preservation Programs," *Journal of Experimental Psychology*: *Applied* 2, 60–83.

Desvousges, William, F. Reed Johnson, Richard Dunford, Kevin Boyle, Sarah Hudson, and K. Nicole Wilson. (1992). *Measuring Non-Use Damages Using Contingent Valuation*: *An Experimental Evaluation of Accuracy*. Research Triangle Institute Monograph 92-1.

Diamond, Peter. (1996). "Testing the Internal Consistency of Contingent Valuation Surveys." *Journal of Environmental Economics and Management* 30, 337–347.

Diamond, Peter, John Hausman, Gregory Leonard, and Michael Denning. (1993). "Does Contingent Valuation Measure Preferences? Experimental Evidence." In J. A. Hausman (ed.), *Contingent Valuation*: *A Critical Assessment*. Amsterdam: North-Holland.

Diamond, Peter and John Hausman. (1994). "Contingent Valuation: Is Some Number Better Than No Number?" *Journal of Economic Perspectives* 8, 45–64.

Eagly, Alice and Shelley Chaiken. (1993). *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace.

Eagly, Alice and Shelley Chaiken. (1996). "Attitude Structure and Function." In Gilbert, D., S. Fiske, and G. Lindzey (eds.), *The Handbook of Social Psychology*, 4th ed. New York: McGraw-Hill.

Fazio, Russell, David Sanbonmatsu, Martha Powell, and Frank Kardes. (1986). "On the Automatic Activation of Attitudes." *Journal of Personality and Social Psychology* 50, 229–238.

Fischhoff, Baruch. (1991). "Value Elicitation: Is There Anything in There?" *American Psychologist* 46, 835–847.

Foster, Vivien, Ian J. Bateman, and David Harley. (1997). "Real and Hypothetical Willingness to Pay for Environmental Preservation: A Non-Experimental Comparison," *Journal of Agricultural Economics* 48, 123–138.

Frederick, Shane and Baruch Fischhoff. (1998). "Scope (In)Sensitivity in Elicited Valuations," *Risk Decision and Policy* 3, 109–123.

Green, Donald, Karen Jacowitz, Daniel Kahneman, and Daniel McFadden. (1998). "Referendum Contingent Valuation, Anchoring, and Willingness to Pay for Public Goods." *Resource and Energy Economics* 20, 85–116.

Griffin, Dale and Amos Tversky. (1992). "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology* 24, 411–435.

Hanemann, Michael. (1994). "Valuing the Environment Through Contingent Valuation," *Journal of Economic Perspectives* 8, 19–43.

Hastie, Reid, David Schkade, and John Payne. (1999). "Juror Judgments in Civil Cases: Effects of Plaintiff's Request and Plaintiff's Identity on Punitive Damage Awards." *Law and Human Behavior* 23, 445–470.

Hoehn, John and Alan Randall. (1987). "Too Many Proposals Pass the Benefit Cost Test," *American Economic Review* 79, 544–551.

Hsee, Chris. (1996). "The Evaluability Principle: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives," *Organizational Behavior and Human Decision Processes* 67, 247–257.

Jacowitz, Karen and Daniel Kahneman. (1995). "Measures of Anchoring in Estimation Tasks," *Personality and Social Psychology Bulletin* 21, 1161–1166.

Jones-Lee, Michael, Graham Loomes, and P. Philips. (1995). "Valuing the Prevention of Non-Fatal Road Injuries: Contingent Valuation vs. Standard Gambles," *Oxford Economic Papers* 47, 676–695.

Kahneman, Daniel. (1986). "Valuing Environmental Goods: An Assessment of the Contingent Valuation Method." In R. Cummings, D. Brookshire, and W. Schulze (eds.), *Valuing Environmental Goods*: *An Assessment of the Contingent Valuation Method*. Totowa, NJ.

Kahneman, Daniel. (1995). "Extension Neglect and Violations of Monotonicity in Judgment and Preference: Three Examples," Bartlett Lecture to the Experimental Psychology Society (UK).

Kahneman, Daniel, Barbara Fredrickson, Charles Schreiber, and Don Redelmeier. (1993). "When More Pain Is Preferred to Less," *Psychological Science* 4, 401–405.

Kahneman, Daniel and Jack Knetsch. (1992). "Valuing Public Goods: The Purchase of Moral Satisfaction," *Journal of Environmental Economics and Management* 22, 57–70.

Kahneman, Daniel, Jack Knetsch, and Richard Thaler. (1991). "The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, 5, 193–206.

Kahneman, Daniel and Dale Miller. (1986). "Norm Theory: Comparing Reality to Its Alternatives," *Psychological Review* 93, 136–153.

Kahneman, Daniel and Ilana Ritov. (1994). "Determinants of Stated Willingness to Pay for Public Goods: A Study in the Headline Method," *Journal of Risk and Uncertainty* 9, 5–38.

Kahneman, Daniel, David Schkade, Elana Ritov and Cass Sunstein. (1999). "Reversals of judgement; The effect of cross-category comparisons intendedly absolute scales," manuscript order review.

Kahneman, Daniel, David Schkade, and Cass Sunstein. (1998). "Shared Outrage and Erratic Awards: The Psychology of Punitive Damages," *Journal of Risk and Uncertainty* 16, 49–86.

Kahneman, Daniel and Amos Tversky. (1972). "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3, 430–454.

Kahneman, Daniel and Amos Tversky. (1973). "On the Psychology of Prediction." *Psychological Review* 80, 237–254.

Kahneman, Daniel and Amos Tversky. (1979). "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263–291.

Kahneman, Daniel, Peter Wakker, and Rakesh Sarin. (1997). "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics* 112, 375–405.

Kemp, Michael and Christopher Maxwell. (1993). "Exploring a Budget Context for Contingent Valuation." In Hausman (ed.), *Contingent Valuation*: *A Critical Assessment*. Amsterdam: North-Holland.

Koehler, Jonathan. (1996). "The Base-Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges." *Behavioral and Brain Sciences* 19, 1–53.

Levav, Jonathan. (1996). "Questioning Contingent Valuation: Maximality and Violations of Monotonicity in Willingness-to-Pay for Public Goods," Unpublished undergraduate thesis, Princeton University.

Lodge, Milton. (1981). "Magnitude Scaling: Quantitative Measurement of Opinions." In J. Sullivan (ed.), *Quantitative Applications in the Social Sciences*, Vol. 25. Beverly Hills, CA: Sage Publications.

Loomis, John, Armando Gonzalez-Caban, and Robin Gregory. (1994). "Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates?" *Land Economics* 70, 499–506.

McFadden, Daniel. (1994). "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics* 76, 689–708.

McFadden, Daniel. (1999). "Rationality for Economists," *Journal of Risk and Uncertainty* 19, 73–106.

McFadden, Daniel and Gregory Leonard. (1993). "Issues in the Contingent Valuation of Environmental Goods: Methodologies for Data Collection and Analysis." In Hausman (ed.), *Contingent Valuation*: *A Critical Assessment*. Amsterdam: North-Holland.

Mitchell, Robert and Richard Carson. (1989). *Using Surveys to Value Public Goods*. Washington, DC: Resources for the Future.

Neill, Helen. (1995). "The Context for Substitutes in CVM Studies: Some Empirical Observations," *Journal of Environmental Economics and Management* 29, 393–397.

Nickerson, Carol. (1995). "Does Willingness-to-Pay Reflect the Purchase of Moral Satisfaction? A Reconsideration of Kahneman and Knetsch," *Journal of Environmental Economics and Management* 28, 126–133.

NOAA panel report. U.S. Department of Commerce, National Oceanic and Atmospheric Administration. (1993). "Natural Resource Damage Assessments under the Oil Pollution Act of 1990," *Federal Register* 58, 4601–4614.

Novemsky, Nathan and Shirit Kronzon. (1999). "How Are Base-Rates Used, When They Are Used: A Comparison of Bayesian and Additive Models of Base-Rate Use," *Journal of Behavioral Decision Making* 12, 55–69.

Osgood, Charles, George Suci, and Percy Tannenbaum. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.

Payne, John, James Bettman, and Eric Johnson. (1992). "Behavioral Decision Research: A Constructive Processing Perspective." *Annual Review of Psychology* 43, 87–131.

Payne, John W., James R. Bettman, and David A. Schkade. (1999). "Measuring Constructed Preferences: Towards a Building Code." *Journal of Risk and Uncertainty* 19, 243–270.

Payne, John, David Schkade, William Desvousges, and Chris Aultman. (1999). "Valuation of Multiple Environmental Programs: A Psychological Analysis," Unpublished manuscript, Duke University.

Pratto, F. (1994). "Consciousness and Automatic Evaluation." In Paula M. Niedenthal and Shinobu Kitayama (Eds.), *The Heart's Eye*: *Emotional Influences in Perception and Attention*. Academic Press, Inc, San Diego, CA.

Quattrone, George and Amos Tversky. (1984). "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion," *Journal of Personality and Social Psychology* 46, 237–248.

Randall, Alan and John Hoehn. (1996). "Embedding in Market Demand Systems," *Journal of Environmental Economics and Management* 30, 369–380.

Ritov, Ilana, Jonathan Baron, and John Hershey. (1993). "Framing Effects in the Evaluation of Multiple Risk Reduction," *Journal of Risk and Uncertainty* 6, 145–159.

Ritov, Ilana. (1996). "Anchoring in a Simulated Competitive Market Negotiation," *Organizational Behavior and Human Decision Processes* 67, 16–25.

Rosch, Eleanor and Barbara Lloyd. (1978). *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Rottenstreich, Yuval and Amos Tversky. (1997). "Unpacking, Repacking, and Anchoring: Advances in Support Theory," *Psychological Review*, 104, 406–415.

Rutherford, Murray, Jack Knetsch, and Thomas Brown. (1998). "Assessing Environmental Losses: Judgments of Importance and Damage Schedules," *Harvard Environmental Law Review*, 22, 51–101.

Schreiber, Charles and Daniel Kahneman. (2000). "Beyond the Peak and End Hypothesis: Exploring the Relation between Real-Time Pleasure and Retrospective Evaluations," *Journal of Experimental Psychology*.

Seip, Kalle and Jon Strand. (1992). "Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment," *Environmental and Resource Economics* 2, 91–106.

Slovic, Paul. (1995). "The Construction of Preference," *American Psychologist*, 50, 364–371.

Smith, V. Kerry. (1992). "Arbitrary Values, Good Causes, and Premature Verdicts," *Journal of Environmental Economics and Management* 22, 71–89.

Strack, Fritz and Thomas Mussweiler. (1997). "Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility," *Journal of Personality and Social Psychology* 73, 437–446.

Stevens, Stanley S. (1975). *Psychophysics*. *Introduction to Its Perceptual*, *Neural*, *and Social Prospects*. Wiley: NY.

Sunstein, Cass, Daniel Kahneman, and David Schkade. (1998). "Assessing Punitive Damages," *Yale Law Journal* 107, 2071–2153.

Tesser, Abraham and Leonard Martin. (1996). "The Psychology of Evaluation." In E. T. Higgins, and A. Kruglanski (eds.), *Social Psychology*: *Handbook of Basic Principles*, pp. 400–432. New York: Guilford Press.

Thaler, Richard. (1992). *The Winner's Curse*: *Paradoxes and Anomalies of Economic Life*. New York: The Free Press.

Tversky, Amos and Daniel Kahneman. (1971). "Belief in the Law of Small Numbers," *Psychological Bulletin* 76, 105–110.

Tversky, Amos and Daniel Kahneman. (1982). "Evidential Impact of Base Rates." In D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment under Uncertainty*: *Heuristics and Biases*. New York: Cambridge University Press.

Tversky, Amos and Daniel Kahneman. (1983). "Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review* 90, 293–315.

Tversky, Amos and Daniel Kahneman. (1986). "Rational Choice and the Framing of Decisions," *Journal of Business* 59, 251–278.

Tversky, Amos and Richard Thaler. (1992). "Preference Reversals." In R. H. Thaler (ed.), *The Winner's Curse*: *Paradoxes of Economic Life*. New York: The Free Press.

Varey, Carol and Daniel Kahneman. (1992). "Experiences Extended Across Time: Evaluation of Moments and Episodes," *Journal of Behavioral Decision Making* 5, 169–185.

Varian, Hal. (1984). *Microeconomic Analysis*. New York: Norton.

Wilson, Thomas, Christopher Houston, Kathryn Etling, and Nancy Brekke. (1996). "A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents," *Journal of Experimental Psychology*: *General* 125, 387–402.

Zaller, John. (1992). *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.