Prospects for utilitarian decision analysis

Jonathan Baron[*]

Department of Psychology

University of Pennsylvania

Abstract

Decision analysis evaluates options in terms of their expected utility, the expected amount of good. Many regulatory decisions require estimates of utility, which is not the same as monetary outcomes or other objectively measurable quantities. I review some of the problems of utility measurement for consequences (hence "utilitarian") and suggest possible solutions. Contingent valuation (CV), a widely used method for obtaining money equivalents, is usually insensitive to the quantity of the good being evaluated, but this problem might be avoided by using per-unit pricing and/or by requiring responses on both attributes (money and the good in question).

Valuation of consequences can also be distorted by provision of information about how the consequences are achieved. I report a study showing how valuation can also be improved by stripping away information other than the consequences of interest; most subjects did not think that removal of the extra information was unfair.

Conjoint analysis, another method in which respondents rate objects that vary in different attributes such as cost and benefit, seems promising because it is often sensitive to quantity. I report a study suggesting that it can suffer from the problem that people limit their attention to a few attributes, possibly even one. I conclude with a discussion of the role of utility measurement in regulation.

Keywords: contingent valuation, public goods, conjoint analysis, biases, regulation.

[*] This paper is an extended version of a presentation at a conference on "New ideas for risk regulation," at Resouces for the Future, Washington, DC, June 22–23, 2009.) Email: baron@psych.upenn.edu.

1 Introduction

Decision analysis is a form of applied decision theory that originally attempted to apply utility theory more or less directly to the analysis of decisions (e.g., Brown, 2006; Keeney & Raiffa, 1993; Raiffa, 1968). In one form, the analyst and client assigned utility numbers to possible outcomes, estimated the probabilities of these outcomes under each option, and then multiplied the probabilities by the utilities in order to calculate an expected utility for each option. In another form, multi-attribute analysis, options are analyzed into attributes, each corresponding to an objective or sub-objective. Each option is given a utility number on each attribute, and each attribute is given a weight. The weights are multiplied by the respective attribute utilities for each option and added up across attributes, to produce an overall utility number for the option. Attributes must have independent effects on utility in order for the addition to serve its intended purpose.

Decision analysis can be applied to the analysis of risk regulation, as well as many other problems. Indeed, some of the textbook examples involve decisions about risk. Decision analysis has been applied extensively in medicine, where risks are those of side effects, various medical conditions, and iatrogenic harms. (The journal *Medical Decision Making* contains many examples.) Often, in medicine, analysis is applied to the health benefits and harms of treatment options, with their monetary cost put aside. The decision analysis yields an expected utility for each treatment, and this may be compared to its cost to yield something like a utility-per-dollar estimate. This is a form of cost-effectiveness analysis, where the "effectiveness" is described in terms of (expected) utility numbers.

Most decision analyses make a strict distinction between consequences, options, and unknown states of the world (to which probabilities are assigned). Utilities are assigned to the consequences, not to properties of the options that yield them. In essence, a consequence is a description of some state of the world (or category of states) to which we assign value, regardless of what led to that state (Baron, 1993, 1996). Equivalently, we often speak of utility as the degree of goal achievement. We could assign utilities to acts or options as well as to consequences, if we have value for them. But such a move has two problems. First, the values we assign to options themselves might be derived from beliefs about the effectiveness of each option in achieving our goals. If our beliefs are incorrect, then we could be misled. (These are "means values" in the sense described by Keeney, 1992.) Second, the values for options could result from *opinions* about what option should be chosen, that is, provisional answers to the question about what to do. If we insert our prior opinions into an analysis this way, then the chance of getting a true second-opinion from the analysis is reduced.

Thus, I want to concentrate here on what I take to be the most useful form of analysis, which is based on consequences alone. I want to put aside various efforts to incorporate opinions about properties of options in the analysis (e.g., Deber & Goel, 1990; Volk, Lok, Ubel, & Vijan, 2008). I call this rarified form of decision analysis "utilitarian" exactly because it is about consequences, and it puts aside any deontological principles that might limit or discourage certain types of acts. If, faced with a discrepancy between the results of a well-done utilitarian decision analysis and our moral principles, we will at least be able to estimate the cost we pay — in reduced utility — for following those principles.

Note that, if our utilitarian analysis of a decision is correct, then any other option than the one it recommends is likely to cause harm to someone without any compensating benefit to anyone else. In saying this, I assume that utility numbers are not just the output of some black box but rather measures of good, of what we value, of the extent to which our goals are achieved (Baron, 1993, 1996, 2008). Moreover, I assume that it makes sense to speak of utility differences. We can say meaningfully that the difference between outcomes A and B is larger than that between B and C. Thus, utility is an interval scale. (But it is not a ratio scale; it has no natural zero.)

In order to apply decision analysis in any form, we must rely on judgments of the utility of consequences, or on decisions — real or hypothetical — from which such judgments can be inferred. We face a problem like that faced by those who tried to measure time throughout most of human history. The concept of time as an interval scale may have existed before it could be measured very well. Early attempts, such as sundials, suffered from consistent biases. Yet, they were better than nothing, and often quite useful. That is our situation now. When we measure utility for purposed of decision analysis, it may not matter much if we are off by a power of 10. In some cases, it seems that regulatory decisions are “off” by several powers of 10 (Breyer, 1993; Tengs et al., 1995).

On the other hand, studies of judgments and decisions suggest that our inferences about utility may be off by more than a power of 10. Decisions, both real and hypothetical, are affected by biases such as the prominence effect (Baron & Greene, 1996; Tversky, Sattath & Slovic, 1988), in which judgments of trade-offs of attributed are affected by the relative “importance” of the attributes without regard to the extent to which each attribute varies. (For example, people respond as if they thought that “life is more important than money” no matter how much money or how low the risk of a single lost life.) Or, people attend to risk ratios rather than risk differences; the latter are relevant for expected utility. If a proposal quadruples the risk of harm A, raising it from .0001 to .0004, but reduced the risk of harm B by only 10%, from .5000 to .4500, people often pay more attention to the effect on harm A (Baron, 1997b).

2 Scope and range effects

Another major bias is the lack of attention to numerical variation in the good to be valued. At first this was found in between-subject experiments in which different subjects were asked how much they would pay for different quantities of a public good, such as cleaning up the pollution in some number of lakes (e.g., Kahneman & Knetch, 1992); Baron (1997a) reviews the literature. The method of asking about willingness to pay (WTP) is a form of matching, in which one attribute is provided, the quantity of a particular good, and the subject is asked to provide a quantity on another dimension, money. In the literature on “contingent valuation” (CV), however, the use of money took on a privileged status because WTP was thought to represent an economic value — not a utility judgment as such — although we might want to conclude that the monetary difference in wealth and the difference in the quantity of the good had equal utility. The use of between-subject designs was standard practice in contingent valuation, but the same result was found in within-subject designs (in which each subject answers questions about both quantities; Baron, 1997a).

A panel convened to establish good practice for CV recommended a "scope test," a demonstration that the method being used was sensitive to the magnitude of the good being evaluated (NOAA, 1993). The NOAA panel required *some* sensitivity, but did not require proportionality, in which 10 times the amount of the good should have approximately 10 times the monetary value. Arguably, the utility of money is marginally declining, so a loss of 10 times X might have much more disutility than 10 times the disutility of losing X. Thus some insensitivity would be reasonable, although in most studies the degree of insensitivity is quite large, without even a doubling of WTP when the good is multiplied by 10. Baron and Greene (1996) pointed out that the utility function for money cannot begin to account for the large insensitivity that is typically found: the same degree of insensitivity is found in willingness to accept (WTA), which should show super-sensitivity according to the hypothesis based on the utility function for wealth.[1]

3 Conjoint analysis as a possible solution

One solution to the problem of neglect of ranges is conjoint analysis, a technique common in marketing research. The subject's task is typically to rate a number of items in a single session. The items vary in attributes of interest, such as a car's price, fuel efficiency, safety rating, and repair record. Ideally, each subject sees all possible combinations of a few levels of each attribute, but usually it is necessary to use a smaller set (Louviere, 1988). We can fit a statistical model to predict each subjects's ratings from the levels of the attributes. Unlike CV or other matching tasks, where one attribute is given and the subject must provide another attribute (money) to match the given attribute, all attributes vary, so they are all on an equal footing in their potential influence. From a model of responses as a function of attribute levels, we can determine how much of a change in one attribute is required to compensate for a given change in another attribute so that the rating stays the same. We could think of this as a rate of substitution, e.g., substitution between price and repair record.

An important question is whether the rate of substitution depends on the actual numbers given, or whether it is also influenced by the range of numbers on each attribute within the task. The results conflict. Beattie and Baron (1991), using such a holistic rating task, found no effects of relative ranges on rates of substitution with several pairs of attributes, but we found range effects with some attributes, particularly those for which the numerical representation

[1] Some CV studies do show complete sensitivity. For example, Corso, Hammitt, and Graham (2001) found that the use of a visual aid led to proportional WTP for mortality risk reduction. However, others have failed to find proportionality with similar visual aids. In general, James Hammitt and his collaborators have been among the few researchers to accept the importance of proportionality, rather than a simple scope effect of any size, as a criterion for the validity of CV.

was not clearly connected to fundamental objectives, e.g., numerical grades on an exam. (The meaning of exam grades depends on the variance.) This gave us hope that holistic ratings could provide consistent and meaningful judgments of tradeoffs. Lynch et al. (1991) also found mostly no range effects for hypothetical car purchases, except in one study with novice consumers. (They used correlations rather than rates of substitution, however, so it is difficult to tell how much their results were due to changes in variance.) Mellers and Cooke (1994), however, found range effects in tasks where the relation of the numbers to fundamental objectives was clear, e.g., distance to campus of apartments.

It is possible that, when people must attend to several attributes varying at once, they are less likely to attend to the attributes they consider less important. If so, it might be better to present two attributes at a time, using, if necessary, all possible pairs. Indeed, for some time it has been suggested that weights derived from such multi-attribute tasks are more variable than weights derived from direct ratings (von Winterfeldt & Edwards, 1986, section 10.4).

4 Experiment 1: Attention in conjoint tasks (with J. Yadavaia)

Beattie and Baron used two attributes at a time. It is possible that subjects have trouble attending to several attributes, as argued by Slovic (1969), Hoffman, Slovic, & Rorer (1968), and Ebbesen and Konecni (1975). This experiment compared two- and three-attribute tasks. It is possible that, with three attributes, subjects will attend less to the attribute that would get the least weight.

We [2] constructed a three-attribute stimulus, each attribute with three levels. Each subject rated four sequences of cases. In one sequence of 27 trials, all three attributes varied orthogonally. In three other sequences of 9 trials each, one attribute was held constant at its middle level, and the two others varied orthogonally; thus, in these sequences the subject needed to attend to only two attributes at a time. We computed weights for each attribute in the 2- and 3-attribute conditions and normalized these so that the three weights summed to 1. Then we estimated a weight for each of the three attributes by summing the 2- and 3-attribute weights for that attribute. Of interest is what happens to attributes with low weights. If subjects attend less to them in the 3-attribute condition than in the 2-attribute condition, then the 3-attribute weight will be lower than the 2-attribute weight. We thus hypothesize that the *difference* of 2-attribute-weight minus 3-attribute-weight will be higher, and positive, when the overall weight (the sum of the two weights) is low. We test this by regression across subjects. Happily, subjects differed considerably in the weights they assigned to the three attributes.

[2] This experiment was done as an undergraduate course project by James Yadavaia, who is now a graduate student in psychology at the University of Nevada, Reno.

4.1 Method

Fifty-three subjects completed a questionnaire on the World Wide Web for $2 each. Five were dropped because their response patterns suggested haphazard guessing. Of the remaining 48, 33 were female and 15 were male. Their ages ranged from 21 to 68 years (median 37.5) They were part of a panel that did questionnaires for pay. In general they were typical of the U.S. population in median education and income (as determined from other studies of the same panel). The questionnaire began:

> Drug evaluation
>
> Imagine that you have been suffering for about a month from hour-long periods of nausea about four times a week. Over-the-counter medications are not very effective, so your quality of life is noticeably diminished. You go to the doctor, and she tells you that the condition is not dangerous and will not get any worse. But there is a pill she can prescribe that, if taken daily, will relieve your nausea most of the time. The only side effect of the medication is an occasional itchy rash on your arms. And your insurance will only cover part of the cost. The prices we give are what you pay after insurance.
>
> You will be asked to rate the desirability of several versions of this drug. The versions differ by price ($20/month, $40/month, or $60/month), reliability (relieves nausea 70%, 80%, or 90% of the time it might occur), and duration of side effect (itchy rash on arms for one, four, or seven days per month).
>
> In one section of the study, all three of these attributes (price, reliability, and side effects) may vary from item to item. In the other three sections, only two attributes will vary, while one remains constant.
>
> The sections will be presented to you in a random order.
>
> Your task is to rate each drug on a 9-point scale from Awful to Excellent. There are 54 screens.

The study had four parts. One part contained all 27 combinations of the levels of all 3 attributes. For example, a page might read:

> Price: $20/month
>
> Reliability: effective 90% of the time
>
> Side Effects: itchy rash on the arms seven days a month

This was followed by a nine-point rating scale from “Awful” to “Excellent.” Each of the other three parts contained the nine combinations of the levels of only two of the attributes while the third attribute remained constant at its middle level. These pages begin with a statement of the form: “For all the items in this section, the drug causes an itchy rash on the arms four days a month.” The four groups appeared in an order chosen randomly for each subject. Similarly, the versions of the drug in each group appeared in a random order chosen for each subject.

Table 1: Parameters for regression of difference between weights on their sum.

Attribute	Parameter	t-value	p
price	slope = $-.18$	-3.5	.0010
	intercept = .10	2.5	.0170
reliability	slope = $-.34$	-4.8	.0000
	intercept = .22	4.5	.0000
side-effects	slope = $-.15$	-2.8	.0077
	intercept = .12	2.7	.0096

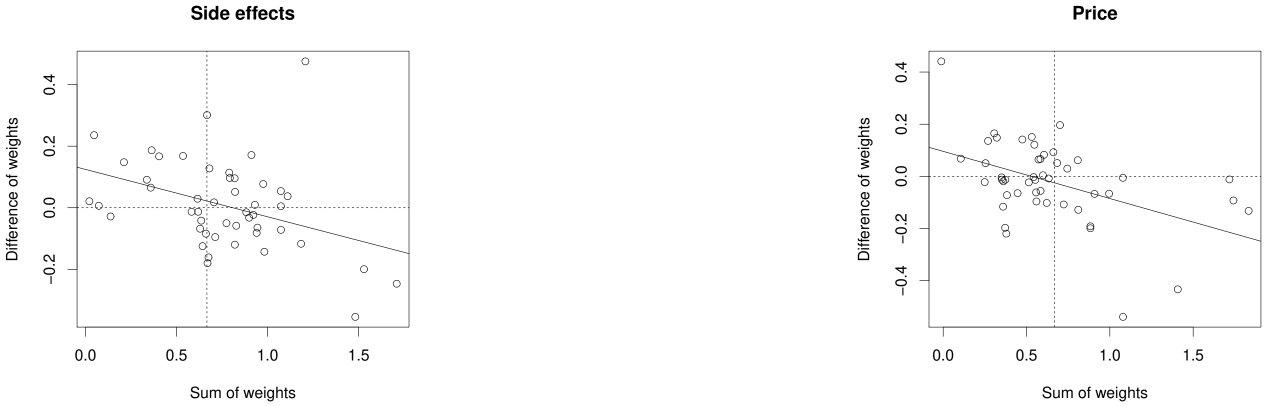

Figure 1: Regression of the difference between the two- and three-attribute weights for each attribute on their sum. Each circle represents one subject. The horizontal dashed line is the null hypothesis that weights are unaffected by the number of attributes. The vertical dashed line is the sum that would represent equal weighting of the three attributes. The solid line is the best fitting regression line.

4.2 Results

We used conjoint analysis to determine the weight of each attribute for each subject in two- and three-attributeal sections. Specifically, we transformed the three levels of each attribute, and the rating responses, so that the rating was as close as possible to a weighted linear sum of the three attributes, using alternating conditional expectations (the ACE package in R: Spector, Friedman, Tibshiranim, & Lumley, 2009; R Development Core Team, 2009; see Gurmankin Levy & Baron, 2005, for a similar use of this approach). The raw weight was the difference between the estimated maximum and minimum scale values. We then rescaled the weights so that they summed to one (i.e., each weight roughly represents the fraction of the raw ratings that resulted from the subject's considering that attribute).

For each attribute, we regressed the difference between each subject's two- and three-attribute weights onto the sum of those weights. Figure 1 plots the regressions for price, reliability, and side-effects respectively. A negative slope and positive intercept indicated that subjects gave less weight to their least important attribute in the three-attribute section than they did in the two-attribute section. Table 1 shows that for all three attributes, the slope was negative, and the intercept was positive. All results were significant at $p < .01$, except for the intercept of price ($p = .017$).

We also did a regression within each subject, with the three differences (2-attribute minus 3-attribute) as the dependent variable and the rank of the three sums as the predictor. The slope is, in essence, the difference for the most "important" attribute minus the difference for the least important attribute. Consistent with the hypothesis, the mean of the slopes was negative, -0.06 ($t_{47} = -3.07$, $p = 0.0036$).

These results shed light on previous findings showing that weights derived from multi-attribute judgment tasks like those used in conjoint analysis are more variable, with more very low weights, than weights derived from direct judgments. This discrepancy alone does not tell us which method is better, only that they disagree. It is possible that direct judgments provide weights that are two close to the middle of the scale, as implied by Parducci's range-frequency theory (Parducci, 1965; and see Baron, Wu, Brennan, Weeks, & Ubel, 2001, for supporting evidence). The present evidence suggests that the problem is at least in part in the conjoint-type tasks, when more than two attributes are used.

However, even two-attribute tasks may show range effects on rates of substitution. Or the weights derived from using all possible pairs may be internally consistent. An additional analysis suggested that internal inconsistency may remain, even in the two-attribute tasks. For each subject I classified the attributes as largest, middle, and smallest according to their effect on ratings as determined by their overall weights (using ACE). I tested the null hypothesis:

$$log[(largest/middle) * (middle/smallest)/(largest/smallest)] = 0$$

Each of three ratios was based on the two-attribute trials using just the two relevant attributes. The null hypothesis expresses consistency. The alternative hypothesis of interest is that subjects tend to exaggerate what should be small differences in attribute weights, thus giving the lower-weighted attribute less weight than consistency would imply, and/or if subjects minimize large differences, giving the lower-weighted attribute more weight when the difference is largest. The null hypothesis was clearly rejected in the direction of the alternative (mean 0.72; $t_{37} = 3.72$, $p = .0007$, eliminating subjects with negative weights). A plausible account of this result is that some subjects exaggerate small differences (the first account just provided) because they base their response largely on a *single* attribute. Examination of the data found many cases of zero or near-zero weights for one attribute (which, together with random variation, explains the existence of some negative weights).

In sum, if the present results are replicated — and they need to be replicated — conjoint analysis suffers seriously from problems of attention.

5 Direct trade-offs of two attributes

A third type of judgment also suffers from internal inconsistency, the judgment of the relative utility of two intervals. In health contexts, respondents are often asked to evaluate some condition, such as blindness in one eye, on a scale anchored at normal health and at death. Implicitly, they are asked to compare two intervals: the difference between being normal and being blind-in-one-eye, and and the difference between normal and death. What happens when we change the standard, the second interval? Normatively, the judgment should change in proportion. For example, keeping normal at one end of each attribute, the utility of blind-in-one-eye relative to death should be the product of two other proportions: the utility of blind-in-one-eye relative to blindness (in both eyes), and the utility of blindness relative to death. In fact, people do not adjust sufficiently for changes in the standard, just as they do not adjust sufficiently for changes in the magnitude of other attributes involved in other judgments of tradeoffs. People show "ratio insensitivity" (Baron & Ubel, 2002).

5.1 Direct trade-offs with explicit ranges or no ranges

Undersensitivity to range can be reduced. Fischer (1995) found complete undersensitivity to range when respondents were asked simply to assign importance weights to ranges, but they were more sensitive to the range, with their weights coming closer to the required doubling with a doubling of the range, when the question was asked in a way that called attention to both ends of each relevant range, e.g., "the difference between 5 and 25 vacation days is what proportion of the difference between $25,000 and $35,000" (for evaluating a job).

In the direct tradeoff method, the range is given for one attribute only. The task is thus analogous to a free-response CV judgment, so we might still expect some insensitivity, regardless of the direction of matching, as found by Fischer and by Baron and Greene (1996, expt. 7).

Baron and Greene (1996, expt 11) found that this insensitivity could be reduced still further by giving no specific ranges for *either* attribute. Respondents were asked to produce two intervals, one on one attribute and one on the other, that were equally large in utility. For example, instead of asking "How much would you be willing to pay in increased taxed per year to prevent a 10% reduction in acquisition of land for national parks?", the two-interval condition asked subjects to give a amount of taxes to be increased [reduced] and a percentage of reduction [increase] in land acquisition that they would find equivalent. Then subjects did this again with higher or lower amounts for each attribute. Insensitivity to magnitude essentially disappeared.

5.2 Double gamble

One method commonly used is the "standard gamble," in which the subject is asked to compare a gamble with a certain outcome. For example, at what probability P would you be indifferent? This task suffers from at least two biases: the certainty effect, and the tendency to consider the certain outcome as a reference point, hence thinking of the gamble as a gain versus a loss (with loss aversion making the loss seem worse: Hershey & Schoemaker, 1986).

In the "double gamble" method, a possible solution, subjects are asked to set one probability to match another one, when both are less than 1 (Pinto-Prades & Abelllán-Perpiñán, 2005). For example, stroke treatment A results in a 25% chance of death and a 75% chance of complete recovery. Treatment B results in a P% chance of permanent brain damage and a 1−P chance of complete recovery. At what probability P are the two treatments equally attractive? Still, the subject is doing a matching task and may tend to give a response that is undersensitive to the quantity of the attribute to be matched, as just discussed. Asking the subject to set both attributes might help.

5.3 Unit-price evaluation

A second solution is to ask subjects for the WTP per unit of the good. This gives a direct answer to the question of interest (most of the time) and insures proportionality, in a trivial way. Baron and Greene (1996, expt. 8), gave subjects items like "Suppose there are small amounts of a cancer-causing chemical in your drinking water. The type of cancer has a 50% cure rate. The chemical can be removed by a filter, which must be replaced once a year." In the unit-price condition, subjects first indicated how much they would pay per year "*for each death prevented out of 1,000,000* for a filter that removes the chemical if the rate of cancer caused by the chemical is 10 per million people who drink the water?" Then they were asked how much they would pay for the device, and they were given a hint that they should multiply their last answer by 10. Finally, they were asked how much they would pay if the rate were 1 per million.

The first and third answers did not differ significantly, indicating that subjects were happy with proportional responses. Although this study is encouraging, we still need to test for sensitivity to the size of the unit! (For example, the filter could provide protection for 5 years instead of 1, but this is perhaps not the best example.)

6 Attention to irrelevant information

In CV studies, people ask "How much do [double hulled tankers] cost?" That is beside the point, which is the value of preventing oil spills. People often confuse value and cost. (For market goods, they should be correlated, but for public goods we have no reason to think that they are correlated.)

Baron and Maxwell (1996) asked students how much they would pay in higher tuition for reduction in the rate of violent crime on campus from 50 to 25 crimes per year, for 50,000 students. When the subjects were told that this would involve an increase of the number of campus police from 50 to 100, their geometric mean WTP was $169. When they were told that the increase would be from 50 to 200, the geometric mean was $247.

It seems obvious that, in this case, they would ignore the irrelevant information if we did not provide it, and we have no reason to think that the subjects would regard it as crucial in this case. More subtle cases are those that involve decision biases. If we remove the information that triggers the bias, would people think we had tricked them by removing something necessary for a valid judgment? The next section reports an experiment in which the answer is not obvious.

7 Experiment 2: Debiasing by removing distracting information

To elicit judgments, we may need to abstract the essentials, removing realistic context that can lead to biases, such as whether an outcome is caused by nature or human activity. (But we can also include emotions as outcomes.)

This may mean giving our subjects *less* information when they make judgments for us. I report here an experiment looking at the effects of removing information on four non-utilitarian biases in judgments about allocation of goods. Items are first presented in the standard condition, then immediately in a stripped-down condition that concerns consequences only. We can expect some bias to carry over, but we can also ask whether biases are reduced. Finally, subjects indicate whether they thought the stripped-down version is a fair summary.

The experiment concerned four biases. Omission bias is the judgment that harmful actions are worse than omissions that are even more harmful (Ritov & Baron, 1990; Spranca, Minsk, & Baron, 1991; see Baron & Ritov, 2009, for a recent review).

What I shall call zero-bias is the bias toward complete solutions, i.e., zero risk or zero harm (Baron, Gowda, & Kunreuther, 1993; Ritov, Baron, & Hershey, 1993). For an extreme example, people would rather reduce a risk from 10% to 0 than from 30% to 10%. This bias is related to the proportionality bias, in which people judge risk reductions by proportion rather than absolute change (Fetherstonhaugh et al., 1997; Baron, 1997b); reduction to 0 amounts to complete (100%) reduction, even if the risk is small.

The ex-ante bias is the finding that people want to equate ex-ante risk within a population even when the ex-post risk is worse (Ubel, DeKay, Baron, & Asch, 1996a). For example, many people would give a screening test to everyone if the test would prevent 1000 cancer cases rather than give a test to half (picked at random) that would prevent 1200 cancer cases.

The equality bias is the preference for equal treatment of two groups, even when unequal treatment would be better on the whole (Baron, 1995; Ubel, DeKay, Baron, & Asch, 1996b). For example, people would rather help 50% of each of two equal-sized groups than 80% of one group and 40% of the other.

In a "minimal" presentation condition, the scenario was reworded so as to emphasize the consequences. A "minimal-test" condition was included to determine whether subjects thought that the minimal description was fair.

7.1 Method

Ninety-five subjects completed a questionnaire on the World Wide Web. Their ages ranged from 18 to 62 (median 38); 34% were male; 15% were students. The questionnaire began:

> **Health insurance**
>
> This questionnaire concerns decisions made by health insurance companies about which treatments to

> cover.
>
> On each screen, you will see some information about two treatments for serious conditions. The two treatments have the same cost, which is high. An insurer cannot afford to cover both treatments, so it chooses one.
>
> All conditions are chronic, making for a low quality of life and usually a shorter life too. Examples of such chronic conditions are severe arthritis, senility, emphysema, Parkinson's condition, and heart disease.
>
> The questions just talk about "conditions" without specifying which conditions. Imagine that the questions refer to serious conditions, that are all **equally** serious.
>
> Sometimes the treatment leads to a different condition as a side effect. This different condition is just as serious as what the treatment cures: no more, no less. . . .

The 32 screens were presented in a different random order for each subject. They constituted a 4x4x2 design: type of bias (omission, zero, ex-ante, equality), type of de-biasing (control condition, minimal, minimal-test, and a fourth condition that will not be discussed here[3]), and action of the company (two levels, choosing the option with the best consequences, the "better" option, or the other option). As an example, the omission bias case was [with comments in brackets]:

> Treatment A cures 50 people out of 100 who come in with condition X each week, and it leads to no other conditions.
>
> Treatment B cures 80 of the people with condition X, but it leads to condition Y (randomly) in 20 of the 100 patients. X and Y are equally serious.
>
> [added for minimal de-biasing]
>
> In other words, treatment A leads to 50 people with condition X and nobody with any other condition, and
>
> treatment B leads to 20 people with condition X and 20 people with condition Y (which is equally serious).
>
> Which treatment should the company choose?

Certainly A	Probably A	Probably B	Certainly B

> The company chose treatment A [B on half the trials].
>
> Would this choice make your more likely or less likely to choose this company as your insurer?

More likely	Probably more	Probably less	Less likely

> [In minimal-test, the following was added.]

[3] The fourth condition involved expansion of the scenario with extra information, rather than removal of information.

A critic of the company argues against the company's decision by pointing out that the consequences were worse. The critic says that the decision amounts to a very simple choice:
treatment A leads to 50 people with condition X and nobody with any other condition, and
treatment B leads to 20 people with condition X and 20 people with condition Y (which is equally serious).

[Or the following, if the company made the choice with the better consequence.]

The company argues for it's decision by pointing out that the consequences were better. The company says that the decision amounts to a very simple choice: . . .

[The wording is identical to minimal de-biasing]

Is this a fair summary of what the decision is about?

Yes, completely	Basically yes	Not really	Not at all

The basic form of the other three biases was as follows:

Zero. X and Y are two forms of a condition.

Treatment A can be given to 100 people with form X who come in each week, and it cures 60 of them.

Treatment B can be given to 50 people with form Y, and it cures all 50 of them.

Equality. A condition has two forms. Each week, 100 people come in with form X, and 100 with form Y.

Treatment A cures 80 of the 100 with form X and 40 of the 100 with form Y.

Treatment B cures 50 of the 100 with form X and 50 of the 100 with form Y.

Ex-ante. Treatment A can be given to 100 patients with a condition each week, and it cures 30.

Treatment B is in short supply, so it can be given only to 50 patients picked at random each week. It cures 40 of these 50.

7.2 Results

7.2.1 Biases

The items were designed so that one answer was optimal. Choice of the other answer could result from random error (misreading, or making an unintended response) as well as bias. However, we can at least compare the four different biases. Table 2 shows the frequency of each of the response options, with "worst" and "best" replacing A and B, according to consequences. It is apparent that the omission and zero-risk biases were stronger than the other two, where the "worst" response was rarely chosen. The table includes responses from all conditions, however, including the de-biasing conditions.

Table 2: Frequency of responses, in percent, for question about what to do in the first experiment ("worst" and "best" indicate consequences).

	Certainly worst	Probably worst	Probably best	Certainly best
Choice question				
Omission	15.7	37.5	42.5	4.3
Zero	14.3	30.0	37.5	18.2
Ex-ante	2.5	14.9	51.4	31.2
Equality	2.4	10.1	61.7	25.8
Trust question				
Omission	16.6	34.6	41.7	7.1
Zero	13.6	28.9	39.5	18.0
Ex-ante	5.0	13.2	55.3	26.6
Equality	2.9	12.1	59.3	25.7

Table 3: Mean rating (-1.5 to 1.5, with 0 indicating neutrality) as a function of type of bias and de-biasing condition. Positive numbers favor the optimal choice.

	Control	Minimal-test	Minimal
Choice question			
Omission	-0.12	-0.16	-0.23
Zero	0.08	0.11	0.30
Ex$-$ante	0.56	0.58	0.71
Equality	0.60	0.66	0.66
Trust question			
Omission	-0.12	-0.12	-0.14
Zero	0.12	0.12	0.28
Ex-ante	0.52	0.46	0.63
Equality	0.60	0.62	0.57

7.2.2 De-Biasing effects

In general, the minimal de-biasing manipulation reduced bias. Table 3 shows the mean ratings on a scale on which 0 is neutrality and each step is 1. We compared each de-biasing manipulation to the combined results of the control condition and the minimal-test condition, which was identical to the control condition up to the point of the argument question. The minimal-test and control conditions did not differ significantly. Also, we combined the trust question and the choice question; these did not differ either.

The ratings for the minimal manipulation were significantly higher than the (combined) control condition ($t_{94} = 2.80$, $p = 0.0062$, two tailed). For individual biases, only the effects on zero and ex-ante were significant ($t_{94} = 3.87$, $p = 0.0002$, and $t_{94} = 3.22$, $p = 0.0018$, respectively). The result for the ex-ante bias suggests that this bias is in fact present, even though ratings were generally high.

Table 4: Frequency of responses, in percent, for question about whether the minimal condition was a "fair summary."

	Yes, completely	Basically yes	Not really	Not at all
Omission	20.0	50.0	25.3	4.7
Zero	30.5	38.9	25.8	4.7
Ex-ante	35.8	48.9	14.7	0.5
Equality	30.5	52.1	13.7	3.7

7.2.3 Acceptance of arguments

The effectiveness of the minimal manipulation suggests that at least people saw the summary argument in terms of consequences as fair. Table 4 shows the ratings for the four bias conditions. A majority of subjects seemed to think that the summary was fair, although fewer thought so in the omission and zero conditions, where the biases were strongest.

7.2.4 Individual differences in the minimal argument

In general, subjects who thought the argument in the minimal condition was more fair were those who were less biased overall ($r = .59$ across all biases, combining choice and trust, $p = .0000$). This correlation was large and highly significant for each of the biases except for omission ($r = .16$).

These correlation could result from effects of prior beliefs on the evaluation of the minimal argument. Or it could result from the de-biasing effect of the argument on the biases in the minimal condition (and perhaps transferring elsewhere), or both. To assess the effect of the argument on the bias, I regressed the bias in the minimal condition on the bias in the (combined) control conditions and the evaluation of the argument, combining all biases and both choice and trust questions. Although the effect of the control condition was highly significant ($t_{92} = 12.25$, $p = .000$), the effect of argument evaluation was not significant ($t = 0.42$). In sum, we have no evidence that individual differences in the perception of argument fairness lead to individual differences in the effect of the minimal de-biasing manipulation. Rather, the evaluation of the argument seems to be affected by prior bias.

8 Discussion

Elicitation of values is sometimes necessary. We cannot always compare benefits of regulation by using objective quantities like money. Money has different value (utility) to different people, and some goods, like wilderness, have no market price that we can use to assess their monetary value.

The most promising approach to utility measurement is to stick to consequences. Even though people care about other things, such as whether consequences are reached through acts or omissions, it is helpful to know how our policies will affect those consequences, and the value people place on them, regardless of what those policies might be (e.g., whether they involve acts or omissions).

Utility measurement is difficult, but I have tried to outline some promising approaches. We can focus on the consequences by eliminating much of the detail about means that is often included in valuation surveys. People do not object strongly even when the means consist of factors that they consider morally relevant (and the objections that I did find might be reduced further if the full elicitation condition, with means as well as ends, were not presented first). Although contingent valuation (CV) is usually insensitive to quantity, sensitivity can be improved by asking people their willingness to pay *per unit*, or by asking them to produce both a monetary amount and a quantity of the good.

Another promising approach is conjoint analysis. But, before we can rely on this, we must make sure somehow that people are attending to all the relevant attributes that vary, rather than using a simplifying heuristic such as relying on a single attribute. The idea is to measure trade-offs. This might be accomplished by asking people to rate each attribute before rating the overall item, and by using two attributes at a time.

8.1 Decision analysis vs. democracy

The use of decision analysis for regulation seems to conflict with democracy, especially when we limit the analysis to consequences. It seems paternalistic for government to over-ride what people want. Yet this is a special kind of paternalism. Ordinary paternalism, like banning marijuana to protect people from harming themselves by using it, or preventing people from giving up their right to sue their doctors in return for lower insurance fees, is justified largely in terms of the good of individuals who, presumably, would otherwise make decisions that undercut their own good. In other cases, like restrictions on smoking, government control is also justified by externalities, effects on others. But paternalism in regulation is about policies concerning public goods. What is being overridden here may well not involve individual preferences (utilities) but rather something more like opinions. Of course, once citizens form opinions about what government should do, they come to prefer that government accept their opinion, so they have a utility for government policy. But this utility may not be primary. It may be more labile, more dependent on beliefs that might be wrong, than more basic and stable preferences.

More generally, democracy does not guarantee optimal outcomes, especially when it is combined with misunderstanding and cognitive biases on the part of citizens (Baron, 2009; Caplan, 2007; Hirschleifer, 2008). And neither does decision analysis. Like any tool, it can be used incorrectly and is not very helpful for some kinds of problems. When majority opinion and decision analysis disagree, that is a sign that both need further probing, to find out why.

Just as governments override individual preferences when it regulates drug use or prohibits the selling of kidneys, governments are set up to make decisions in terms of their own analysis of what is best. Democracy — even in Switzerland, where referendums are common — is mostly representative. We elect politicians, who then make laws and supervise government agencies staffed by career employees as well as political appointments. The agencies, in turn, try to carry out their mission. The politicians might be satisfied even if they do not understand everything that the government does, so long as they are convinced that it is trying to do the right thing as best it can. This argument is consistent with the arguments made by Breyer (1993) and Sunstein (2002), who advocated heavier reliance on experts in risk regulation. An example is the U.S. Federal Reserve, which has taken the issue of "tight money," a preoccupation of politicians in the 19th century, almost completely out of politics. Government based on expertise can work if it builds trust over the long run, even if citizens are sometimes puzzled by its decisions.

It is also difficult to imagine how any formal method of analysis would dictate policy by itself. The Federal Reserve surely relies on economic models and predictions, but it surely also tweaks the policies that the models imply. The kind of cost-benefit analysis that I recommend here can provide a clear prediction of the utility of consequences. The implied policies can be overridden by politicians if not by government officials. Yet, if the government does something else, at least it has a second opinion, and knows what it is doing.

8.2 Taking inequality into account

One of the ways in which policy makers want to tinker with formal analyses is that the analyses often seem unfair in the distribution of benefits and burdens. Regulation of arsenic in drinking water, for example, may be beneficial on the whole but may impose costs on the poor that, for some of them, outweigh the benefits (Sunstein, 2002). (They might do better to spend the money on other ways of reducing risk or improving health.)

A traditional argument is that specific regulations and legal rules should not be used as a tool of redistribution (e.g., Kaplow & Shavell, 2002) because such tools are crude, and it is better to use progressive taxation and direct subsidies for redistribution, at least because these may be more precisely targeted. Yet, the urge to tinker with specific laws and regulations to improve their distributional fairness seems irresistible, at least because regulators, judges, and legislators become impatient waiting for the redistributive Utopia that never seems to arrive (arguably, in part, because the legislators don't allow it to arrive).

As I suggested in Experiment 2, some of the preference for fairness may be the result of a bias. (See also Baron, 2008.) In other cases, though, it is a simple recognition of a distributional problem, as in the arsenic case just described. Some have argued for an explicit correction of the results of cost-benefit analysis, in order to deal with such issues. A correction would clearly be needed if the analysis is done entirely in terms of monetary values, as the poor have a higher utility for money than the rich. But a correction may not be needed if different populations are considered as part of a decision analysis based on utility rather than money.

Specifically, it might be possible to elicit from respondents different utilities as a function of rich vs. poor, for example:

> John is single and lives in lower Manhattan. He makes $200,000/year and his 2-bedroom apartment has 15,000 square feet, and a garage for his car.

> Jake, also single, lives on the outskirts of Capetown, South Africa, in a house with 2 rooms and 5,000 square feet. His job pays $5,000/year (at the current exchange rate), which enables him to have a TV and refrigerator, something many of his neighbors lack.
>
> Consider the effect of changes in income on John and Jake. For example, who would be affected more by a change in income of $1,000? Now give two numbers, a change for John and a change for Jake, around their current income. Choose the numbers so that the effect on the two would be the same, in terms of achievement of their goals.

Such a method, using the double-response approach described in section 5.1, might elicit a judged utility function for money as a function of current income. Of course, it is based on judgments, but any method of correcting a cost-benefit analysis would also be based on judgments.

8.3 A short note on methodology

Most CV surveys use large stratified samples, and the number of critical responses collected from each respondent is small, sometimes even a single WTP response. Traditionally, decision analysis is done with very few respondents, who may be interviewed for several hours each, over several days. These interviews often elicit utility functions that are then checked and corrected so that they are internally consistent (Keeney & Raiffa, 1993).

The results I have cited here suggest that the error in typical large surveys is large and systematic, e.g., the error that results from neglect of quantity. These surveys may still be better than nothing. But the use of large samples seems to be false precision. Large samples reduce the error of estimation of the population mean, but that error may be very small relative to the systematic error resulting from flaw in the method itself. It may be better to use smaller numbers of respondents and to spend more time with each one doing the kind of follow-up checking that decision analysts typically do. Even with a sample of 100, it is still possible to get a variety of respondents so that important effects of gender and income, for example, would be detected.

References

Baron, J. (1993). *Morality and rational choice*. Dordrecht: Kluwer.

Baron, J. (1995). Blind justice: Fairness to groups and the do-no-harm principle. *Journal of Behavioral Decision Making, 8*, 71–83.

Baron, J. (1996). Norm-endorsement utilitarianism and the nature of utility. *Economics and Philosophy, 12*, 165–182.

Baron, J. (1997a). Biases in the quantitative measurement of values for public decisions. *Psychological Bulletin, 122*, 72–88.

Baron, J. (1997b). Confusion of relative and absolute risk in valuation. *Journal of Risk and Uncertainty, 14*, 301–309.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

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., & Maxwell, N. P. (1996). Cost of public goods affects willingness to pay for them. *Journal of Behavioral Decision Making, 9*, 173–183.

Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral Judgment and decision making*, Vol. 50 in B. H. Ross (series editor), *The Psychology of Learning and Motivation*, pp. 133–167. San Diego, CA: Academic Press.

Baron, J., Wu, Z., Brennan, D. J., Weeks C., & Ubel, P. A., (2001). Analog scale, ratio judgment and person trade-off as utility measures: biases and their correction. *Journal of Behavioral Decision Making, 14*, 17–34.

Baron, J., & Ubel, P. A. (2002). Types of inconsistency in health-state utility judgments. *Organizational Behavior and Human Decision Processes, 89*, 1100–1118.

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.

Breyer, S. (1993). *Breaking the vicious circle: Toward effective risk regulation.* Cambridge, MA: Harvard University Press.

Brown, R. V. (2006). Making decision research useful — not just rewarding. *Judgment and Decision Making, 1*, 162–173.

Corso, P. S., Hammitt, J. K., & Graham, J. D. (2001). Valuing mortality risk-reduction: Using visual aids to improve the validity of contingent valuation. *Journal of Risk and Uncertainty, 23*, 165–184.

Deber, R. B., & Goel, V. (1990). Using explicit decision rules to manage issues of justice, risk, and ethics in decision analysis. *Medical Decision Making, 10*, 181–194.

Ebbesen, E., & Konecni, V. (1975). Decision making and information integration in the courts: the setting of bail.

Journal of Personality and Social Psychology, 32, 805–821.

Fetherstonhaugh, D., Slovic, P., Johnson, S., & Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty, 14*, 283–300.

Fischer, G. W. (1995). Range sensitivity of attribute weights in multiattribute value models. *Organizational Behavior and Human Decision Processes, 62*, 252–266.

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., Slovic, R., & Rorer, L. (1968). An analysis of variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin, 69*, 338–349.

Kahneman, D. & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management, 22*, 57–70.

Kaplow, L., & Shavell, S. (2002). *Fairness versus welfare.* Cambridge, MA: Harvard University Press.

Keeney, R. L. (1992). *Value-focused thinking: A path to creative decisionmaking*. Cambridge, MA: Harvard University Press.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preference and value tradeoffs.* New York: Cambridge University Press. (Originally published, 1976.)

Louviere, J. J. (1988). *Analyzing individual decision making: Metric conjoint analysis.* Newbury Park, CA: Sage.

Lynch, J. G., Jr., Chakravarti, D., & Mitra, A. (1991). Contrast effects in consumer judgments: Changes in mental representation of in the anchoring of rating scales. *Journal of Consumer Research, 18*, 284–297.

Mellers, B. A., & Cooke, A. D. J. (1994). Tradeoffs depend on attribute range. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1055–1067.

NOAA (National Oceanic and Atmospheric Administration). (1993). Report of the NOAA panel on contingent valuation. *Federal Register, 58* (10), 4602–4614.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 401–418.

Pinto-Prades, J.–L., & Abelllán-Perpiñán, J.–M. (2005). Measuring the health of populations: The veil of ignorance approach. *Health Economics, 14*, 69–82.

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. `http://www.R-project.org`.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*, 263–277.

Slovic, P. (1969). Analyzing the expert judge: a descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology, 53*, 255–263.

Spector, P., Friedman, J., Tibshiranim, R. & Lumley, T. (2009). acepack: ace() and avas() for selecting regression transformations. R package version 1.3-2.2.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.

Sunstein, C. R. (2002). *Risk and reason: Safety, law, and the environment.* New York: Cambridge University Press.

Tengs, T. O., Adams, M. E., Pliskin, J. S., Safran, D. G., Siegel, J. E., Weinstein, M. E., & Graham, J. D. (1995). Five-hundred life-saving interventions and their cost-effectiveness. *Risk Analysis, 15*, 360–390.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95*, 371–384.

Ubel, P. A., DeKay, M. L., Baron, J., & Asch, D. A. (1996a). Cost effectiveness analysis in a setting of budget constraints: Is it equitable? *New England Journal of Medicine, 334*, 1174–1177.

Ubel, P. A., DeKay, M. L., Baron, J., & Asch, D. A. (1996b). Public preferences for efficiency and racial equity in kidney transplant allocation decisions. *Transplantation Proceedings, 28*, 2997–3002.

Volk, M., Lok. A. S., Ubel, P. A., & Vijan, S. (2008). Beyond Utilitarianism: A method for analyzing competing ethical principles in a decision analysis of liver transplantation. *Medical Decision Making, 28*, 763–772.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research.* New York: Cambridge University Press.