

Biases in the Quantitative Measurement of Values for Public Decisions

Jonathan Baron
University of Pennsylvania

Measurement of personal values in terms of money or utility can promote efficient public decisions about environmental and risk regulation, health care, and so forth. Current measures are subject to several biases. Quantitative judgments of value are often based on a concept of importance that ignores the quantity of the good being valued. They are sensitive to irrelevant factors, such as cost of the good (vs. its benefit) and whether it has been reduced by human action or nature. Some judgments are based on moral opinions about actions rather than on the value of consequences. Some of these problems seem solvable by methods that remove irrelevant information or force attention to relevant information. Other problems are less tractable. Their solution should be a high priority for research.

Around the world, governments and other institutions are trying to make decisions more rationally. More officials, and more of the citizens they represent, are asking which of the options available to them does the best job of trading off relevant values against each other. For example, should women under Age 50 get yearly mammograms to detect breast cancer, despite the extra cost and risk? Is the reduction of air pollution from vehicle inspection and maintenance worth the extra cost and time?

Many of these decisions require measurement of people's values for specific consequences, such as breast cancer or the respiratory symptoms of air pollution, in a way that allows these consequences to be compared quantitatively with other costs and benefits. Such values can be measured with surveys or interviews, in which respondents are asked to make comparisons from which their values can be inferred. For example, they might be asked how much they are willing to pay to avoid 2 days per year with certain respiratory symptoms. Or they could be asked how the number of symptom days trades off with other things they value, such as commuting time per day.

In this article, I examine several biases discovered recently in attempts to measure values. When I call a measure *biased*, I mean either that responses are insensitive to manipulations that should affect them, given the purposes of measurement, or that the responses are sensitive to what should not affect them.

First, I briefly describe the context in which values are measured and the main methods of measurement. Then I discuss the main biases. Some findings demonstrate that measures are insensitive to the quantity of the good being evaluated. People seem to attend to the "importance" of the type of good but not the quantity at issue. This phenomenon affects judgments of willingness to pay (WTP), judgments of benefit of goods (per

dollar or for the whole good), and judgments of the trade-off between two different dimensions, where judgments are insensitive to the ranges on the dimensions.

Other demonstrations show that most measures are sensitive to factors that should not affect them, given the purpose of measurement. Respondents attended to cost of a good as well as its benefit (when the point was to assess the benefit so that it could be compared with the cost). The monetary value of a good depends too much on the direction of the imagined transaction: Willingness to accept (WTA) is often much higher than WTP. People find it difficult to distinguish means and ends, valuing the means even when the relevant ends are held constant. They take into account the cause of a change in the level of a good, in particular whether it was caused by human intervention or by nature.

In the next section of the article, I review a different tradition concerned with the measurement of values for health states in particular. These methods also suffer from various biases peculiar to the hypothetical trade-offs that respondents are asked to make, but these methods may point to some promising directions for value measurement in general.

I conclude with discussions of two outstanding problems that have received little attention. One of these is the need to ensure that respondents focus on fundamental values, the ones of ultimate importance, rather than on values that represent means to these ends. I make some suggestions about how this might be done. The other problem is that many respondents want to "protect" certain values from trade-offs, regarding them as absolutes. I suggest that this problem may be part of a general unwillingness to distinguish consequences from the actions that produce them.

Importance of Value Measurement

Much of the interest in value measurement comes from cost-benefit analysis. When values can be classified into costs and benefits, we can ask whether the benefits of some project exceed the costs, and we can use the answer to this question as an argument for or against the undertaking of a project. With the increased interest in cost-benefit analysis has come the recogni-

This research was supported by grants from the National Science Foundation. I thank Michael DeKay and Carol Nickerson for helpful comments.

Correspondence concerning this article should be addressed to Jonathan Baron, Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Pennsylvania 19104-6196. Electronic mail may be sent via Internet to baron@psych.upenn.edu.

tion that some benefits and costs are not directly measurable in terms of market prices. It was recognized decades ago that the economic value of a human life should not be measured in terms of life-insurance payments or lost wages, or even the sum of these two, because these measures neglect the value of the life to the person who lives it (Schelling, 1968). Happily, there is no market in which people buy their own lives. Although there are markets in which people buy protection against the risk of death (e.g., air bags), many economists recognized that other methods are needed to estimate the value of such risks—methods in which people are directly questioned about their WTP for safety (e.g., Jones-Lee, 1989).

Environmental controversies provide another source of interest. Environmental regulations often protect natural resources, such as wilderness and ecosystems. Again, there is no market for the protection of endangered species. Yet, one must trade off the value of saving them against other values. The idea of measuring the value of natural resources through questioning of respondents was introduced into U.S. law in the context of legal obligations for polluters, but now it is becoming more relevant to regulatory decisions as well.

Another major application of value measurement for public policy is in medicine, where cost sometimes limits allocation of resources. Because most medical care is paid for by insurance that bundles various services together, no market prices exist for most treatments or for insurance against the need for particular treatments. Accordingly, subjective measurement of values for health states has become more widely used. One common goal is to measure treatment effects in terms of quality-adjusted life years (QALYs). A QALY is the difference between good health and death for 1 year, that is, the benefit of prolongation of a healthy life by 1 year. If the quality of life is only half as good as the state of being in good health, then prolongation of life for 1 year is worth 0.5 QALY. Likewise, to raise a person with this quality to good health for 1 year is also worth 0.5 QALY. Current U.S. medical practices typically balk at treatments that cost more than \$50,000 per QALY. QALYs are usually measured by asking knowledgeable people to make various kinds of judgments about health states (e.g., Torrance, 1986).

Quantitative value measurement is often used to decide what a benefit is worth in money. But I use the term *quantitative* to include measurement of general utility or utility of health states as well. In health policy decisions, value measurement is used more to decide among treatments, given a fixed budget, than to decide on the appropriate size of the budget. In other applications of value measurement, monetary expenditures themselves may be seen as varying in utility, depending on who does the spending.

The potential for value measurement in the formation of public policy is much greater than the examples suggested so far (Portney, 1994). The democratic process, even when voting is supplemented by opinion polls and communication between constituents and representatives, is a crude instrument for the tailoring of government policy to the values of those affected (e.g., Peretz, 1983). Value measurement promises more responsive and efficient government if the methodological difficulties can be overcome. This is largely a problem for the discipline of psychology.

Methods of Value Measurement

I consider three types of methods to measure values: contingent valuation (CV), direct utility measurement, and multiattribute analysis (MA). In the CV method, respondents indicate their WTP for an increase in the good (e.g., their WTP increased taxes or fees in order to reduce the rate of species destruction by a specified amount or their WTA a payment for a reduction in the quantity of the good; Mitchell & Carson, 1989). CV was devised by economists, and it converts all values into monetary values. Its theoretical justification is provided by the economic theory of the consumer, which, in turn, is based on the idea of maximization of a utility function of consumption of various goods (Mitchell & Carson, 1989, ch. 2). CV is used in several countries to assess environmental damage from oil and chemical spills and to evaluate environmental and safety effects of government policies. CV is a matter of contentious debate and litigation (National Oceanographic and Atmospheric Administration [NOAA], 1993; Portney, 1994).

Direct utility measurement refers to a class of methods used mostly in medicine, including ratings, standard gambles, time trade-offs, and person trade-offs. Instead of using money as the standard, these methods define the unit as the difference between life with good health and death for one person (often for 1 year).

In MA, respondents provide trade-off functions for several attributes, one or more of which could be monetary (Keeney, 1992; von Winterfeldt & Edwards, 1986). The ultimate dimension used for decision making is utility rather than money. *Utility* can be defined in various ways, but I take it to be an interval measure of the extent to which goals are achieved. By *interval*, I mean that one can compare utility differences meaningfully. Money itself has utility, but the utility of money need not be a linear function of the amount of money for an individual. Moreover, even the same amount of money may have different utilities in different contexts, depending on one's alternative options for spending it. MA is widely used, but it does not have the same legal force as CV, which has been accepted by the legal field as a standard for assessing environmental damage. MA grew out of several disciplines, including operations research, economics, statistics, and psychology. Its theoretical justification comes most directly from the mathematical psychology of measurement theory (Krantz, Luce, Suppes, & Tversky, 1971). It is related to two other methods covered briefly in this article: conjoint analysis and functional measurement.

MA uses a broader range of approaches than does CV, with several checks for internal consistency. For example, if a certain medical treatment is worth \$50 of the people's money or 2 hr of their time, then they ought to value their time at approximately \$25/hr. Proponents claimed that inconsistent answers to such checks can be eliminated by asking respondents to think about such answers (von Winterfeldt & Edwards, 1986). The assumption is that such inconsistencies result from error or mis-specification of the question rather than any inherent problem in the method.

Many psychologists claimed that values are "constructed" in response to questions and that they do not exist before they are measured (e.g., Slovic, 1995). If this view is correct, then biases in value measurement are to be expected, but an elimina-

tion of obvious biases would not solve the problem of valid measurement. I take the view that it is too early for researchers to give up trying to measure values reliably and validly. I discuss these problems to encourage other researchers to try to solve them, not to promote despair.

The idea of values is that we have some sort of ultimate standards by which we evaluate states of affairs (Baron, 1996). We define our good in terms of these standards. In most cases, we have difficulty applying these standards because we lack information about how to apply them to the objects we are asked to evaluate; for example, we may value the lives of wild animals, but we may have difficulty evaluating the prevention of oil spills because we lack quantitative and qualitative information about how the spills affect the animals. In these cases, we must construct our responses to questions by making guesses about what we are missing or by using other heuristics. But the existence of this sort of construction does not imply that we do not have relevant values or that these values cannot be measured, at least in some ideal world in which we have plenty of time for measurement. In some cases, such as values for our health, the problem of missing information seems manageable in real life.

The concept of value is too important in our view of human good for us to give up trying to measure values prematurely, on the basis of the difficulties I review. At least, we might think of value as something we can approximate in various ways, with different sorts of error. By reducing the error, we can make our approximations more useful. I now turn to a discussion of the problems in detail.

Insensitivity to Amount of the Good

CV judgments are often remarkably insensitive to quantity or scope of the good provided. This insensitivity provides ammunition for critics who say that CV does not measure economic value (e.g., Diamond & Hausman, 1994). Defenders of CV have argued that insensitivity does not necessarily imply faulty measurement (e.g., Hanemann, 1994) or that CV can be made more sensitive if properly conducted (e.g., Schuman, 1995). At issue is the explanation of the insensitivity effects themselves. The source of the effects has implications for whether the problem of insensitivity is serious and, if it is, how to make people more sensitive. Insensitivity is a pervasive phenomenon, not dependent on the details of method. Researchers have identified several types of effects.

Types of Effects

Embedding effect. Kahneman and Knetsch (1992) asked some respondents their WTP for improved disaster preparedness and other respondents their WTP for improved rescue equipment and personnel. The improved equipment and personnel were thus "embedded in" the improved disaster preparedness, so the preparedness included the equipment and personnel and other things as well. The average WTP was, however, about the same for the larger good and the smaller good included in it. When respondents were asked their WTP for the smaller good after they had just been asked about the larger one, they gave much smaller values for the smaller good than the larger one and much smaller values than those given by respondents who were asked

just about the smaller good. Thus, a good seen as embedded in a larger good has a reduced value. M. A. Kemp and Maxwell (1993) replicated this effect, starting with a broad spectrum of public goods and narrowing the good down in several steps, thus obtaining WTPs for an embedded good that were $\frac{1}{300}$ of WTP for the same good in isolation.

Adding-up effect. In a related demonstration, respondents provided WTP values for timber harvest prevention in federally protected wilderness areas. WTP for the prohibition in three areas of about the same size was not much (if any) higher than different respondents' WTP for prohibition in one of the areas alone (Diamond, Hausman, Leonard, & Denning, 1993; following an unpublished paper by Hoehn, Randall, & Tolley, described by Mitchell & Carson, 1989, p. 44). This insensitivity did not result from respondents' thinking that protection of one area was sufficient: When other respondents were asked their WTP to protect one area, assuming that another was already protected, or a third area, assuming that the other two were protected, their WTP values were just as high as those for protection of the first area. In general in this kind of *adding-up effect*, respondents are asked their WTP for Good A (e.g., a single wilderness area); for Good B, assuming that Good A has been provided already; and for Goods A and B together. The WTP for Goods A and B together is much lower than the sum of the WTP for Good A and Good B with Good A provided. This effect is also found within respondents (Baron & Greene, 1996).

Quantity effect. Jones-Lee, Loomes, and Philips (1995) asked respondents to evaluate hypothetical automobile safety devices that would reduce the risk of road injuries. WTP judgments were, on the average, only 20% higher for a risk reduction of 12 injuries for every 100,000 drivers than for a reduction of 4 injuries for every 100,000 drivers. Other results show similar small changes in WTP, even in response to changes of several orders of magnitude in the quantity of the good (e.g., Boyle, Desvousges, Johnson, Dunford, & Hudson, 1994; Carson & Mitchell, 1993, p. 1265). Such results imply that the rate of substitution between money and the good, the dollars per unit, depends strongly on the amount of the good. (If a risk reduction of 12 is worth \$120 and a risk reduction of 4 is worth \$100, then the dollars per unit of risk reduction are 10 and 25, respectively.) This makes it difficult to generalize results to different amounts—a generalization that is nearly always required (Baron, 1995b). Even when such a generalization is not required, such extreme insensitivity over small ranges raises questions about validity of any single estimate.

Does Dichotomous Choice Help?

Could insensitivity result from the difficulty of the task? More recent applications of CV tend to use dichotomous choice rather than free response. In dichotomous choice, the respondents simply indicate whether they are willing to pay some given amount for a given good. They do not have to figure out their maximum WTP. Different respondents respond to different amounts, so the distribution of WTP can be inferred. Of course, this method is much more costly because more respondents are required to estimate a median WTP. When the mean WTP is sought, the cost is even greater because the mean depends more heavily on the tail of the distribution of WTPs.

The use of choice questions was more recently endorsed by a panel convened by the NOAA (1993). This NOAA panel argued for the superiority of choice on the grounds of greater familiarity to respondents, despite its obvious disadvantages in the extra cost of data collection. Baron (1995b) argued that familiarity need not affect validity and that, in any case, the greater familiarity of choice is assumed rather than demonstrated. Moreover, studies of choice versus free response suggest that choice is excessively sensitive to global assessments of the importance of attributes (Tversky, Sattath, & Slovic, 1988; Zakay, 1990). Thus, in choosing between two safety programs that differed in cost and the number of lives saved, respondents chose the one that saved more lives, although many of these respondents were not willing to pay the cost when asked in a free-response format how much they would pay (to make the two programs equivalent overall).

Direct tests of the effect of method on biases do not support the NOAA recommendations. McFadden (1994) found no improvement in sensitivity to quantity from dichotomous choice in a large survey of WTP to preserve wilderness areas. In the free-response choice, respondents' WTP to preserve 57 areas was 57.0% higher than their WTP to preserve 1 area; in the dichotomous choice, their WTP to preserve 57 areas was 33.8% higher than their WTP for 1 area (p. 702). Moreover, in the dichotomous choice, the respondents' WTP to preserve 1 area was actually more than their WTP to preserve 3 areas (including the 1 area).

Fischhoff et al. (1993) claimed to have found that dichotomous choices were more sensitive to quantity. However, their choice condition was one between two programs (to clean up rivers) rather than between no action and a single program at a given price. Baron and Greene (1996; see *Explanations of Insensitivity*) found comparison of two expenditures to induce greater sensitivity to quantity, even when one program's size was specified by the respondent's free response. (Fischhoff et al. also showed that choice was significantly sensitive to quantity, but they did not test the difference between these methods.) In summary, no solid evidence supports the claim that dichotomous choice increases respondents' sensitivity to quantity, and one major study shows no benefit for dichotomous choice.

Explanations of Insensitivity

Five explanations of insensitivity have been proposed. Kahneman and his colleagues considered the first two—contribution and warm glow—to be a single hypothesis, which they called *moral satisfaction*. They have suggested that WTP judgments are insensitive to quantity because the WTPs are expressions of attitudes about the respondents' satisfaction obtained from contributing rather than the amount of benefit. They found that WTP responses are highly correlated with judgments of satisfaction and of importance (Kahneman & Knetsch, 1992; Kahneman, Ritov, Jacowitz, & Grant, 1993) and with the judged importance of the issues and the respondent's feeling of responsibility for them (Guagnano, Dietz, & Stern, 1994). The three other hypotheses I consider are budget constraints, availability, and importance.

Contribution. According to the contribution hypothesis, respondents misunderstand the WTP task as asking about a chari-

table contribution. A contribution increases the size of the good by an amount roughly proportional to the amount contributed, regardless of the size of the good before the contribution is made, so the present size of the good is roughly irrelevant to the effect of the contribution. WTP is limited, mainly because the loss of each additional dollar matters more as more money is paid. By this hypothesis, respondents are more sensitive to quantity if they understand that the size of their contribution does not affect the size of the good provided but is, instead, part of a decision about whether to provide a good of a certain size. To convey this idea, Baron and Greene (1996) told respondents that the size of the good is fixed and that their WTP will be compared with their share of the cost of the good. If more than half of the respondents are willing to pay at least their fair share, then the good will be provided; otherwise, it will not. This manipulation did not increase respondents' sensitivity to quantity. This result argues against a role for the contribution hypothesis.

Warm glow. By this hypothesis, WTP depends on personal agency or participation rather than on the expected consequences of a contribution (Andreoni, 1990; Margolis, 1982). Respondents get a good feeling from the idea of contributing, and the strength of this feeling depends only on type of good and size of the contribution, not on size of the good. This hypothesis predicts that respondents' sensitivity to quantity would increase if they simply provide ratings of the importance of various public goods because questions about ratings do not mention participation in providing the good. Kahneman and Ritov (1994) found that ratings of importance and support for interventions were insensitive to quantity. Jones-Lee et al. (1995) suggested that their demonstration of the quantity effect for auto safety is difficult to explain in terms of moral satisfaction (presumably the warm glow version) because the good is private. These results speak against a role for this hypothesis.

Budget constraint. People may be insensitive to quantity of the good because the monetary value of the good declines as more of the good is available or as more money is spent on it. Such effects may be perceived rather than real. For example, people may think of their spending in terms of separate, limited accounts (Thaler, 1993), including, perhaps, accounts for public goods, so they may overestimate the amount of utility they lose from paying more. The declining perceived effect of size of the good may also occur in several ways. One proposal is that it results from substitution effects (Carson & Mitchell, 1995; Hanemann, 1994). People may think of wilderness areas as substitutable, so it is most important to save one, less important to save the second one, and so on. The term *budget constraint* emphasizes the common idea that the perceived utility of additional payment declines as payment increases (if only because more payment is correlated with more of the good).

This general hypothesis predicts that WTA judgments (willingness to accept payment for loss of a good) would be sensitive to quantity of the good. People have no reason for a budget constraint on what they are willing to accept. WTA judgments could even be oversensitive. For example, if people think that it is most important to preserve one wilderness (as discussed earlier), then they would require more to give up the second of two rather than the first. In fact, WTA judgments show undersensitivity to quantity (Baron & Greene, 1996, Experiments 4–6;

Dubourg, Jones-Lee, & Loomes, 1994, p. 128). The WTA for two goods is thus less than twice the WTA for one good. This result contradicts the budget constraint hypothesis.

Baron and Greene (1996) reported other results that contradict this hypothesis. In Experiment 7, they asked respondents two kinds of questions about risk reduction. One was the standard WTP question (e.g., "What is the largest amount that you would pay per year for a safety device if it prevented a type of accident that caused 1 [or 10] death per year for every 1 million drivers?"). In the other, they gave respondents a monetary amount and asked for an equivalent risk reduction, thus asking for a matching of risk to money rather than money to risk (e.g., "How high would the chance of death from this kind of accident have to be for you to be willing to pay \$500 [or \$50] per year? Answer in number of deaths per year for every 1 million drivers.") The budget constraint hypothesis predicts that the rate of substitution between good and money, the risk reduction per dollar, would be about the same for the two questions. Thus, the good-to-money judgment would show oversensitivity to the amount of money; that is, when the amount of money changed by a factor of 10, the matching risk reductions would differ by a greater multiple. In fact, respondents did not show any oversensitivity, and the rates of substitution were very different between the two tasks.

In Experiment 8, Baron and Greene (1996) asked respondents to give their WTP per unit of the good for a given amount of the good (again, a risk reduction; e.g., "What is the largest amount that you would pay per year for 1 death prevented for every 1 million drivers for a device that prevented a type of accident that caused 10 deaths per year for every 1 million drivers?") In this task, respondents' WTP per unit did not depend on the total amount of the good (10 vs. 1 deaths for every 1 million drivers). The quantity effect disappeared. This also contradicts the budget constraint hypothesis.

The budget constraint hypothesis implies that respondents express their true rates of substitution. The manipulations just described—matching good to money and WTP per unit—affected the respondents' expressed rate of substitution substantially. Notice that the contribution and warm glow hypotheses also imply that these manipulations should have no effect on the rate of substitution between good and money. The effect of these manipulations, then, casts doubt on the contribution and warm glow hypotheses as well as on the budget constraint hypothesis.

Availability. This hypothesis holds that respondents do not think of other goods of the same type as the good evaluated, unless these other goods are explicitly presented. This account is similar to the *part-whole bias* described by Mitchell and Carson (1989). This account implies that insensitivity is a real problem, but it might be cured by a reminder to people about other goods. It cannot explain the quantity effect, but it can explain the embedding and adding-up effects.

Baron and Greene (1996, Experiment 9) manipulated the presence of a full context (reminding the respondent of other goods) and found a small effect on the size of the embedding effect for two elements of a medical insurance package: insurance for transplants and for expensive cancer treatments. With the full context followed by one of the goods and then both goods, the geometric mean was \$163 for WTP for both goods

together and \$165 for the sum of WTP for both goods separately. Without the full context, the respective means were \$155 and \$180. Thus, the embedding effect can be reduced by a reminder to respondents about alternative goods of the same type.

Importance. People seem to have a concept of importance of types of goods that is independent of quantity. They respond to this attribute when asked about WTP, so they are insensitive to quantity information. For example, when people are asked which is more important, life or money, most confidently answer that life is. If they are then asked whether they would spend \$100,000 (borrowing if necessary) to extend a random patient's life by 1 hr, they feel tricked. In a sense, it is meaningless to ask whether life or money is more important, so long as there is some amount of money that is too much to spend for some amount of life, unless the amount of money and life is specified at least implicitly. Yet, people seem not to be bothered by such questions; according to this hypothesis, they even ignore quantity information when it is provided.

The difference between the importance hypothesis and the warm glow hypothesis is that the warm glow one depends on payment made as well as on type of good. WTP judgments based on importance result from a kind of matching of monetary amounts to importance levels, without reference to quantity of either the good or the contribution. These judgments do not even depend on a role for the respondent, although the warm glow hypothesis concerns exactly this role. The importance hypothesis predicts that people will respond similarly about their WTP per unit of a 10-unit good and their WTP for the entire good or for 1 unit alone (as Baron & Greene, 1996, found), whereas the warm glow hypothesis predicts that they would pay less per unit for 10 units than for 1 unit because their total payment will depend on the number of units and a sufficient warm glow might be obtainable from a relatively small payment.

The matching of responses to felt importance may have little to do even with money as such. Slovic, Lichtenstein, and Fischhoff (1979) asked respondents to express the seriousness of various types of death (e.g., caused by smoking vs. caused by alcoholism) by stating a ratio of each type of death to a standard monetary unit of \$10,000 or \$1 million (for different respondents). The ratio depended on the type of death, but it did not depend on the monetary unit. Respondents seemed to match the numbers to the types of death, without considering the types' translation into dollars.

Baron and Greene (1996) found other support for the importance hypothesis, which they called the *prominence hypothesis*, by showing that respondents became more sensitive to quantity when it was harder for them to ignore. In one effective manipulation, respondents were asked about cuts in government programs. They were asked to specify both the size of the cut in terms of a percentage (within a range defined separately for large and small cuts) and the amount of increased taxes they would pay to prevent the cut (WTP). Their sensitivity, shown by the effect of cut size on WTP, was greater when respondents specified the percentage themselves than when it was presented to them. The former condition made the numerical value more salient, presumably. Respondents also became more sensitive to the quantity of goods when they were given the percentage of one cut and asked for the percentage of another cut that was equally harmful. This response mode was presumably more

compatible with the stimulus (because both were sizes of cuts), and this too made respondents more attentive to the size of the cut (Tversky et al., 1988).

In summary, the importance hypothesis is consistent with most results that show insensitivity to quantity (the exception is effect of provision of full context, which supports availability). Responses are governed largely by the perceived importance of issues, which is independent of quantity.

The importance hypothesis has two practical implications. First, respondents must be induced to attend to quantity of the good. Perhaps the simplest way is to ask respondents for an entire function relating their WTP (or some other attribute) to the quantity of the good. This is standard practice in MA. Second, it is necessary to check that respondents have taken quantity into account adequately. One way is to reverse the judgment task, for example, to give an amount of money and ask for an equivalent amount of the good. Another is to ask for trade-off judgments between two goods and between each good and money and then to check that the three functions are consistent. If the checks fail, then the respondents must be asked to repeat their judgments. These methods, too, are commonly used in MA.

Unit Pricing and Number of Units

Another type of study supports the hypothesis that respondents attend to importance of a type of good rather than its quantity. The response measure is a judgment of utility or benefit rather than a judgment of a monetary equivalent. Respondents compare goods as wholes or utilities of the amounts of goods provided by a constant amount of money (e.g., per dollar). S. Kemp and Willetts (1995) asked respondents to rate the value of government services in New Zealand, including items that varied in the money spent on them (e.g., government retirement income = NZ\$4,314 million, universities = NZ\$577 million, and the New Zealand Symphony Orchestra = NZ\$8 million). Respondents rated "how useful or worthwhile" each item was to New Zealand or "how much value it would be to New Zealand" (p. 5). One group of respondents rated the total utility of each service and the utility of a 5% increase in spending (marginal utility of percentages). The other group rated the utility per dollar spent and the utility of each extra dollar spent on the program (marginal utility of dollars). Respondents were not told the actual cost of each service, but it was reasonable for them to think that the amounts differed substantially.

We might expect that total utility would correlate with cost, whereas per-dollar utility would not. (Ideally, based on economic theory, per-dollar utility should be equal for all services.) In fact, both total and per-dollar ratings correlated moderately across services with $\log(\text{cost})$, $r = .55$ in both cases. (Correlations are based on mean ratings for each service across respondents.) Respondents did not distinguish between total utility and utility per dollar.

Marginal-utility-of-dollars ratings should be uncorrelated with cost or with any other ratings, although one might expect positive correlations between cost and marginal-utility-of-percentages ratings. The correlation of marginal utility of dollars with cost was .38, and the correlation of marginal utility of percentages was .41. Marginal ratings also correlated over .90

with total ratings in both conditions, and they correlated .99 with each other (across programs).

In summary, respondents did not distinguish among different kinds of questions. Their ratings make more sense if one thinks of them as ratings of total utility (or perhaps of utility of the first dollar spent) because the ratings were correlated with cost. This may correspond to importance in the sense discussed in *Explanations of Insensitivity*. This kind of importance judgment is insensitive to quantity of the good because total amount of the good is different from amount purchased with a fixed amount of money.

Range Insensitivity in Weights for Multiattribute Analysis

Insensitivity to quantity is also found in MA, in the form of insensitivity to ranges when weights are assigned to attributes. The importance hypothesis can also explain this effect. This bias is typically found when assessment of utilities is divided into two parts, assessment of utility functions on each dimension and then assignment of relative weights to the dimensions. To assess utility on a single dimension, researchers can use several methods, such as those described in *Conflict Among Methods for Valuing Health States*. In some cases, utility may be assigned on the basis of some reasonable assumption (e.g., if one of the dimensions of a decision about a contraceptive method is "probability of pregnancy," then it is reasonable to assume that the utility is a linear function of probability because expected utility is a linear function of probability).

After utilities are assigned on a single dimension, several methods are commonly used to assign relative weights to the different dimensions (Edwards & Barron, 1994; von Winterfeldt & Edwards, 1986). The idea is that each dimension is initially scaled from 0 to 1, but then the utility of each dimension must be multiplied by a weight, so after multiplication the utility units are comparable. For example, if one dimension is the probability of pregnancy, ranging from 0% to 100%, and another is cost, ranging from \$0 to \$1,000/year, suppose a person feels that the range of the difference for pregnancy is greater than the range for cost. The person might feel that the difference between 0% and 20% chance of pregnancy was equivalent to the difference between \$0 and \$1,000 in cost. In this case, the cost dimension should be weighted 0.2 if the pregnancy dimension is weighted 1.0, on the assumption that the utility function is linear with probability of pregnancy. Notice that the weights depend on the ranges. The weight of cost, relative to pregnancy, would be larger if the cost ranged from \$0 to \$2,000, approximately twice as large.

People make judgments about the relative importance of dimensions even without knowing the ranges. For example, people say that life is more important than money in public decisions about transportation safety, without knowing what amount of money or how many lives. When asked to choose between two proposals described in terms of cost and lives saved, respondents chose the one that was better on the important dimension (lives saved), despite its higher cost. Similar respondents were given the same two programs but without the cost of the one that saved more lives. They were asked to assign a cost to this program that would make the two programs equally attractive.

The cost they assigned was lower than the cost given to the first group of respondents, thus implying that, if they had been given the cost received by the other group of respondents, they would have chosen the program that saved fewer lives (Tversky et al., 1988). Both the choices and judgments of importance, then, seemed to be influenced by a concept of importance, which was somewhat unrelated to the explicit trade-offs made by the second group of respondents.

When people assign weights, they often use this concept of importance rather than pay attention to the ranges. Keeney (1992) called underattention to range "the most common critical mistake" (p. 147). Unless a researcher is very careful, respondents can easily fall into this mistake, even when the ranges are called to their attention. As a result, respondents in experiments on weight assignment are often undersensitive to the range (Weber & Borchering, 1993). Doubling the range of money, for example, should approximately double its relative importance, but this rarely happens.

Weber and Borchering (1993) pointed out that this insensitivity to ranges is related to another bias in weight assessment. If one attribute is split into two parts, respondents assign more weight to the total. For example, if risk of sexually transmitted diseases is split into risk of AIDS and risk of other diseases, then the combined weight increases. Again, respondents may have assigned weights to each part based on the type of risk (risk of disease) rather than the quantity of it, which was less when only a subset of diseases was considered.

Weber and Borchering (1993) also suggested that all of these effects may result from respondents making an insufficient differentiation among dimensions. People are reluctant to say that the importance of one dimension is only a tiny fraction of the importance of another. They tend to assign high weights to less important dimensions. This hypothesis does not explain all the results (e.g., those of Tversky et al., 1988) in which judgments of importance made in the absence of quantity information were related to choices. The latter result supports a role for perceived importance in some tasks, in addition to this non-differentiation of dimensions.

This inadequate differentiation might be understood as an example of underadjustment following anchoring. The task is to compare two intervals to establish what proportion the smaller interval is of the larger (or vice versa). Respondents may anchor on a judgment that the two intervals are equal in utility (i.e., the smaller interval is 100% of the larger). They may then adjust this figure downward, insufficiently.

Underattention to range can be reduced. Fischer (1995) found complete undersensitivity to range when respondents were asked simply to assign weights to ranges (e.g., to the difference between a starting salary of \$25,000 and \$35,000 and between 5 and 25 vacation days or between 10 and 20 vacation days for a job). When the range of vacation days doubled, the judged importance of the full range of days (10 vs. 20) relative to the range of salaries (\$10,000) did not increase. Thus, respondents showed inconsistent rates of substitution, depending on the range considered. Respondents were more sensitive to the range, with their weights coming closer to the required doubling with a doubling of the range, when they used either direct trade-offs or swing weights. In a direct trade-off, the respondent changed one value of the more important dimension so that the two

dimensions were equal in importance, for example, by lowering the top salary of the salary dimension. (Weights then must be inferred by either measuring or assuming a utility function on each dimension.) In the swing weight method, respondents judged the ratio between the less important and more important ranges; that is, they replied that the difference between 5 and 25 vacation days is $\frac{1}{5}$ of the difference between \$25,000 and \$35,000.

In the direct trade-off method, the range is given for one dimension only. The task is thus analogous to a free-response CV judgment, so we might still expect—and Fischer (1995) still found—some insensitivity. Baron and Greene (1996) found that this insensitivity could be reduced still further by giving no specific ranges for either dimension. Respondents were asked to produce two intervals, one on one dimension and one on the other, that were equally large in utility. Of course, to test sensitivity to quantity, Baron and Greene had to make sure that respondents actually produced different quantities, so they gave respondents suggestions about the ranges. It seems plausible that these suggestions created some residual insensitivity to range; if they had been absent, respondents might have been completely sensitive. Hence, in principle, it is likely that the problem of range insensitivity can be solved.

In general, one might avoid anchoring and range effects of many sorts by not presenting any tempting anchors. If respondents then anchor on anything, it might be on their initial answer to the question, which would be sensitive to quantity if the respondents attended to quantity. Empirical tests are still required to determine whether judgments are consistent, such as tests of range sensitivity (when one range is given) or tests of transitivity of relative weights among three dimensions (in which the trade-off between two dimensions is predicted from the trade-offs of each with a third).

Range Insensitivity in Holistic Judgments

Other techniques for the assignment of weights include researchers asking respondents simply to evaluate multidimensional stimuli holistically with ratings or rankings. For example, each respondent rates many antipollution policies that differ in yearly deaths prevented and in yearly cost per driver for inspections. Utilities on each dimension and relative weights of dimensions are inferred from these responses. A distinction between utility on the dimension and weight of the dimension is typically not even needed. It is simply assumed that respondents will answer in terms of total utility, so the utility measures inferred for each dimension can be assumed to use the same utility scale across all the dimensions. A great variety of methods use this approach. The two most common ones are functional measurement (e.g., Anderson & Zalinski, 1988) and conjoint analysis (P. E. Green & Wind, 1973; P. E. Green & Srinivasan, 1990; Louviere, 1988).

In the method of holistic judgments, the trade-off between a given change on one dimension and a given change on the other should be unaffected by the range of either dimension used within the experimental session. If a change from 50 to 100 lives is worth a change from \$20 to \$40, then this should be unaffected, regardless of whether the dollar range is from \$20 to \$40 or from \$2 to \$40. Beattie and Baron (1991), using

such a functional measurement task, found no effects of relative ranges on rates of substitution, using several pairs of dimensions, so long as (arguably) the dimensions were understandably related to fundamental values, the values that really mattered (e.g., effect on the winning of a basketball game vs. injury rate of the players and risk of endometrial cancer vs. risk of osteoporosis for women considering estrogen replacement). They found range effects only when the dimensions were presented as arbitrary numbers (e.g., a score on a test) so that the range itself provided information about the numbers' relation to the values of fundamental interest (e.g., ability). Mellers and Cooke (1994) found range effects even under conditions like those in which Beattie and Baron found none. The source of the discrepancy is unclear, although one possibility is that Beattie and Baron presented two attributes at a time, whereas Mellers and Cooke presented three. The use of three attributes may encourage the use of heuristics, such as the ignoring of less important dimensions, and perceived importance could be affected by range. Still, the results of Beattie and Baron suggest yet another way to avoid biases in the comparison of dimensions.

In a few studies, researchers have compared weights derived from holistic ratings with those derived from pairwise comparisons of attributes (Hoffman, 1960, Figures 3–7; von Winterfeldt & Edwards, 1986, p. 365). Typically, weights derived from holistic judgments are more variable across attributes than are those derived from direct comparisons; the ratio of the higher weights to the lower weights is greater for holistic judgments. One explanation of these results is that people tend to ignore the less important dimensions in holistic judgments (Shepard, 1964), which is consistent with the results of Tversky et al. (1988) who found that choice (holistic) responses were more sensitive than matching responses (pairwise comparisons of attributes) to perceived importance. Alternatively, the matching responses may be an example of insufficient differentiation of dimensions, as discussed in *Range Insensitivity in Weights for Multiattribute Analysis*. In that case, the holistic ratings might be better.

Sensitivity to Irrelevant Factors

Although judgments of value are often insensitive to quantity—a relevant factor—they are also affected by irrelevant factors. Most research on these factors is on CV.

Cost and Fair Price

CV researchers typically assume that the value of the benefits is a function of the effects of the good in question, not its cost. Although cost is usually a good guide to value, excessive attention to cost as a guide can cause trouble. For example, suppose that government Program X has a little more benefit than Program Y, so people's WTP for Program X would be a little higher if the costs of the two programs were the same. But suppose that respondents express higher WTP for Program Y than for Program X if they learned that Program Y was more costly for the government to implement. Then, if the government used WTP as an index of preference and cost was not a major issue, the government would assume that people preferred Program Y. A government that acted on this information would

choose Program Y over Program X. The citizens would pay more and get less benefit.

Reported WTP increases with cost of the good, even when benefit is constant. In one study, WTP for a hypothetical bottle of beer, to be consumed on the beach, depended on the source of the beer (Thaler, 1985). WTP was higher if the beer came from a fancy resort hotel than from a mom-and-pop store. Because the beer was to be consumed on the beach, none of the atmosphere of the hotel would be "consumed." Baron and Maxwell (1996) asked WTP questions concerning hypothetical public goods. They gave respondents information about the benefits and costs to provide various forms of public risk reduction (e.g., removal of chemicals from drinking water). They found that respondents were influenced by both benefits and costs. Respondents were willing to pay more for goods that were more expensive when benefit was held constant. Cost information affected WTP when it took the form of estimated cost or was simply implied by past expenditures or by descriptions of how a good would be provided. Cost affected WTP when each respondent judged two cases that differed only in cost and when different costs were presented to different respondents. The results can be understood as an overextension of a somewhat useful heuristic: Things that cost more often yield more benefit. The findings suggest that CV methods may be improved by the elimination of information from which costs could be inferred, so respondents can focus more easily on benefits alone.

Other evidence indicates that WTP is affected in part by judgments of what a "fair price" should be (Kahneman, Knetsch, & Thaler, 1986). In general, people like to pay what something is worth. They do not want to be taken advantage of by paying much more than the cost, and they are reluctant to take advantage of excessively low prices (Winer, 1986).

Other evidence points to a role of fair prices in WTP judgments for public goods. When respondents provided justifications for their WTP responses, they often referred to the cost of providing the good or the cost per household (Schkade & Payne, 1994, p. 99). D. P. Green, Kahneman, and Kunreuther (1994) asked respondents about their WTP donations for a program to teach English to immigrants. Respondents who were reminded that 20 million other households would be asked the same question differed from respondents who did not receive the reminder in two ways. First, they thought that it was less appropriate not to contribute at all. Second, they thought that the appropriate contribution was lower. Apparently, the reminder of the other contributors made them feel obliged to do their share, but it also may have made them wary of contributing more than was needed to cover the cost of the program. In this way, they reported something about how much they thought their share of such a program would cost.

Such results can be understood in terms of a *disutility for unfairness*, the exploitation of the provider of a good by too little or too much payment. Inefficiency can still result, however, even if one considers this disutility; that is, people can pay more for what they want less. For example, suppose that you have a slight preference for Heineken over Miller beer. Your friend asks you your WTP for Miller from the resort hotel and your WTP for Heineken from the mom-and-pop store. You say \$5.00 and \$4.00, respectively, because you think that the fair prices are \$5.01 and \$3.99, respectively, and you are somewhat affected

by your taste. Moreover, your disutility for unfairness depends on the absolute value of the departure from the fair price. If the prices turn out to be \$4.75 and \$4.25, respectively, then your friend would get you the Miller for \$4.75. You pay more for what you like less, and the degree of unfairness (\$0.26) would be the same with either transaction, so you are not compensated for your loss by engaging in a fairer transaction. The problem is that anyone using a reasonable decision procedure based on WTP responses must assume that lower prices are always better and that higher WTPs represent more utility from the good itself.

Many CV studies provide considerable information about how the public good is provided; this information might lead respondents to focus on the cost of provision rather than the extent to which their own utility is affected. Inefficient decisions could thus arise.

WTA Is Larger Than WTP

A general result in CV is that WTA is larger than WTP (Mitchell & Carson, 1989). The size of this effect varies. Sometimes it is not significant (Baron & Greene, 1996, Experiment 4); at other times, WTA is several orders of magnitude greater than WTP (Thaler, 1985). The mere existence of such an effect need not be a bias because the utility for each additional dollar (or each additional unit of the good) declines with the amount of money (or good) that one has; thus, part of the WTA–WTP difference could be that the loss of \$X has more disutility than the gain of \$X has utility and likewise for the good involved. The discrepancy between WTP and WTA is found, however, even when these effects are controlled, so this effect might be irrelevant to true value. Kahneman, Knetsch, and Thaler (1990) found that “choosing” between a good and money led to lower evaluations of the good than selling a good for money—the only difference was whether the status quo was to have the good or not. Baron (1992) and Irwin (1994) found a parallel difference between WTP and willingness to forgo a loss of money by the giving up of a hypothetical good. Again, the only difference was the status quo (not even any experience of the good).

The usual explanation of the WTP–WTA discrepancy is that losses loom larger than gains (Kahneman & Tversky, 1979). But the variability of the effect from case to case, although not contradicting this account, suggests that it is not a raw phenomenon but one that can be further understood and manipulated.

Several researchers have succeeded in manipulating the effect. Irwin (1994) found that this kind of effect was greater for environmental goods than for market goods; she presented evidence that the difference is the result of people’s moral concerns about making the environment worse, although she did not manipulate moral factors while holding the good constant. Marshall, Knetsch, and Sinden (1986) found that WTA versus WTP effects do not occur for advisers (vs. decision makers). A possible explanation of this result is that the norm that prohibits the accepting of money in return for losses (even personal losses) might be “agent relative,” that is, dependent on the role of the decision maker rather than on the outcome alone. I discuss similar factors in *Protected Values*.

Failure to Warn

DeKay and McClelland (1993) studied dichotomous decisions about safety, such as a warning issued to nearby residents about the possibility that a dam would break. They asked respondents to rate the desirability of four possible outcomes (with the dam scenario as an example): warning issued and dam broke, warning not issued and dam broke, warning and no break, and no warning and no break. Many respondents rated the outcome of the false alarm (warning, evacuation, and no break) as better than the outcome of the correct rejection (no warning, no evacuation, and no break). In the scenario in question, needless evacuation in response to a false alarm could lead to mistrust of future warnings. Yet, respondents thought that a warning was the better decision in this case and rated the outcomes of the false alarms as better than the correct rejection. Their valuation of the outcome was contaminated by their valuation of the decision that led to it. Such contamination could, in principle, be reduced either by asking the respondents to evaluate the decision and the outcome separately (thus giving them a chance to express their opinion about the decision) or by removing the description of the decision context (as DeKay & McClelland, 1993, found).

Human Versus Nature

Studies of environmental values (e.g., Kempton, Boser, & Harley, 1995) indicate that much concern about the environment results from a belief that people have no right to interfere with it. People are willing to pay more to prevent damage caused by humans than the same damage caused by nature (Kahneman & Ritov, 1994; Kahneman et al., 1993). For example, they are willing to pay more to save sea birds from oil spills than from a new epidemic disease. Likewise, people place a higher priority on saving endangered species threatened by humans than by nature (DeKay & McClelland, 1996). These results are evidence against the validity of CV for measurement of the value of nature itself. The purpose of valuation for decision making is to separate the consequences from the acts that partially determine them.

In summary, judgments are affected by several irrelevant factors, many of which involve the means rather than the ends (e.g., the means by which animals are killed or whether an outcome resulted from a good decision or a bad one). The means are irrelevant because they are often not the same as those that would be used in applications. That is, the values measured are supposed to be transportable across different means. Later I discuss possible solutions to this problem.

Conflict Among Methods for Valuing Health States

Researchers have studied several methods of value measurement, largely in the context of evaluation of health states or conditions, such as heart disease, sensory difficulties, and so forth. These methods are also biased by contamination from irrelevant influences, such as the use of various heuristics connected with decision making. Quantity sensitivity is not a problem because the conditions compared are things that could happen to a person. In general, studies of these methods may be

able to teach researchers useful things that they could apply more generally.

Health utilities are assessed on a scale of 0 to 1, where 0 = *death* and 1 = *good health*. Ideally, the condition is described in some detail. If possible, respondents already familiar with the condition, either as patients or care providers, make the evaluations. Utilities assessed in this way are used to measure the efficacy of medical interventions. For example, if a condition has a utility of .6, then curing that condition (completely) has a utility of .4. Utilities are typically multiplied by the duration of the condition (or cure), so utility is integrated over time. It is natural to think of utility in terms of QALYs, where 1.0 QALY is the utility resulting from the prevention of a person in good health from dying for 1 year (Kamlet, 1992; Pliskin, Shepard, & Weinstein, 1980).

Several methods are used for these ratings (Froberg & Kane, 1989; Nord, 1992; Torrance, 1986). The main ones are standard gamble, analog scale, time trade-off, and person trade-off. The standard gamble method asks for a probability that makes the respondent indifferent between the condition of interest and a gamble with a P probability of good health and a $1 - P$ probability of death; P is thus the utility of the condition on a scale in which 0 = *utility of death* and 1 = *utility of good health*, assuming that the respondent equates the expected utility of the two options. The analog scale method simply asks for a rating of the condition on a scale from 0 = *death* to 100 = *good health*. The time trade-off method asks respondents for Age A, such that they would be indifferent between living to Age A in good health and living to Age B (greater than Age A) with a condition. The person trade-off method is typically used to compare treatments of different conditions; it asks for a number N such that respondents are indifferent between curing, for example, 10 patients with the more severe condition and N patients with the less severe condition. An answer of 100 implies that the utility of the severe condition is 10 times as far from 1 as that of the less severe condition.

Several researchers have compared these methods (Dolan, Jones-Lee, & Loomes, 1995; Froberg & Kane, 1989; Jones-Lee et al., 1995; Nord, 1992; Ubel, Loewenstein, Scanlon, & Kamlet, 1996). In general, the standard gamble and time trade-off methods tend to minimize differences among the conditions other than death, putting them all very far from death relative to the analog scale. Loomes (1994) compared standard gambles with CV to avoid death and other conditions, finding again that standard gambles tend to put conditions closer to good health and CV tends to put them closer to death. I cannot tell from this inconsistency which method is more accurate. The analog scale yields the highest correlation with other measures of health (Bosch & Hunink, 1996), but this may be because it is easier to use. Distortion may still be present.

All of these methods (at least under some circumstances) may be susceptible to the bias found by Weber and Borcherding (1993), the tendency to assign numbers too close to each other when ranges are compared. Possibly this results from anchoring on equality of weights. In the person trade-off method, equality has the direct interpretation of giving everyone equal priority regardless of the seriousness of his or her condition, the expected benefit, or the cost of treatment (Nord, Richardson, Kuhse, & Singer, 1995). Of course, anchoring on equality cannot explain

the extreme differences among conditions in the person trade-off method found by Ubel et al. (1996). I return to this result later.

The standard gamble method is subject to the certainty effect, in which respondents overweight outcomes with a probability of 1 (Kahneman & Tversky, 1979). This effect might be understood in terms of respondents using the certain option as a reference point, thus viewing the risky option as leading to potential loss, which is then weighted more heavily than the potential gain (Hershey & Schoemaker, 1980, 1986; Loomes, 1993, 1994). But distortions of probability do not seem to be limited to 100% versus everything else (de Neufville & Delquie, 1988; Kahneman & Tversky, 1979). Nor can these distortions be eliminated by a monotonic transformation of the utility function elicited with gambles (Richardson, 1994). Standard gambles are thus incapable of yielding consistent utility estimates across different probabilities. In particular, the utilities of health states vary as a function of the probabilities used in the gambles, so utilities elicited with one probability cannot be assumed to apply to decisions with other probabilities for the outcomes.

The time trade-off method is affected by subjective discounting of events in the distant future; although discount rates can be measured and corrections can be made, most users of this method do not attempt this. Perhaps they are right not to make this correction because many respondents do not seem to answer the question as asked; rather they answer as if they were being asked about a ratio (Stalmeier & Bezembinder, 1996). Discounting itself is not properly classified as a bias because it is arguably normative, unless precautions are taken to induce the respondent to care equally about all future times (e.g., by instructing the respondent that the decision is being made for someone else). Another problem is that some respondents ignore the severity of conditions for durations of a few months, giving duration of survival absolute priority (Miyamoto & Eraker, 1988).

The person trade-off method asks for judgments that are directly relevant to the kind of rationing decisions that result from cost-utility analysis. But respondents seem to use various heuristics connected with their view of fairness, and this view is not generally utilitarian (Baron, 1994, 1995a). One heuristic is that patients with equally serious conditions should have equal priority, regardless of the expected benefit of treatment or its cost. Nord (1992) found that, when people make person trade-off decisions from the point of view of a patient who is choosing hospitals on the basis of their policies, they favored policies that treated the sicker patients first, regardless of the degree of improvement expected from treatment. (Patients who are very sick at the outset may still be sick after treatment but less so.) When people made decisions from the point of view of a policy maker, however, they took improvement into account. It seems that the policy maker's viewpoint encouraged more utilitarian reasoning.

Another heuristic is that people with more serious conditions should be given priority, almost without regard to cost or the number that could be treated within a given budget constraint. Use of this heuristic could explain the intransitivity found by Ubel et al. (1996): for three conditions—A, B, and C—in which A is worse than B and B is worse than C, the C:A ratio in number of people treated was less than the product of B:A

and C:B. For example, if respondents said that treating 10 Bs was equivalent to 1 A and that treating 10 Cs was equivalent to 1 B, they said that the number of Cs equivalent to 1 A is less than 100. Judgments thus cannot be interpreted as consistent estimates of utility. If they were consistent, respondents would say that 100 Cs are equivalent to 1 A. Instead, respondents may base their judgments on an estimate of what *high priority* means when converted into numbers (e.g., 10:1), and this estimate may depend little on the perceived seriousness of the conditions involved.

The major biases found with standard gambles, time trade-offs, and person trade-offs seem to result from principles for decision making itself. Methods that rely on direct judgments rather than hypothetical decisions might avoid these biases. In difference measurement, respondents directly equate differences between pairs of outcomes or find an outcome midway between two others. Consistency is checked by researchers looking for *monotonicity*: If AB (the difference between A and B) = DE and $BC = EF$, then $AC = DF$ (Krantz et al., 1971, ch. 4). One way to implement difference measurement is to find midpoints: On a scale between good health and death, for example, what condition is equidistant from both? What other condition is equidistant from *that* condition and good health? and so forth. Barron, von Winterfeldt, and Fischer (1984) and Krzysztofowicz (1983) used this method. Guiot and Lefoll (1988) found that such judgments were consistent with analog-scale judgments for utility of monetary income. Difference measurement may be seen as a modification of analog scaling in which the significance of intervals on the scale is given a precise interpretation.

The major limitation of difference measurement is that it requires a large number of conditions. It also may be subject to various context effects that result from respondents' tendency to spread out judgments evenly (Poulton, 1979). This is typically found with the analog-scale method. In principle, however, this problem can be avoided (with sufficient time or numbers of respondents) by the researcher checking initial judgments of midpoint bisection and by presenting only the two ends and the purported midpoint for a second judgment, after a delay or with different respondents.

Another problem with difference measurement is that judgments may depend on how the question is posed. For example, Varey and Kahneman (1992) told respondents that Person A must carry a 30-pound (13.61 kg) suitcase for 200 yards (182.88 m), B must carry it for 550 yards (502.92 m), and C must carry it for 900 yards (822.96 m). When respondents were asked whether B's overall discomfort was closer to A's or C's, most respondents thought it was closer to C's. When the same respondents were asked whether an individual would suffer greater discomfort while walking the interval from 200 to 550 yards (182.88 to 502.92 m) or the one from 550 to 900 yards (502.92 to 822.96 m), most thought that the second interval was worse. It seems that, in this case, the second way induced respondents to think about the experience, but the first way evoked a general heuristic that people generally adapt to experiences. This sort of problem is doubtless not peculiar to difference measurement.

The method of bisection can be seen as an addition to the analog-scale method, a way to clarify to respondents the idea

that intervals are supposed to be comparable. Possibly after respondents understand this, they can proceed to make analog-scale judgments, without the need for the bisection of intervals. Such a mixed procedure would reduce the need for a large number of conditions. Respondents must still be encouraged to attend to outcomes rather than to rely on general heuristics.

Another approach to value elicitation would be to ask for direct judgments of ratios, such as in the swing-weight method of measuring weights of attributes. For example, how many times worse is being blind and deaf than just being blind? In principle, this is the kind of judgment that the person and time trade-off approaches are supposed to tap, but direct questions about ratios may avoid heuristics connected with the decision itself, such as those concerned with fairness and with discounting of the future.

Both difference-measure and direct-ratio judgments are promising ways of avoiding some of the problems of value measurement discussed earlier. (More research is still needed to "get the bugs out," though.) For example, to avoid problems of quantity insensitivity, respondents could carry out difference measurement with several items at once by ranking the items, then finding the midpoint, and so on. Items in the set could vary in quantity as well as in type. Difference measurement is typically used when the items vary in quantity, so sensitivity might be found. If some of the items in the set already had monetary values assigned, this method could replace CV for some purposes.

I now turn to two final problems of value measurement: the problem of making sure that fundamental values are brought to bear on judgments and the problem of some respondents being unwilling to think in terms of trade-offs.

Types of Values

Fundamental Values

An important distinction made in both philosophy (e.g., Brandt, 1979) and MA (particularly by Keeney, 1992) is that between fundamental values and proxy values. Values are the criteria that we use to evaluate outcomes. *Fundamental* values express our most important concerns. *Proxy* values are related to fundamental values through beliefs about the extent to which our satisfying the former will satisfy the latter. Most of the values of interest to policy makers are proxy values, at least in part. If people's beliefs about the relation between proxy values and the underlying fundamental values are incorrect, then their expressions of value are invalid indicators of their real concerns (Baron, 1995b, 1996). If people are unsure of the relation between what they are asked about and their fundamental values, they may have difficulty answering questions with reference to fundamental values; they may then either make guesses about the missing facts or else use various heuristics to answer the questions.

For example, suppose I am asked to evaluate the saving of 10 sea otters from death in an oil spill. This event would be a means to other things that I care about: the pain that the sea otters would experience; the shortening of their lives and the consequent loss of their pleasures of existence, such as they are for sea otters; the effects on prey and predators of sea otters;

and the effect (together with other events) on the stability of the ecosystem, which in turn is a means to the achievement of many other goals. These are some of the important values I have that make me care about sea otters. Yet, I have *almost no idea* how any of these values are affected by the death of 10 sea otters in an oil spill. For example, I do not know how the sea otters would die and how much worse this death is than any other type of death of sea otters. Perhaps it is no worse at all. Perhaps they usually get eaten by sharks. A serious attempt to measure my values for sea otter deaths would tell me some of these things I need to know. Of course, people must make decisions even in the absence of complete information, so it is unreasonable for me to demand that I get all the answers, even the ones that nobody knows. But it is not unreasonable to demand that I get some expert opinions about the major issues.

The usual way to look at CV and other methods of valuation bypasses questions of this sort. Economists are happy with valuation methods if they predict real economic decisions. By this view, it does not matter whether responses reflect fundamental values. Yet, even consumer decisions can ignore these values; they are often made on the basis of easily correctable false beliefs, such as beliefs about the risks of eating apples that contain small amounts of Alar (a ripening agent that may cause cancer).

The failure to make respondents consult their fundamental values seriously may be the largest source of error in valuation of natural resources, as practiced. It is these values that give meaning to the whole enterprise. They are precisely the reasons why people care about sea otters or anything else. We do not need to measure values just to have a way of making decisions. We already have legislatures, courts, and regulatory officials for that. If methods of valuation have any special status compared with these other institutions, it would seem to have something to do with their ability to measure fundamental values.

The problem may be solvable by providing respondents with summaries of expert opinion. DeKay and McClelland (1996) provided respondents with summaries of various dimensions of ecological importance of endangered species. They found a shift away from more superficial attributes, such as similarity to humans—which, they argued, were being used as heuristics in the absence of information about ecological effects, which was their main concern.

Protected Values

Theories of rational decision making in general require trade-offs among values, including moral values. People value human life, and they value the other things that money can buy. They rationally spend some amount of money on safety each year but not an infinite amount. Some risks are just too small and too costly to reduce. Although this limitation is a fact of life, it is not one that everyone is happy with. It goes against some people's values, which hold that human lives—or human rights or natural resources—are infinitely more important than other economic goods. These people hold protected values (Baron & Spranca, in press). Some of their values, as they conceive them, are protected against being traded off for certain other values. People who hold such values may behaviorally trade them off for other things, such as risking lives or sacrificing nature or

human rights, but they are not happy with themselves for doing so, if they are even aware of what they are doing. These values appear in CV survey responses (Mitchell & Carson, 1989). When asked for their WTA or even their WTP for natural resources, some respondents say "zero" or "no amount" because they think that "we shouldn't put a price on nature."

If we are trying to maximize utility through policy decisions, we must take everyone's values into account. We cannot, for example, rely on the median valuation, as we might do if we were trying to simulate a referendum. (All that matters in a referendum is whether median monetary value of the good is more or less than cost of the good. Only if the median is higher than cost will a majority vote for it, knowing the cost.)

Under the assumption that people strive to take the degree of everyone's values into account, protected values cause trouble for policy makers because these values imply that one value is infinitely important, relative to others. If the value of forests is infinite, we will simply not cut them down, and we will have to find substitutes for wood. Even if only a few people place such an infinite value on forests, their values will trump everyone else's values and everyone else will spend more money on dwellings and furniture, at least. Protected values can also conflict. If some people have protected values for yew trees while others have protected values for the rights of cancer patients to the drug that is produced from the yew trees, no solution seems possible. Of course, we could honor one side or the other, ignoring the values of patient lovers or tree lovers. (Or we could compromise, leaving everyone dissatisfied.) But the choice of the solution would be unaffected by the number of those who favored patients versus trees. We are not voting; rather, we are asking which of two infinite values is larger.

Such a situation violates apparent normative principles of decision making. For example, it is reasonable to think that, for two options L and T, people either prefer L, prefer T, or are indifferent. In the situation just described, they would be indifferent because either solution is "optimal," in the sense that any improvement for one person would make someone else worse off by at least the same amount. Yet, a doubling of the number of people who preferred L or T would not change the decision. This seems to violate a principle of dominance, which could be stated roughly as, "If people are indifferent between L and T and then get an additional reason for L (or T), they should then favor L (or T)." The same problems arise within an individual who holds conflicting protected values. An additional argument for one option or another will not swing the decision. To avoid problems of this sort, most normative theories of decision making assume that values can be traded off. That is, for any pair of values, a sufficiently small change in the satisfaction of one value can be compensated by a change in some other value. (Technically, this amounts to a form of an Archimedean axiom [Krantz et al., 1971].) The values may be held by the same person or by different people. Values are called *compensatory* when they are part of such a pair.

The seriousness of these problems and the possibility of solutions to them may depend on the nature of protected values themselves. Baron and Spranca (in press) proposed that protected values derive from deontological rules, rules that prescribe action and inaction "whatever the consequences." An example of such a rule is "Do not destroy natural processes

irreversibly.' Such a rule would prohibit people from destroying species, but it would not, for example, oblige them to prevent natural extinctions. When people who try to follow such rules are asked about their values, they are reminded of the rules. They take the value questions to bear on the actions that the rules prohibit.

Most methods of value elicitation, such as WTA, allow the respondents to interpret questions as referring to their actions, such as accepting money in return for allowing something to happen or giving consent for a group to make such a deal. In some cases, the respondents might reasonably think that their answers would be used to make real decisions; by answering the questions, they play a role in determining the outcome. Even when the questions are hypothetical, respondents interpret them in this way because hypothetical questions are clearly simulations of real questions.

In several studies, Baron and Spranca (in press) defined protected values in terms of absoluteness, that is, unwillingness to make trade-offs. Such absoluteness was correlated—across different values within respondents—with several other properties derived from the theory that protected values stem from deontological rules. The first three properties follow directly from the proposal that protected values arise from deontological rules.

1. *Quantity insensitivity.* Quantity is irrelevant for protected values. Destroying 1 species is as bad as destroying 100. The protected value applies to the act, not the result (although a compensatory value may apply to the result as well). Some opponents of abortion seem to feel this way when they oppose government spending money on international family-planning programs that perform abortions, even if the money does not pay for the abortions and even if other expenditures actually reduce the number of abortions performed. It is not the number of abortions they care about. Of course, I have already noted that insensitivity to quantity is a more general problem. Baron and Spranca (in press) found that it was greater when protected values were involved: Respondents who said that a trade-off should not be made would sometimes still make it, but they would neglect quantity.

2. *Agent relativity.* Protected values are *agent relative*, as opposed to *agent general* (McNaughton & Rawling, 1991). This means that participation of the decision maker is important, as opposed to the consequences themselves. This follows from the assumption that protected values arise as rules about action. For example, in one item, respondents would not buy stock in a company that violated a protected value, even if someone else bought the stock anyway and if the price of the shares were unaffected.

3. *Moral obligation.* The actions required or prohibited by protected values are seen as moral obligations, as explained by Turiel (1983). Moral obligations are not just conventions or personal preferences. They are seen as universal and independent of what people think. They are also seen as objective obligations: People should try to carry them out even if they do not think they should.

Three other properties follow from those listed above, along with other assumptions:

4. *Overstatement.* People who hold protected values may feel that it is morally correct to overstate the strength of these

values in public discussion. This property was not highly correlated with other properties in several analyses.

5. *Denial of trade-offs.* People may resist the idea that anything must be sacrificed at all for the sake of their value. People generally tend to deny the existence of trade-offs (Jervis, 1976, pp. 128–142; Montgomery, 1984), and this tendency may be particularly strong when one of the values involved is not supposed to trade off with anything. People may desire to believe that their values do no harm.

6. *Anger.* People may become angry at the thought of a violation of a protected value. This is a consequence of its being a moral violation.

Conclusion

In this review, I have identified several biases, which can be classified into two types: insensitivity to quantity and contamination of judgments by irrelevant factors. I have also identified some paths worth further exploration as ways to eliminate these biases. Many of these paths involve efforts to make the task easier, to exclude irrelevant information from the task, or to force respondents to attend to relevant information.

Can Making the Task Easier Reduce Biases?

Respondents in most value elicitation studies find the tasks difficult. For example, most would find it much easier to rate importance on a 7-point scale—as long as they do not stop to think about the quantity of each item they rate. It is almost painful to be asked questions such as, How much would you pay in auto-inspection costs to reduce the death rate in Philadelphia, PA (a city of 1.5 million people) by 10 deaths per year?

The difficulty has many sources. One, often forgotten, is that psychophysical judgments are imprecise, and the task is a kind of psychophysical judgment. It is similar to presenting two tones that differ greatly in loudness and asking, "If the first has a loudness of 10, how loud is the second?" Respondents cannot find reasons for choosing one response over another within a broad range of possible responses. In principle, the task might be easier if respondents could give a range of answers, although assigning the ends of the range might pose the same problem. Dubourg et al. (1994) found that assigning ends did not reduce the large discrepancy between WTA and WTP. In fact, the ranges given did not even overlap in most cases.

In psychophysics, we often ask observers to make the judgments over and over; the errors then average out, and we are left with orderly data from a single respondent. In value elicitation, we either question many respondents or else we do not worry about being off by less than an order of magnitude. For many decisions, it is the order of magnitude that matters. For example, the cost per life saved by regulations of chemicals varies across different regulations by several orders of magnitude (Breyer, 1993). Even if an estimate of the monetary value of a life were as imprecise as "somewhere between \$1 million and \$10 million," it would suffice to decide whether almost all regulations were cost effective.

Another source of difficulty is the imagining of very large or very small numbers. If death rates from air pollution are translated into probabilities per year, the numbers are tiny. It may be

helpful to compare them with a scale of infrequent events, such as lightning strikes of people. Or one might try to translate the problem into something meaningful: "A football stadium is full of people. Someone, selected at random, will die within 1 year unless all of the people pay some amount of money. How much should they pay?" Another solution is to think about the problem in a series of smaller steps, although, of course, errors could compound. Little research has been conducted on the efficacy of such methods to improve reliability or reduce biases.

Quantity Insensitivity and the Concept of Importance

Insensitivity to quantity in CV and insensitivity to ranges in MA are closely related. Each may have several causes. For example, insensitivity in CV is exacerbated by protected values. But both kinds also seem to involve a focus on an undifferentiated concept of importance. The same undifferentiated concept may explain the failure to differentiate total value from value per dollar of government expenditures.

The concept of importance seems to be undifferentiated (Schoemaker, 1981). It is analogous to young children's concepts of quantity. Young children do not distinguish number and length. When asked which of two rows of dots is longer or has more, children give the same answer, no matter what they were asked; the answer they give seems to depend on how easy it is to count the dots (Baron, Lawson, & Siegel, 1975). Likewise, when adults are asked about values in monetary terms, they seem to answer in terms of the same dimension, no matter what they are asked. The dimension they use may depend on various factors but not much on the question.

What concept of importance do people use when they are asked about values? One possibility is that, when asked about trade-offs between some value and money, for example, they focus on the price of particular examples that represent the value in question. This seems unlikely to explain all the results because quantity insensitivity is found for ratings as well as trade-offs with money (Kahneman & Ritov, 1994).

Another possibility is that respondents think of value questions as questions about individual differences. So when asked how much they value life versus money, people say that they value life more if they think that their value for life, relative to money, is higher than that of others. This seems insufficient too because people often agree on which of two values is more important, even when quantity is not involved (Tversky et al., 1988); but almost everyone claims to value life more than money. However, this individual-difference interpretation may still apply to results concerning government expenditures.

A similar possibility for importance judgments of public goods is that respondents judge a good to be important (or unimportant) if they think more (or less) of the good should be provided than is currently being provided, taking into account the cost of a provision as they perceive it. Of course, this is not the judgment that is needed for policy decisions because people are supposed to judge the benefits of proposals alone, not their benefit-cost ratio.

Finally, the undifferentiated concept of importance could be a manifestation of unreflective generalizations. Although the statement, Life is more important than money, is either meaningless or inaccurate as a general statement about most people's

values, they may still believe that it is true, just as they believe in the truth of such generalities as, Honesty is the best policy, until they are faced with counterexamples. Such generalities are necessary for quantitative measurement of values, but they may interfere with tasks designed to measure these values.

The existence of such generalities does not mean that people cannot make quantitative judgments. People can still meaningfully answer questions, such as, Which matters more to you: the difference between spending nothing and spending \$200/year or the difference between your current annual risk of death and your annual risk if you took one additional 1,000-mile (1,609.35 km) airplane trip per year? Some values researchers have suggested that, after we strip away all the irrelevant influences on tasks such as this, nothing remains. It is indeed possible that some people simply have no values relevant to certain decisions, but we do not need to conclude this for most public policy questions. Moreover, people with no relevant values should be indifferent, thus willing to let others decide.

To assess values, we may need to teach respondents about the purpose of quantitative value measurement. Students in courses I have taught on MA seem to understand this idea. Although their judgments are not always consistent, the inconsistencies seem to result from sheer difficulty of imagination rather than from any fundamental insensitivity to quantity. This is an extreme case. We will not, I hope, have to make all our respondents sit through a course on MA.

We must also make sure to ask questions in a way that avoids distortion from anchoring and underadjustment. We may need to use tasks that do not provide anchors, even if these tasks are more difficult and more prone to error. The error, although larger, may be unbiased. For example, we may ask respondents simply (and repeatedly) to pick two intervals, one on each of two dimensions, that are equal in utility, with no hints about what values to use. Another promising approach to multiattribute comparison is to ask for holistic ratings of attribute pairs. (But more than two attributes at a time may cause difficulty.)

Contamination by Other Judgments

Some of the problems with quantity, as just noted, involve contamination, that is, giving the right answer to the wrong question. This includes failing to differentiate total value from value per dollar in the case of government expenditures.

I have reviewed two other kinds of contamination—(a) influence by judgments of cost (to the government) rather than value (to those who benefit) and (b) influence by judgments of the cause of harm, human versus natural. In the latter case, judgments about the value of the harm itself may be contaminated by judgments about the extent to which someone should pay a penalty for having caused it (Baron, 1993). In a similar vein, judgments based on protected values can be seen as a contamination of judgments of the value of consequences by judgments of the rightness of acts that might lead to those consequences.

The idea of contamination by opinions about actions raises a question: How do we draw the line between measuring values for outcomes and measuring political opinions about the means of producing them? Consider, for example, the decision about whether drivers should have to get their cars inspected and main-

tained to reduce air pollution. We might try to decide this question by measuring values for the time and money spent on inspections and repairs and for the various health and aesthetic consequences of air pollution. But what if people value "autonomy" because they have a prior political commitment to freedom from government coercion? What if they think that disease caused by pollution is much worse than the same disease caused by nature? Or they consider the haze caused by pollution ugly but the same appearance caused by a volcanic eruption beautiful? If we take the value of autonomy or of human interference with nature as part of our analysis, then it seems that we are doing something more like an opinion poll about what should be done than a measurement of values for outcomes. The purpose of value measurement for policy is to try to decide on the policy that is best in terms of the outcomes it yields.

However, actions are themselves consequences of policies. Policies affect what people do. If inspection and maintenance are required, then people will be coercively punished for not following the law. This is a government action that some people do not want to encourage. Conversely, to not have a program will mean that drivers cause others to suffer and die in ways they could easily prevent (through an inspection of their cars). This kind of action itself might be disvalued.

Perhaps we can deal with this by separating the values for actions as consequences, such as those just mentioned, from opinions about which policies should be adopted. Thus, if necessary, we could measure values for countable acts of coercion or pollution, but we could try to avoid contamination from respondents' prior opinions about whether government is too coercive or not coercive enough.

Even the values for actions, however, might not be fundamental. They might be derived from a belief that government coercion does no good or that government coercion is needed to prevent people from harming each other (e.g., by polluting). In theory, we could determine whether these values are proxy values or fundamental values by asking people about their beliefs and about what values they would hold if their beliefs were different. In practice, two problems may prevent this. First, proxy values may become fundamental over time, detached from their origins yet still irrationally held if they arose because of false beliefs (Baron, 1996). Second, respondents may be reluctant to admit that their values depend on beliefs that may be incorrect. Beliefs and values tend to correlate highly when they support the same opinion about what should be done (Ellsworth & Ross, 1983), so it seems that people try to convince themselves that these beliefs are true. People may have a hard time answering questions such as, Would your value for autonomy change if you believed that government coercion were effective?

One practical solution to this problem is to try to do the analysis both ways, with and without values for actions. This requires that we induce respondents to separate their values concerning actions from their values concerning other consequences. If the results lead to the same policy recommendations, then the issue would not have to be resolved. If they do not, then some other solution might be warranted. For example, perhaps extensive discussion among two contrasting sides could make the parties realize how much their values did in fact depend on their beliefs.

More generally, two approaches are possible to remove contamination. One is to remove possible contaminants from the questions that we ask, for example, remove discussion of how a consequence would be brought about. The other is to make the contaminant explicit and ask the respondent to evaluate it separately, in hopes that the respondent, having this opportunity, will then be able to avoid contamination in the judgment of interest. Little research has been conducted on the effectiveness of either method.

The effectiveness of the latter approach depends to some extent on whether respondents can become aware of the contamination effects. Breaking up the judgment into parts succeeds only if the respondents can separate one part from another. Whether they can do this may depend on their theories about the source of their judgments (Wilson & Brekke, 1994). It may also depend on whether their judgments were true expressions of value or blind application of heuristics. These conditions may depend on the situation, so they may need to be checked in each case.

References

- Anderson, N. H., & Zalinski, J. (1988). Functional measurement approach to self-estimation in multiattribute evaluation. *Journal of Behavioral Decision Making*, 1, 191-221.
- Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal*, 100, 464-477.
- Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology*, 63, 320-330.
- Baron, J. (1993). Heuristics and biases in equity judgments: A utilitarian approach. In B. A. Mellers & J. Baron (Eds.), *Psychological perspectives on justice: Theory and applications* (pp. 109-137). New York: Cambridge University Press.
- Baron, J. (1994). Nonconsequentialist decisions (with commentary and reply). *Behavioral and Brain Sciences*, 17, 1-42.
- Baron, J. (1995a). Blind justice: Fairness to groups and the do-no-harm principle. *Journal of Behavioral Decision Making*, 8, 71-83.
- Baron, J. (1995b). Rationality and invariance: Response to Schuman. In D. J. Bjornstad & J. Kahn (Eds.), *The contingent valuation of environmental resources: Methodological issues and research needs* (pp. 145-163). London: Edward Elgar.
- Baron, J. (1996). Norm-endorsement utilitarianism and the nature of utility. *Economics and Philosophy*, 12, 165-182.
- Baron, J., & Greene, J. (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.
- Baron, J., Lawson, G., & Siegel, L. S. (1975). Effects of training and set size on children's judgments of number and length. *Developmental Psychology*, 11, 583-588.
- Baron, J., & Maxwell, N. P. (1996). Cost of public goods affects willingness to pay for them. *Journal of Behavioral Decision Making*, 9, 173-183.
- Baron, J., & Spranca, M. (in press). Protected values. *Organizational Behavior and Human Decision Processes*.
- Barron, F. H., von Winterfeldt, D., & Fischer, G. (1984). Empirical and theoretical relationships between value and utility functions. *Acta Psychologica*, 56, 233-244.
- Beattie, J., & Baron, J. (1991). Investigating the effect of stimulus range on attribute weight. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 571-585.
- Bosch, J. L., & Hunink, M. G. M. (1996). The relationship between

- descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Medical Decision Making*, 16, 217–225.
- Boyle, K. J., Desvousges, W. H., Johnson, F. R., Dunford, R. W., & Hudson, S. P. (1994). An investigation of part-whole biases in contingent valuation studies. *Journal of Environmental Economics and Management*, 27, 64–83.
- Brandt, R. B. (1979). *A theory of the good and the right*. Oxford, England: Clarendon Press.
- Breyer, S. (1993). *Breaking the vicious circle: Toward effective risk regulation*. Cambridge, MA: Harvard University Press.
- Carson, R. T., & Mitchell, R. C. (1993). The issue of scope in contingent valuation. *American Journal of Agricultural Economics*, 75, 1263–1267.
- Carson, R. T., & Mitchell, R. C. (1995). Sequencing and nesting in contingent valuation surveys. *Journal of Environmental Economics and Management*, 28, 155–173.
- DeKay, M. L., & McClelland, G. H. (1993, November). *Errors in estimating outcome utilities: The flip side of outcome bias*. Paper presented at the conference of the Judgment and Decision Making Society, Washington, DC.
- DeKay, M. L., & McClelland, G. H. (1996). Probability and utility components of endangered species preservation programs. *Journal of Experimental Psychology: Applied*, 2, 60–83.
- de Neufville, R., & Delquie, P. (1988). A model of the influence of certainty and probability “effects” on the measurement of utility. In B. Munier (Ed.), *Risk, decision, and rationality* (pp. 189–205). Dordrecht, The Netherlands: Reidel.
- Diamond, P. A., & Hausman, J. A. (1994). Contingent valuation: Is some number better than no number? *Journal of Economic Perspectives*, 8, 45–64.
- Diamond, P. A., Hausman, J. A., Leonard, G. K., & Denning, M. A. (1993). Does contingent valuation measure preferences? Some experimental evidence. In J. A. Hausman (Ed.), *Contingent valuation: A critical assessment*. Amsterdam: North Holland Press.
- Dolan, P., Jones-Lee, M., & Loomes, G. (1995). Risk-risk vs. standard gamble procedures for measuring health state utilities. *Applied Economics*, 27, 1103–1111.
- Dubourg, W. R., Jones-Lee, M. W., & Loomes, G. (1994). Imprecise preferences and the WTP-WTA disparity. *Journal of Risk and Uncertainty*, 9, 115–133.
- Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, 60, 306–325.
- Ellsworth, P. C., & Ross, L. (1983). Public opinion and capital punishment: A close examination of the views of abolitionists and retentionists. *Crime and Delinquency*, 29, 116–169.
- Fischer, G. W. (1995). Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes*, 62, 252–266.
- Fischhoff, B., Quadrel, M. J., Kamlet, M., Loewenstein, G., Dawes, R., Fischbeck, P., Klepper, S., Leland, J., & Stroh, P. (1993). Embedding effects: Stimulus representation and response mode. *Journal of Risk and Uncertainty*, 6, 211–234.
- Froberg D. G., & Kane R. L. (1989). Methodology for measuring health-state preferences. II: Scaling methods. *Journal of Clinical Epidemiology*, 42, 459–471.
- Green, D. P., Kahneman, D., & Kunreuther, H. (1994). How the scope and method of public funding affects willingness to pay for public goods. *Public Opinion Quarterly*, 58, 49–67.
- Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 45, 33–41.
- Green, P. E., & Wind, Y. (1973). *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, IL: Dryden Press.
- Guagnano, G. A., Dietz, T., & Stern, P. C. (1994). Willingness to pay for public goods: A test of the contribution model. *Psychological Science*, 5, 411–415.
- Guiot, J. M., & Lefoll, J. (1988). Cardinal utility: An empirical test. In B. Munier (Ed.), *Risk, decision, and rationality* (pp. 97–102). Dordrecht, The Netherlands: Reidel.
- Hanemann, W. M. (1994). Valuing the environment through contingent valuation. *Journal of Economic Perspectives*, 8, 19–43.
- Hershey, J. C., & Schoemaker, P. J. H. (1980). Prospect theory’s reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance*, 25, 395–418.
- Hershey, J. C., & Schoemaker, P. J. H. (1986). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, 31, 1213–1231.
- Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, 57, 116–131.
- Irwin, J. R. (1994). Buying/selling price preference reversals: Preference for environmental changes in buying versus selling modes. *Organizational Behavior and Human Decision Processes*, 60, 431–457.
- Jervis, R. (1976). *Perception and misperception in international politics*. Princeton, NJ: Princeton University Press.
- Jones-Lee, M. W. (1989). *The economics of safety and physical risk*. Oxford, England: Basil Blackwell.
- Jones-Lee, M. W., Loomes, G., & Phillips, P. R. (1995). Valuing the prevention of non-fatal road injuries: Contingent valuation vs. standard gambles. *Oxford Economic Papers*, 47, 676.
- Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, 22, 57–70.
- Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76, 728–741.
- Kahneman, D., Knetsch, J. L., & Thaler, R. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98, 1325–1348.
- Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study of the headline method. *Journal of Risk and Uncertainty*, 9, 5–38.
- Kahneman, D., Ritov, I., Jacowitz, K. E., & Grant, P. (1993). Stated willingness to pay for public goods: A psychological perspective. *Psychological Science*, 4, 310–315.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47, 263–291.
- Kamlet, M. S. (1992). *A framework for cost-utility analysis of government health care programs* (Doc. No. 1992-617-025/68280). Washington, DC: U.S. Government Printing Office.
- Keeney, R. L. (1992). *Value-focused thinking: A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.
- Kemp, M. A., & Maxwell, C. (1993). Exploring a budget context for contingent valuation estimates. In J. A. Hausman (Ed.), *Contingent valuation: A critical assessment*. Amsterdam: North Holland Press.
- Kemp, S., & Willetts, K. (1995). Rating the value of government-funded services: Comparison of methods. *Journal of Economic Psychology*, 16, 1–21.
- Kempton, W., Boster, J. S., & Harley, J. A. (1995). *Environmental values in American culture*. Cambridge, MA: MIT Press.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.
- Krzysztofowicz, R. (1983). Strength of preference and risk attitude in utility measurement. *Organizational Behavior and Human Performance*, 31, 88–113.
- Loomes, G. (1993). Disparities between health state measures. Is there a rational explanation? In B. Gerrard (Ed.), *The economics of rationality* (pp. 149–178). London: Routledge.

- Loomes, G. (1994). *Valuing health and safety: Some economic and psychological issues*. Unpublished manuscript, University of York, Department of Economics, York, England.
- Louviere, J. J. (1988). *Analyzing individual decision making: Metric conjoint analysis*. Newbury Park, CA: Sage.
- Margolis, H. (1982). *Selfishness, altruism, and rationality: A theory of social choice*. New York: Cambridge University Press.
- Marshall, J. D., Knetsch, J. L., & Sinden, J. A. (1986). Agents' evaluations and the disparity in measures of economic loss. *Journal of Economic Behavior and Organization*, 7, 115-127.
- McFadden, D. (1994). Contingent valuation and social choice. *American Journal of Agricultural Economics*, 76, 689-708.
- McNaughton, D., & Rawling, P. (1991). Agent relativity and the doing-happening distinction. *Philosophical Studies*, 63, 167-185.
- Mellers, B. A., & Cooke, A. D. J. (1994). Trade-offs depend on attribute range. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1055-1067.
- Mitchell, R. C., & Carson, R. T. (1989). *Using surveys to value public goods: The contingent valuation method*. Washington, DC: Resources for the Future.
- Miyamoto, J. M., & Eraker, S. A. (1988). A multiplicative model of survival duration and health quality. *Journal of Experimental Psychology: General*, 117, 3-20.
- Montgomery, H. (1984). Decision rules and the search for dominance structure: Towards a process model of decision making. In P. C. Humphreys, O. Svenson, & A. Vari (Eds.), *Analyzing and aiding decision processes* (pp. 343-369). Amsterdam: North Holland Press.
- National Oceanographic and Atmospheric Administration. (1993). Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58(10), 4602-4614.
- Nord, E. (1992). Methods for quality adjustment of life years. *Social Science and Medicine*, 34, 559-569.
- Nord, E., Richardson, J., Kuhse, H., & Singer, P. (1995). Who cares about cost? Does economic analysis impose or reflect social values? *Health Policy*, 34, 79-94.
- Peretz, P. (1983). *The political economy of inflation in the United States*. Chicago, IL: University of Chicago Press.
- Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research*, 28, 206-224.
- Portney, P. R. (1994). The contingent valuation debate: Why economists should care. *Journal of Economic Perspectives*, 8, 3-17.
- Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin*, 86, 777-803.
- Richardson, J. (1994). Cost-utility analysis: What should be measured? *Social Science and Medicine*, 39, 7-21.
- Schelling, T. C. (1968). The life you save may be your own. In S. B. Chace, Jr. (Ed.), *Problems in public expenditure analysis* (pp. 127-175). Washington, DC: Brookings Institution.
- Schkade, D. A., & Payne, J. W. (1994). How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation. *Journal of Environmental Economics and Management*, 26, 88-109.
- Schoemaker, P. J. H. (1981). Behavioral issues in multiattribute utility modeling and decision analysis. In J. N. Morse (Ed.), *Organizations: Multiple agents with multiple criteria*. Heidelberg, Germany: Springer-Verlag.
- Schuman, H. (1995). The sensitivity of CV outcomes to CV survey methods. In D. J. Bjornstad & J. Kahn (Eds.), *The contingent valuation of environmental resources: Methodological issues and research needs* (pp. 75-96). London: Edward Elgar.
- Shepard, R. N. (1964). On subjectively optimum selection among multi-attribute alternatives. In M. W. Shelley & G. L. Bryan (Eds.), *Human judgments and optimality* (pp. 257-281). New York: Wiley.
- Slovic, P. (1995). The construction of preferences. *American Psychologist*, 50, 364-371.
- Slovic, P., Lichtenstein, S., & Fischhoff, B. (1979). Images of disaster: Perception and acceptance of risks from nuclear power. In G. Goodman & W. Rowe (Eds.), *Energy risk management* (pp. 223-245). London: Academic Press.
- Stalmeier, T. F. M., & Bezembinder, P. G. G. (1996). Proportional heuristics in time tradeoff and conjoint measurement. *Medical Decision Making*, 16, 36-44.
- Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science*, 4, 199-214.
- Thaler, R. (1993, August). *Mental accounting matters*. Paper presented at Subjective Probability, Utility, and Decision Making Conference 14, Aix-en-Provence, France.
- Torrance, G. W. (1986). Measurement of health-state utilities for economic appraisal: A review. *Journal of Health Economics*, 5, 1-30.
- Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.
- Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, 95, 371-384.
- Ubel, P. A., Loewenstein, G., Scanlon, D., & Kamlet, M. (1996). Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Medical Decision Making*, 16, 108-116.
- Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making*, 5, 169-185.
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, England: Cambridge University Press.
- Weber, M., & Borchering, K. (1993). Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operations Research*, 67, 1-12.
- Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116, 117-142.
- Winer, R. S. (1986). A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research*, 13, 250-256.
- Zakay, D. (1990). The role of personal tendencies in the selection of decision-making strategies. *Psychological Record*, 40, 207-213.

Received March 4, 1996

Revision received August 2, 1996

Accepted August 24, 1996 ■