Various forms of analysis are used to evaluate public-policy options. In many cases, such as risk reduction or preservation of the natural environment, part of the input involves the personal value of outcomes to people, in ways that cannot easily be captured by the assignment of monetary values. The analysis requires judgments from respondents to surveys. One method is "contingent valuation" (CV) in which the respondents are asked how much money they would be willing to pay for some public good (such as reduction in the risks of arsenic pollution). Another method is "conjoint analysis" in which each respondent rates a series of options differing in a few attributes (such as cost and amount of pollution reduction).

These methods, and others, have serious flaws. For example, CV yields judgments that are usually not sufficiently sensitive to the amount of the good (e.g., magnitude of pollution reduction). Conjoint analysis solves this problem, but it seem to lead to underestimation of the weights of the less important attributes because people attend only to the most important one. The proposed research will explore the nature of these problems and several promising approaches to correcting them. For example, CV may be improved by asking for judgments of both dimensions (magnitude of the good and the amount that it would be worth paying for that magnitude). Conjoint analysis may be improved by asking for judgments of the effect of each attribute on overall value.

In addition, the research will explore the nature of judgments when people are asked about paying through taxes vs. other ways such as voluntary donations. In particular, in answering such questions, where do people think their duty lies: defending their self-interest? or trying to do what is best on the whole? I shall also examine willingness-to-pay as a function of the type of good and whether the payment is voluntary or (through taxes) compulsory. I will examine people's reasoning about these issues: for example, do they see compulsory contributions (taxes) for public goods such as pollution reduction to be more justified than compulsory contributions for redistribution? This part of the research will help to arrive at a more general understanding of how people think of their roles as citizens in a democracy.

Finally, I will explore the nature of preferences for equitable distributions of goods, with a view to the implications for cost-benefit analysis. In general, cost-benefit analysis is a potential application of all the work proposed.

1 Results of prior NSF support from NSF 0213409: Inconsistency and bias in thinking about tax reform

Work on this grant (with Edward McCaffery) continues, even though the funding finally had to stop (after two extensions). We have published several papers (some cited below), and we have a draft of a book. Here I summarize some overall themes, with special attention to work that is relevant to the present proposal.

The general issue is how citizens think about tax. Ultimately, in a democracy, the stability of a tax regime must have sufficient public support. Much of our research has concerned framing effects, that is, responses that would lead to different outcomes depending on what question is asked. For example, McCaffery and Baron (2003) asked subjects to design a single, global tax system or to vary one component of a tax system (payroll or income tax) with the other component held constant. The idea was to replicate the effects of income tax reform given a constant payroll tax system. Subjects focused on the component they were asked to manipulate (income tax) and did not respond fully to changes in the other component (payroll), across conditions, reflecting an underadjustment bias as well as a framing effect. Similarly, in McCaffery and Baron (2004), people could be induced to focus on different aspects of the "marriage penalty". In general, if the questions were asked in ways that made the issue salient, people favored "couples neutrality" (equal taxes for couples with the same total income), "marriage neutrality" (marriage of two people should not affect total taxes they pay), and graduated taxes (higher percent tax for higher income). The laws of arithmetic imply that at most two of these principles can be true at once. (The U.S. violates marriage neutrality with a penalty or a bonus, depending on the relative incomes of the pair.)

More relevant to issues described in this proposal, Baron & McCaffery (2006; and McCaffery & Baron, 2006b) found the same effects in judgments about redistribution through taxation. People tended to view the progressiveness of the tax system in isolation. Their judgments of appropriate taxation of different incomes were relatively insensitive to changes in other factors that would affect the distribution of income, such as the use of tax deductions (which favor those with higher marginal rates) or the privatization of services. Also, we observed wide individual differences in preferences for progressiveness, with some favoring a flat tax and others wanting to use the tax system to equate all after-tax income.

Later papers emphasized the "isolation effect", a general term for viewing decisions in isolation, focusing on what is "on screen" and ignoring outside factors that are relevant but not salient. McCaffery and Baron (2006a) showed how this effect could lead to a preference for hidden taxes such as those on business. When people were prompted to think about who actually pays such taxes, their preference for them was reduced, as if they originally viewed the tax as not paid by any actual person.

The work most relevant to the current proposal is Baron and McCaffery (2008), which was titled "Starving the beast" because it was inspired by what was thought to be a conscious strategy pursued by Ronald Reagan (and later by George W. Bush) of cutting taxes as a means to reduce the size of government, in hopes that large deficits will result in later budget cuts. Historically, whatever the intention, the approach did not work: spending was not reduced. We asked subjects to chose general levels of taxation and public spending from various hypothetical starting points. Subjects wanted to reduce both taxes and spending, preferring balanced budgets and even surpluses to deficits. But they

were reluctant to make *specific* spending cuts, even though they were given all possible categories in which cuts could be made, while still wanting cut taxes. In this proposal I suggest that a similar phenomenon will be observed in studies of the valuation of public goods.

2 Proposed research

This proposal is a revision of an earlier proposal with the same title, which was not funded. The major criticisms, and a summary of the changes made to address them are as follows (and the most relevant sections of the proposal are marked with asterisks):

• *Reliance on a single panel of subjects could be misleading.* I will recruit additional samples at the end of the research to test any improvements that result from it, and I will replicate with my panel studies done on other samples.

• *There is no attempt to validate methods externally.* I propose to test both attribute weighting and quantity sensitivity using external criteria. (I thank reviewer #6 for this criticism.)

• *The research relies on ratings, yet discrete choice is more popular these days.* I shall test results with discrete choice.

The research combines two related themes. One (also pursued in the work on tax just described) examines how citizens think about public policies. In a democracy, this thinking helps to determine what policies are put into effect. If the thinking involves cognitive biases that lead people to favor non-optimal policies, then, other things being equal, the policies adopted will tend to be less optimal than they could be without the biases. This approach has been a major interest of mine (e.g., Baron & Jurney, 1993; Baron, 2009, forthcoming), and it is also found in the work of others (e.g., Caplan, 2007; Hirschleifer, 2008).

The second theme is the measurement of values for public decisions, such as regulation of risk or allocation of scarce goods. This was the topic of a much earlier NSF grant proposal ("The measurement and expression of values for public goods," 1993–1995), and several publications (e.g., Baron & Greene, 1996; Baron, 1997a). A widely used method is contingent valuation (CV), in which members of the public are typically asked for their willingness to pay money for some public good such as a reduction in risk from air pollution. The results are used to assess the benefits of policies such as regulations, as part of cost-benefit analysis. This method has serious flaws (Baron, 1997a), and alternatives (or improvements) are needed. The need for improvements became apparent to me at a recent conference on "New ideas for risk regulation," at Resources for the Future, June 22–23, 2009 (http://www.rff.org/Events/Pages/New-Ideas-for-Risk-Regulation.aspx).

The relation between these two approaches is that a value measurement exercise such as CV can be seen as a way for citizens to provide input to government, like voting or other sorts of political action. In some ways, these different research traditions seem to conflict. For example, in CV studies, researchers are often surprised at what seem to be very high willingness-to-pay (WTP) figures for very specific environmental benefits, even when the payment vehicle is a tax increase. Yet, our studies of tax suggest that people want lower taxes overall. The proposed research will extend both approaches and locate other links between them.

2.1 General method

The research proposed will use questionnaires on the World Wide Web. The subjects are a panel of about 1200 adults (at present), most of whom found me by looking for ways to earn money on the Internet. They are almost all U.S. residents. They are quite varied in background and political ideology. Their median age, education, and income is about that of the U.S. Although they are not exactly representative of the U.S. population, the research proposed is not limited to the U.S. in its relevance. The sample is much more varied than college student samples. Usually I test a sufficient number of subjects (80), in within-subject designs, so that I can examine individual differences statistically, correlating responses with each other and with general questions about attitudes.

* Once results are obtained from this panel, I will replicate the most relevant studies on other samples of less experienced subjects selected in other ways. I plan to do this mainly in the last year of the grant, so I have requested extra funds for that year. The availability of other samples may change between now and then, but if I did this now, I would use some of the resources listed in `http://www.sjdm.org/links.html#research`, particularly "Time sharing experiments for the social sciences" (which is "free"), "Knowledge Networks", "Mechanical Turk" (which is inexpensive and draws from a very different population than the one I use, thus allowing a check for robustness of results), and one of the organizations that selects on-line samples.

* In addition, I will test my own panel by replicating, as closely as possible, some relevant results of others using both contingent valuation and conjoint analysis. One example among many of possibly useful studies is that of Cameron and DeShazo (2008) on health risk reduction. Both these replications and those described in the last paragraph would be part of a paper designed to summarize major results for an audience of "practitioners".

For my own panel, I program the questionnaires in JavaScript, which allows considerable flexibility. For example, the order of items is randomized, and it is possible to construct items on the basis of responses to earlier items, to reveal items on a page one by one, use visual displays to collect responses, or reveal information according to the position of a pointer (mouse). Each questionnaire has several items of the same type, so that I can estimate parameters for each subject.

For data analysis, I use R. For many purposes, I have been using the lmer() function (Baayen, Davidson, & Bates, 2008), because it allows crossed random effects for subjects and cases. I also examine individual differences as described in Baron (in press). All data (with identification removed) are available in my web page, with the R scripts used to analyze them, and (separately) the questionnaires that generated them.

2.2 Improvements in contingent valuation (CV)

A major bias in CV is the lack of attention to quantitative 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 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).

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 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 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]

2.2.1 Unit-price CV

A couple of variations in the basic CV method seem to make people more sensitive to quantity. One is to ask people for their willingness to pay *per unit* of a good, e.g., their WTP for a single life saved out of 1,000,000 exposed to a risk. Baron and Greene (1996, expt. 8) found that this measure was essentially the same regardless of whether (for example) the risk reduced was 1/1,000,000 or 10/1,000,000. (Subjects were asked to multiply their response by 10, in the 10/1,000,000 case, to see what their overall WTP would be.)

I propose to extend this method by varying the size of the unit. For example, in the case just given, the risk reduction covered a period of one year. What if it were 5 years? Quantity sensitivity would imply that the WTP should be close to 5 times as much. In other cases, different units could be used, such as inches vs. feet for rise in sea level from global warming. This will further test the internal consistency of the unit-price approach. I shall also ask whether visual representations are helpful in bringing about consistency.

2.2.2 Double-response CV

Baron and Greene (1996, expt. 11) found that insensitivity could also be reduced by giving no specific quantity intervals 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. 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 double-response condition asked subjects to give an amount of taxes and a percentage reduction that they would find equivalent. Then subjects did this again with a higher or lower amount. Insensitivity to magnitude essentially disappeared.

I shall explore this method in ways similar to the last one. If I change the units of one of the measures, are subjects still sensitive, when they adjust the other measure? I shall also combine the two methods, asking for two unit-prices. Although difficult for the subject, this may be the most promising approach overall.

---

[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.

2.2.3 * Validation

The two methods just described are tests of internal consistency. I shall also test the psychological mechanisms involved for prediction, which I assume to be similar to valuation in the cognitive processes involved. It is difficult to do this with ordinary consumption goods because their price is usually *not* a linear function of their quantity. I shall instead use public goods, asking subjects for judgments of actual government expenditures in various categories, as done by Kemp (2002, 2003). For example, I shall ask about spending on Medicare, the number of patients served, and annual spending per patient. These questions will be separated to prevent calculation, and several categories will be used. I shall also ask parallel valuation questions, about how much should be spent in each category.

2.3 Attention and range effects in conjoint analysis

Another solution to the problem of quantity neglect is conjoint analysis. The subject's task is typically to rate several items in a single session, or choose between two items. (I shall use both methods.) 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, or a smaller set if necessary (Louviere, 1988). We fit a model to predict each subject'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 the 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 or choice probability stays the same. We could think of this as a rate of substitution, e.g., substitution between price and repair record of a car.

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) found no effects of relative ranges on rates of substitution with several pairs of dimensions, but did find them when the numerical representation 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. 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 (as Beattie & Baron did). Indeed, for some time it has been suggested that weights derived from such multi-attribute tasks (with several attributes) are more variable than weights derived from direct ratings (von Winterfeldt & Edwards, 1986, section 10.4).

With James Yadavaia, I carried out a preliminary test of the effect of an additional attribute on the weight of the lower-weighted attributes (Baron, draft). We constructed a three-attribute stimulus: price, side-effects, and effectiveness of a drug for relief of periodic nausea. Each attribute had three levels. Each subject rated four sequences of cases on a 1–9 scale. 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 subject knew this in advance), and the two others varied orthogonally; thus, in these sequences the subject needed to attend to only two dimensions 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. The weight was the difference between the ratings corresponding to the highest and lowest attribute levels, after the ratings were transformed by applying to each subject a model that allowed transformations of the stimulus and response scales so as to get the best linear additive fit (ACE, or Alternating Conditional Expectations: Spector, Friedman, Tibshiranim, & Lumley, 2009).

Then we estimated an overall 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 overall 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 found that the *difference* of 2-attribute-weight minus 3-attribute-weight was higher, and positive, when the overall weight (the sum of the weights from the two tasks) was low. This was true both within subjects (comparing the highest ranked attribute to the lowest) and across subjects (correlating the sum of the weights with their difference, for each attribute, across subjects).

In sum, when subjects must consider three attributes, they seem to give a lower weight to the attribute that has (overall, regardless of the number of attributes) less of an effect on their ratings, compared to the weight of that attribute when only two attributes vary at a time.

Even with only two attributes, subjects seem to underweigh the less important of the two. Evidence for this came from a test of internal consistency in the two-attribute task. Consider attributes A, B, and C, such that, overall, A has the highest weight w(A) and C the lowest w(C). If subjects are consistent, $w(A)/w(C) = [w(A)/w(B)] * [w(B)/w(C])$. If, however, subjects attend very little to B when it is paired with A (but attend more when it is paired with C), then $w(A)/w(B)$ will be very high, and $[w(A)/w(B)] * [w(B)/w(C])$ will be higher than $w(A)/w(C)$. In fact, it was much higher, and many subjects appeared to give near-zero weight to the lower-weighted dimension, even in two-attribute tasks.

I propose to extend this experiment, using other sets of attributes, and choice as well as rating. I shall also attempt to examine the cognitive processes used, by recording response times (which I do anyway). It may be possible to predict the magnitude of the effect just described from the effect of number of attributes on response time: subjects who take longer with three attributes than with two may show less of the effect. The next step is to try to manipulate the effect by instructing subjects to take longer (or forcing them to). Finally, to examine cognitive processes in another way, I shall time the attention to each attribute by requiring that the subject hover the pointer over the attribute level in order to see it, and record the time and sequence of viewing. (Note that these are spontaneous times, and subjects do not know they are being timed. In analyzing these data, it is necessary to omit outliers, e.g., by routinely deleting the 25% slowest times in each condition. Data analysis will be carried out primarily using a mixed model approach as described by Baayen, Davidson, & Bates, 2008, which is designed to deal well with missing data and unbalanced designs.)

I will also examine internal consistency as a function of number of attributes, as described above. I shall do this using two attributes at a time, and three attributes at a time. (Adding another attribute D does not affect the logic of the design for any triple of attributes.)

Most importantly, I shall look for ways to induce consistency, ways that are easier than the method used by Baron, Wu, Brennan, Weeks, & Ubel (2001), in which subjects were presented with their inconsistencies and asked to resolve them. One method is simply to force attention to both attributes by requiring that the subject hover the pointer over their values, and not letting the subject proceed until both (or all three) attributes have been examined. It would be encouraging if this experiment had the effect of eliminating the bias that it was designed to study. Forcing subjects to slow down might have a similar effect, e.g., by revealing attribute values sequentially. Most promising might be to require subjects to rate the effect of each attribute on the attractiveness of the option. Clearly several experiments will be needed here, varying both the method for improving consistency, the number of attributes, and the type of material.

* Assuming that one of these modifications improves the distribution of attention, I shall try go validate the modification against an external criterion, again using prediction rather than valuation (but also including, after the prediction task, a valuation task). In this case I can use consumer goods. I shall seek types of goods that subjects are knowledgeable about but with unknown prices. An example is houses, with the subject being potential or recent house buyers. Subjects could judge the asking and/or selling price of houses from descriptions in actual real-estate ads. The main hypothesis is that conditions that increase consistency of attention (as just described) will also increase accuracy of prediction.

2.4 Debiasing by removal of information

In CV studies of, for example, WTP to prevent oil spills, people often 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?

In Baron (draft), I argue that decision analysis and value elicitation should focus on consequences, putting aside the means of achieving them. (In essence, a consequence is a state of the world to which we assign value. It should not matter whether the state is the result of human behavior or nature, although we can also assign value to human behavior itself; see Baron, 1996.)

I also present an experiment, in which subjects were asked about several decisions in which responses are influenced by the means as well as the result. Then they were "de-biased" by presenting the identical case (on the same page) in terms of consequences only. Finally, they were asked if they regarded the consequence-only version as a fair summary. As an example, one version of the de-biasing condition for "omission bias" case was:

> 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.
>
> 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? ...
>
> The company chose treatment A.
>
> Would this choice make your more likely or less likely to choose this company as your insurer? ...
>
> 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 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).
>
> Is this a fair summary of what the decision is about?

For omission bias, 70% of the responses said that the summary was fair, even though the experiment biased against this response by putting the bias-inducing item first.

I propose to carry out additional tests of this kind of manipulation, using other examples of biases, such as the effect of cost, and even magnitude effects. (Magnitude effects can be studied using problem like the jacket/calculator problem of Tversky & Kahneman, 1981, as done by Baron, 2000.) I shall also counterbalance the order of the original and consequence-only versions.

I shall ask whether people have different attitudes depending on the type of bias. For range effects, most people might think that removal of information about ranges would be helpful, but for cases that involved deontolotical moral rules, e.g., rules that distinguish action and omission, more people might think that knowledge of the means by which a consequence is brought about is relevant to any judgment they make about it. At issue here is how seriously people take such rules in their role as providers of evaluations to potential providers of public goods. For example, do some people think it is morally wrong to withhold information about means? In the study I just described (Baron, draft), most subjects seemed to think that the information would not be required morally, but I did not ask this directly.

2.5 Inequality and direct interpersonal comparison

An issue in cost-benefit analysis is equity (in the sense of distributional fairness). When such analysis is based on monetary values, it risks under-weighing the interests of the poor (or, sometimes, over-weighing these interests). The utility of money, and many other goods, is greater for the poor. For example, a cost-benefit analysis of whether an international fund should provide mosquito nets

to prevent malaria might yield a low value. WTP would be relatively low because those most interested would be poor people, for whom even the retail cost would be a considerable sacrifice in utility terms. Because of this sort of problem, many authors have advocated an explicit correction for equity, even when welfare is measured in terms of utility rather than dollars (e.g., Adler, 2008; Adler & Sanchirico, 2006; but see Greene & Baron, 1991, for a critique of this position).

A great deal of literature shows that people prefer fair distributions. (See Baron, 1993, 1995, 2008, for summaries.) Some of these preferences seem to result from principles that are inconsistent not only with utilitarianism but also with any social welfare function based on individual utilities. For example, Baron (1995) found preferences for more equal distributions between groups even when less equal distributions clearly did more good (e.g., saving more lives) and when none of the recipients knew their group membership (and thus could not experience envy or anger at being denied some benefit on the basis of their group membership).

Of course, for money and many other goods, utility is marginally declining with the amount of the good, so, other things being equal (such as incentive effects), increases in equity can increase total utility. This principle alone accounts for many results in the literature (e.g., most of those of Frohlich & Oppenheimer, 1993). At issue is whether, in making judgments about social distributions, people prefer some distributions that are more equal over less equal distributions with a greater total utility.

In order to answer this question we must compare judgments of overall distributions to judgments of individual utilities. Dolan, Edlin, and Tsuchiya (2008) recently attempted such a comparison, and they found considerable departure from a model based on individual utilities. Their method relied on several assumptions, most importantly hometheticity (trade-offs unaffected by proportional changes in all quantities) and, in essence, additivity over time, as implied by their model of Quality Adjusted Life Years.

I propose to approach this question in other ways, extending the methods used by Greene and Baron (2001) and by Pinto-Prades & Abelllán-Perpiñán (2005). I shall present subjects with societal distributions of some quantifiable good, such as income or life expectancy. Each distribution will be described in terms of three equally-likely levels: high, middle, and low. To assess "social utility", I shall ask for evaluations of distributions from a social planner's perspective, e.g., "Rate the desirability of a society in which the top 1/3 had incomes of $2,000,000, the middle had incomes of $40,000 and the bottom had $20,000." (More detail will be given about the meaning of "income.") I shall also ask (with counterbalanced order) for ratings of *being* each person (high, medium, low) in each distribution. The question is how the "social utility" ratings depend on the individual ratings. One question is whether the social-utility judgments can be predicted from the individual judgments by a separable utility function. To assess this, I shall apply ACE (described above) to the data of each subject. This will find transformations of the response scales so as to yield the best linear fit. By looking at the residual errors, I shall test possible deviations from separability; if separability holds, then the errors should not be systematic (e.g., negative when predicted social-utility is high). Because this analysis is done for each subject, statistical tests across subjects are unaffected by the fact that optimization within each subject is a form of data snooping.

The next question is whether the implied social-welfare function is utilitarian. One issue is whether different levels in the distribution are equally weighted. Possibly those worse off could get more weight. This involves a simple test of the weights given by ACE.

Another issue is whether individual utilities require a concave transformation in order to account for social-welfare judgments, as proposed by Adler and Sanchirico (2006), Dolan et al. (2009), and many others. The idea is that utility itself is marginally declining when it comes to computing overall social welfare.[2] This issue cannot be addressed from the data described so far. We need a way of assessing individual utility functions on their own.

One way to do this (used by Pinto-Prades & Abelllán-Perpiñán, and by others) is a "veil of ignorance" gamble, in which a subject is asked for a self-interested decision about a gamble in which she could become any person in the social distribution with equal probability. A problem with this method is that it is very likely affected by risk aversion, which cannot be assumed to result only from declining marginal utility (as is required when using it to establish utility functions from gambles). Moreover, we need to assess utilities *within* the social distribution so that subjects can include whatever utilities result from comparing themselves to others, irrespective of the absolute level of their own outcome.

I thus propose to test a different method involving direct trade-offs across individuals, e.g., "Consider two people, person C in the lowest group, with an income of [whatever that group has], and person B in the middle group [with an income of ...]. Please provide an increase [decrease] for B and an increase [decrease] for C, so that the overall benefit [harm] is the same." Note that this uses the double-response approach described in section 2.2.2. I shall also use other methods, such as direct comparison of intervals — "Who benefits more?" — and matching — "How much would C have to increase [decrease] so that the benefit [harm] to C is the same as that from B's increase [reduction] for B?"

For the social-welfare condition, I shall use exactly the same method but apply it to the distribution, e.g., "Provide an increase [reduction] for group C and an increase [reduction] for group C, so that the overall benefit [harm] is the same." This project will be done in collaboration with Matthew Adler and Kevin Haninger.

Losses (decreases) might be easier to understand than gains, because it is easy to understand why a small loss to a poor person might be the equivalent of a large loss to a rich person. In the case of gains, subjects may want to "give" more to the poor person, which would lead to the opposite response from what would be implied by a concern with equity. These methodological issues require preliminary exploration. When possible, I shall use comparisons of differences of income levels (or levels of other goods) without mentioning which end of the interval is the status-quo or the default.

2.6 Willingness to pay tax

Conflicts between specific programs and overall spending: When people are asked directly about their willingness to pay an increased income tax in order to fund some public good, they often express surprising willingness, as if they neglected all the other possible public goods that might be equally valuable and funded through similar tax increases, which, together, could sum to more than 100% of income. On the other hand, when people are asked about the overall Federal

[2]From a normative perspective, this proposal has several apparent problems. It violates ex-ante Pareto optimality (Adler & Sanchirico, 2006). Also, unlike a utilitarian social welfare function, it requires a defined zero point on the utility scale, because the concavity applies to utility, not utility differences. From a descriptive perspective, Greene and Baron (1991) found that the intuition that "the utility of utility" is declining is not specific to social welfare: it applies to gambles too. Still, the point here is to investigate this idea as a descriptive account of judgments.

budget of the U.S., they generally favor lower taxes, lower spending, and a balanced budget (Baron & McCaffery, 2008). They tolerate deficits when they are asked about government spending on particular programs, but they still resist tax increases.

I propose to extend this research by using methods more like those used in CV. In CV, people are asked about a targeted tax, often a one-time tax, to pay for a particular public good. I shall thus contrast two conditions. In one, people would be ask their WTP for various tax increases to pay for various popular programs (many of which we can determine from previous research, although the most popular — universal health insurance — may soon be moot). In one condition, like CV, I would ask WTP program by program. In the other, I would ask WTP for an entire bundle of programs. The hypothesis is that people will get "sticker shock" for the bundle, even though they are willing to pay a fair amount for each of its components.

For a fair comparison, it is necessary to insure that, in evaluating each program, the subject understands that all the other programs previously evaluated would be put into effect for amounts correspond to his or her WTP, as done by Diamond, Hausman, Leonard, & Denning (1993) and others. In the experimental condition, each proposal is evaluated separately. In one control condition, all proposals together are evaluated as a bundle. In a more conservative control condition, the subject would evaluate each condition separately, as in the experimental condition, but the page would present a running total. These conditions all require the same order of presentation of the separate programs. Each subject will see several groups of proposals, and each group will be presented to different subjects in different random orders. In another method, I will present all the programs together on one page and allow the subject to adjust all the payments at once (as done by Baron & McCaffery, 2008). The control condition would be to display the total, in addition to all the individual WTP's. In another approach, I shall ask for *relative* WTP for the programs in the control condition, in terms of percent of the total increase in spending, then ask separately for the total increase, assuming that the relative increases would be constant. This method would assume that the allocation percentages would not depend on the total, but this assumption could be checked.

Citizen's duty vs. self-interest In recent studies (partly reported in Baron, 2009, and Baron, forthcoming), I have explored peoples concept of their duty as citizens, with particular attention to how this concept relates to their narrow self-interest, to what is good for groups to which they belong, such as nations, and to what is good for all, including outsiders (and potentially those not yet born). I have been particularly interested in "parochialism", defined as a preference for what is good for one's group, even when it causes greater harm to the self and to outsiders. I found that many citizens say that it is their duty to support parochial policies, those that benefit their group. And these citizens say they will vote according to their duty, even when they acknowledge that the policies they support would make things worse overall, including both their own group and outsiders.

In these studies, many subjects said that they had a duty to vote for what was in their self-interest. This was surprising, because it would seem that duty is a moral concept, arising out of obligations to others. Yet, the obligation to pursue self-interest could make sense. People may think of democracy as like a market, in which "consumers" vote for their favored product instead of buying it. As in a market, better products would get more votes (as better products presumably generate more demand). It may be that people are much better at judging their self-interest than at judging social good. But people who think they have a duty to support their self-interest may not

always think this far; they may have nothing in mind but a loose analogy with the market.

I propose to explore the nature of the concept of duty further, asking about the source of the duty. Do people think of it as a moral rule that applies to others? If so, do members of out-groups have the same moral duty to their group? It is contingent on what other people think, or would it be present regardless of what others think? Is it objective or subjective? (These questions are those often used in the study of moral judgment, e.g., Baron & Miller, 2000; Goodwin & Darley, 2008.)

And I shall explore beliefs about the role of self-interest and group-interest. Do people think that voting on the basis of self-interest leads to the best overall outcome? (Note that, although the literature suggests that self-interest plays a only small role in voting, my interest here is in those people who say that it defines their duty. Even if these people are in the minority, their beliefs may be held to some extent by others, and politicians may try to appeal to them.)

I shall explore these issues in the context of voting, of ordinary contingent valuation (CV), and an in-between case: CV may be presented in the form of a referendum in which subjects are told (at greater length than I state here) that they give a WTP for a public good, and, if and only if the median WTP of all the respondents is greater than the actual cost, the good will be provided at the actual cost. The implication is that provision or non-provision at the actual cost would win a majority vote. (Another in-between case of interest is an "assembly of citizens", which acts like a jury to resolve a dispute among governmental factions; see Elmendorf & Leib, 2009.)

Coercion vs. voluntary contributions Another way to explore the self-interest issue is to compare WTP in the following situations (for example):

- a filter to remove a certain type of bacteria from one's tap water (pure self interest, voluntary);
- a paternalistic law requiring everyone to install such filters (compulsory);
- a tax that everyone pays, which would install an equivalent filter in the town's water supply (compulsory public good, with the subject included);
- a contribution toward the same public good (voluntary, subject included);
- a contribution toward the public good when the subject already has a filter or well (voluntary, excluded);
- a tax for the public good when the subject already has a filter or well (compulsory, excluded).

In each of these cases, I shall ask about WTP, duty, and moral justifications.

I shall also examine different types of goods. (Some goods do not make sense in some of the conditions.) Environmental goods and general risk reductions benefit everyone and are true public goods. On the other side are mechanisms of redistribution, such as a tax to help the poor in some way. One possible result of interest is that some people are more willing to pay taxes — and find support for taxes more of a duty — for public goods that benefit everyone than for redistribution. They may regard compulsory taxation for redistribution as forced charity, reducing the freedom of those who give, for the benefit of others, without necessarily obtaining their consent — given that the tax will fall on those who vote against it. Of course, the same could be said for compulsory taxation to support public goods, because each person's tax payments go mostly to benefit others. People may not see it that way because of the morality-as-self-interest illusion (Baron, 1997b). I shall thus test the hypothesis that those who see only redistribution as forced charity are those who are more susceptible to the self-interest illusion, and I shall examine the effect of de-biasing that illusion (as done by Baron, 2001).

Intermediate cases are voluntary and compulsory insurance (including social insurance, which are also forms of redistribution but may not be seen that way, since the beneficiaries are not known in advance. People may be more willing to help the poor through insurance (e.g., unemployment and disability insurance) than through programs described as redistribution, even if the redistribution goes to the same people. The difference is whether the programs are described as "insurance in case someone loses a job" vs. "help for those who have lost jobs." This is ex-ante vs. ex-post, but the people are the same. I shall look for this sort of framing effect.

3 Broader effects and dissemination plan

I expect to publish the results of the proposed studies in a variety of journals, including law journals and those concerned with public policy. I also plan to present the work at meetings, particularly the Society for Judgment and Decision Making.

The proposed research will, if it is successful, make significant advances in our ability to estimate utilities of non-market goods. The focus of the intended applications is regulatory policy, particularly regulation of risk and environmental amenities. However, the same issues arise in health, where valuation of health states is part of cost-effectiveness analysis, now used around the world to increase the efficiency of health expenditures. The measurement problems are very serious. Yet, despite them, valuation is so needed that flawed methods are being used, on the assumption that, despite the flaws, they are better than nothing. Improvements are sorely needed.

The research could also contribute to our understanding of how people think about public policy, public goods, and their roles as citizens. The benefits of such understanding are less direct and immediate, but they may ultimately contribute to improved thinking about citizenship by teachers, politicians, and citizens themselves.

Some current students will also benefit. Although I do not list specific collaborations, I have a history of collaborating with younger researchers, including undergraduates, graduate students, and post-doctoral researchers, as well as other faculty here and elsewhere, and their students.

4 Human subjects

The issues addressed are well suited to the methods I have used since 1995 without any adverse events (so far as I know). Subjects complete questionnaires on the World Wide Web for pay.

Historically, my research has exposed subjects to a serious risk that was never a concern of the Institutional Review Board (IRB), the risk of identity theft because of the need to submit Social Security Numbers (SSNs) in order to be paid. In the last few years, the University has clarified its policy, and my subjects no longer need to provide their SSNs. (The payments I make to each individual are assumed to be below the threshold required.) Thus, I now feel that my research falls in the category of "minimal risk." The IRB has accepted this category even when I thought it was not strictly true.

The subjects are now from a panel of about 1,200 that has been recruited mostly as a result of their finding my site and asking to be included. (I used to encourage this.) They are diverse. From various studies over the years, I know that their median income is about that of the U.S. adult population, as is their median age. A problem has been that most are women. I am trying to correct

this problem by asking current members of the panel to recommend me to males they know. (But it is not getting enough, fast enough, so I plan to look for other ways to expand the panel in ways that are biased toward males.)

When a study is ready, I send email to (usually) 1/3 of the panel, thus saving the others for modifications of the same study. The email contains a link to the study, and a statement of the payment, usually $4 for a study that takes about 20 minutes (although times are highly variable). I keep records of what each person has earned, and, periodically, I send payments through PayPal (with a minimum of $7).

When I was trying to recruit subjects, my site was included in various Web pages that advertised ways to make money on the Internet. Comments were always favorable. I was considered honest and responsive in dealing with errors and questions.

I identify subjects by their email addresses, which is what I provide to PayPal. Before I analyze the data, I strip the email addresses from the rest of the data. Although the data are mostly just numbers, I story them securely on personal computers and now do not leave them on any public server.

References

Adler, M. D. (2008). Risk equity: A new proposal. *Harvard Environmental Law Review, 32*, 1–47.

Adler, M. D., & Sanchirico, C. W. (2006). Inequality and uncertainty: Theory and legal applications. *University of Pennsylvania Law Review, 155*, 279–377.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effect modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baron, J. (1993). Heuristics and biases in equity judgments: a utilitarian approach. In B. A. Mellers and J. Baron (Eds.), *Psychological perspectives on justice: Theory and applications*, pp. 109–137. New York: Cambridge University Press.

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). The illusion of morality as self-interest: a reason to cooperate in social dilemmas. *Psychological Science, 8*, 330–335.

Baron, J. (2000). Measuring value tradeoffs: Problems and some solutions. In E. U. Weber, J. Baron, & G. Loomes (Eds.) *Conflict and tradeoffs in decision making: Essays in honor of Jane Beattie.* New York: Cambridge University Press.

Baron, J. (2001). Confusion of group-interest and self-interest in parochial cooperation on behalf of a group. *Journal of Conflict Resolution, 45*, 283–296.

Baron, J. (2009). Cognitive biases in moral judgments that affect political behavior. (Special issue of *Synthese* on the foundations of the decision sciences, edited by H. Arló-Costa & J. Helzner).

Baron, J. (in press). Looking at individual subjects in research on judgment and decision making (or anything). *Acta Psychologica Sinica*. (Special issue on "Methodological concerns of the experimental behavioral researcher: Questions and answers," edited by S. Li).

Baron, J. (forthcoming). Parochialism as a result of cognitive biases. To appear in A. K. Woods, R. Goodman, & D. Jinks (Eds.), *Understanding social action, promoting human rights*. Oxford: Oxford University Press.

Baron, J. (draft). Prospects for utilitarian decision analysis. Presentation at "New ideas for risk regulation," Resources for the Future, Washington, June 22–23, 2009. To be submitted for a special issue of *Risk Analysis*.

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. & Jurney, J. (1993). Norms against voting for coerced reform. *Journal of Personality and Social Psychology, 64*, 347–355.

Baron, J., & McCaffery, E. J. (2006). Unmasking redistribution (and its absence). In E. J. McCaffery & J. Slemrod (Eds.), *Behavioral Public Finance*, pp. 85–112. New York: Russell Sage Foundation.

Baron, J., & McCaffery, E. J. (2008). Starving the beast: The political psychology of budget deficits.

In E. Garrett, E. A. Graddy, & H. E. Jackson (Eds.), *Fiscal challenges: An inter-disciplinary approach to budget policy*, pp. 221–239. New York: Cambridge University Press.

Baron, J. & Miller, J. G. (2000). Limiting the scope of moral obligation to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology, 31*, 705–727.

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.

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.

Cameron, T. A., & DeShazo, J. R. (2008). A generalized empirical model of demand for health risk reductions. Manuscript, Department of Economics, University of Oregon.

Caplan, B. (2007). *The myth of the rational voter: Why democracies choose bad policies.* Princeton, NJ: Princeton University Press.

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.

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., Edlin, R., & Tsuchiya, A. (2008). The relative society value of health gains to different beneficiaries. Final report. Health Economics and Decision Sciences Discussion Paper, SCHARR: Sheffield.

Elmendorf, C., & Leib, E. J. (2009). Op-Ed Contributors: Budgets by the People, for the People. *New York Times, July 28*. http://www.nytimes.com/2009/07/28/opinion/28leib.html?emc=eta1

Frohlich, M., & Oppenheimer, J. A. (1993). *Choosing justice: An experimental approach to ethical theory.* Berkeley: University of California Press.

Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106*, 1339–1366.

Greene, J., & Baron, J. (2001). Intuitions about declining marginal utility. *Journal of Behavioral Decision Making, 14*, 243–255.

Hirschleifer, D. (2008). Psychological bias as a driver of financial regulation. *European Financial Management, 14*, 856–874.

Kemp, S. (2002). *Public goods and private wants: A psychological approach to Government spending*. Cheltenham, U.K.: Edward Elgar.

Kemp, S. (2003). The effect of providing misleading cost information on the perceived value of government services. *Journal of Economic Psychology, 24*, 117–128.

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

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

McCaffery, E. J., & Baron, J. (2003). The Humpty-Dumpty blues: Disaggregation bias in the evaluation of tax systems. *Organizational Behavior and Human Decision Processes, 91*, 230–242.

McCaffery, E. J., & Baron, J. (2004). Framing and taxation: evaluation of tax policies involving household composition. *Journal of Economic Psychology, 25*, 679–705.*

McCaffery, E. J., & Baron J. (2005). The political psychology of redistribution. *UCLA Law Review, 52*, 1745–1792.

McCaffery, E. J., & Baron J. (2006a). Isolation effects and the neglect of indirect effects of fiscal policies. *Journal of Behavioral Decision Making, 19*, 1–14.*

McCaffery, E. J., & Baron J. (2006b). Thinking about tax. *Psychology, Public Policy, and Law, 12*, 106–135.

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.

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

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

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.