Appendix B: Stated Preference Quality Assurance and Quality Control Checks

Table of Contents
1  Introduction ............................................................................................................................. 2
2  Risk comprehension verification............................................................................................. 2
3  Minimization of biases associated with omitted substitutes.................................................. 2
4  Minimization of hypothetical bias.......................................................................................... 3
5  Minimization of distortions from provision rules and order effects...................................... 3
6  Tests for the effects of program scope.................................................................................. 3
7  Minimization of yea-saying.................................................................................................... 3
8  Basic tests for theoretical validity ......................................................................................... 4
9  Respondent learning and fatigue ........................................................................................... 4
10 Heuristics and metric recoding ............................................................................................... 5
11 Concerns about choice inconsistency .................................................................................... 6
11.1 Reed Johnson’s VALIDTST program .............................................................................. 6
12 Unobserved recoding and scope insensitivity ...................................................................... 8
13 References ............................................................................................................................ 9
1 Introduction
We undertook numerous ex ante measures to minimize biases through careful survey design and also seek to evaluate our data ex post for the presence of remaining biases. Our survey includes a verification of respondents’ risk comprehension well as features to limit the extent of biases associated with the hypothetical nature of the choice questions, distortions due to the omission of relevant substitutes, order effects across the choice questions for any one individual, and yea-saying tendencies. Our choice set design is structured to provide ample opportunity for external “scope” testing, as well as for general evaluation of the theoretical validity of results in relation to economic theory. These efforts are described in this Appendix with additional detail for some design issues and tests described in Appendices A, C, and D.

We also include in this appendix some information in response to the concerns of reviewers of previous versions of the main paper, Cameron and DeShazo (2009). In particular, we discuss respondents’ potential use of choice heuristics and their potential recoding of attributes, and whether respondents to our survey can be assessed with respect to the consistency of their choices.

2 Risk comprehension verification
After we administer an extensive risk tutorial and present the risk changes in three forms (textually, graphically and mathematically), we test the individual's risk comprehension. This comprehension test requires individuals to rank the sizes of the risk reductions associated with two risk mitigation programs. Approximately eighty percent of the individuals demonstrated adequate comprehension of the relative risk size reductions of the programs, which is a rate consistent with risk comprehension levels documented in other surveys (Alberini, et al., 2004 and Krupnick et al., 2002).1

3 Minimization of biases associated with omitted substitutes
In contrast with many valuation studies that focus on just one or two risks and their associated risk-reduction programs, we endeavor to reduce biases associated with so-called bracketing (Read, et al, 1999) via inclusion of nearly all major competing health risks (and specific programs to reduce them) in at least one of each individuals' choice sets.2

Presentation of a broad spectrum of major health threats and mortality risks increases the generality of our estimates. Of course, a potential disadvantage of this approach is the cognitive complexity associated with the choice task, which we seek to minimize through careful survey design, and which we evaluate ex post.3

---

1 As Harrison and Rutstrom (2006) argue, reliable estimates of the monetary value of risk reductions hinge on respondents' comprehension of mortality risks. Their research suggests that it is indeed possible to elicit subjective beliefs about mortality risks from individuals. We discuss the effects on the estimated parameters of including and excluding individuals from the sample based on their risk comprehension in an appendix.

2 Ashenfelter and Greenstone (2004) also address the problem of omitted variables and other biases in measuring the value of a statistical life. Competing risks are addressed in Dow et al. (1999).

3 We assess this concern directly in the survey. After each choice set we ask individuals how difficult each choice was. On a scale of 1 to 5 (very easy to very difficult), the average response for the first choice set was 3.2. This rating fell with each subsequent choice set, suggesting that the choice task became easier with increasing familiarity.
4 Minimization of hypothetical bias
At the beginning of the valuation module, we include a "cheap talk" reminder—to ensure that respondents carefully consider their budget constraints and to discourage them from overstating their willingness to pay (Cummings and Taylor, 1999; List, 2001). Individuals are instructed, "In surveys like this one, people sometimes do not fully consider their future expenses. Please think about what you would have to give up, to purchase one of these programs. If you choose a program with too high a price, you may not be able to afford the program when it is offered."

The second strategy comes from the mechanism design literature which involves convincing respondents that their answers may actually effect the provision or pricing of the good under study (Carson, Machina, and Groves, 2004).

5 Minimization of distortions from provision rules and order effects
To clarify provision rules for each choice set (Taylor, et al, 2005) and to avoid potential choice set order effects (Ubel et al., 2002; de Bruin and Keren, 2003), we instructed individuals to assume that every choice is binding and to evaluate each choice set independently of the other choice sets. Our empirical analyses show that the first four choice sets appeared largely free of choice task order effects. Individuals did exhibit a slightly higher propensity to select a program from the last choice set, an effect that has also been demonstrated in other similar settings (Bateman, et al, 2004).

6 Tests for the effects of program scope
We explore whether individual choices are sensitive to the scope of the illness profile and the scope of the risk mitigating program (Hammit and Graham, 1999; Yeung et al., 2003). We show, even in the simplest possible choice models, that individuals readily pass the "scope test." Our subjects are highly sensitive to differences in the scope of our key choice-scenario attributes across the 7520 different choice scenarios considered by our 1801 individuals. In Table 2, Model 1 demonstrates that even a minimal conditional logit choice model, specified in terms of the minimal number of raw program attributes, produce intuitively plausible and strongly significant coefficients on the two most crucial aspects of each program: i.e. a lower cost and a greater risk reduction make a program more attractive. Model 2 shows that the other two most important dimensions of the illness profiles, the number of sick-years and the number of lost life-years for which the risk will be reduced, are also strongly significant determinants of respondents' choices among programs. Respondents are systematically more likely to choose programs which address more serious health threats.

7 Minimization of yea-saying
Another concern, if there are no actual costs to respondents at the time they agree to purchase a hypothetical good, is that they will “yea-say,” that is, agree to purchase the offer good in a effort to be agreeable. We employed a strategy that involves reminding respondents that there exist

---

4 For a complete description, see the annotated survey instrument available from the authors. We note that Hakes and Viscusi (2007) have demonstrated that the value of a statistical life implied by stated preference survey estimates is not statistically significantly different from estimates of the same quantity derived from seatbelt usage.
several “good” and acceptable reasons to reject the offer of the goods under study. This approach is intended to mitigate yea-saying which may lead to the respondents’ inadvertent overstatement of their WTP. Discussing many of the legitimate reasons for opting out of the choice occasion also reinforces the role that economic reasoning should play in their decision making. It also reminds respondents of the importance of substitute goods and binding budget constraints while compelling them to consider more carefully the relative expected value of the goods being offered.

8 Basic tests for theoretical validity
An important test of the validity of individuals’ stated choices is whether their WTP varies as theory would predict it should with specific variables. In statistical analyses that followed the survey (Cameron and DeShazo, 2009; DeShazo and Cameron, 2005b and 2005c), we show the respondents’ stated WTP does vary systematically with their income, age, discount rate, and health status. It also varies (as theory would predict it should) with the latency, duration and severity of the illness profiles as well as the cost and the effectiveness of the program as measured by the size of the risk reduction.

9 Respondent learning and fatigue
In response to the complexity and number of choice tasks, respondents may both learn and become fatigued. Learning about both their own preferences and how to more efficiently choose formats might reduce the amount of time respondents spent on each choice task. Increasing fatigue, in contrast, may increase their time-on-task. These processes are important for us because learning might reasonably be expected to increase the quality of preference information we can recover from their stated choices, while fatigue might reasonably reduce it.

We evaluate these effects in three ways. First, after each choice set we ask individuals about the subjective difficulty of that choice, using a rating scale for difficulty. (See the Survey forms in Appendix A.) On a scale of 1 to 7 (from “easy” to “very difficult”), the average response for the first choice set was 3.2. (See Forms 26, 30, 34, 38, and 42.) We asked respondents to continue to rate the difficulty of each of their choice tasks. The first such subjective rating can be expected to be fairly arbitrary, since the respondent must decide for themselves “relative to what?” However, these difficulty ratings, on average, tend to fall with each subsequent choice set, suggesting that respondents perceived that the choice task became easier with increasing familiarity or learning.

Second, we examined trends in the amount of time respondents spent on selecting an alternative from each choice set. On average, respondents’ time spent on each choice task fell consistently from their first, second and third choice tasks, thereafter remaining relatively constant across their fourth and fifth choice tasks. This result, which is consistent with

---

5 Just prior to the introduction of the first choice set, the survey said “People might choose neither program because they:
- could not afford either program,
- did not believe they face these illnesses or injuries,
- would rather spend the money on other things, or
- believe they will be affected by another illness or injury first.”
respondents’ self-reported difficulty ratings, suggests that fatigue at least did not slow them down.

Third, we explored whether there was systematic variability in individuals’ WTP across the five choice tasks. One might expect that if individuals were becoming fatigued, their answers would become increasingly random. Increasing randomness might also occur if individuals were “rushing” through the choice tasks. (This concern might be heightened in light of the above evidence of progressively falling time spent on each subsequent choice task.) Our test for trends in implied WTP as function of the order of the choice tasks show no discernable trend up or down. The only pattern that clearly emerged was a slight increase in WTP for the very last (i.e., fifth) choice task. Respondents were informed between the fourth and fifth choice tasks that the next choice would be the last program choice they would be asked to consider, so this could be a “home stretch” phenomenon.

10 Heuristics and metric recoding
Respondents’ use of heuristics in decision making is indeed a very important consideration and one to which we devoted a great deal of care to minimize and evaluate through the many iterations of trial versions of the survey format with numerous survey test subjects over several rounds at the Knowledge Networks facility in Menlo Park, CA. But it is important begin by asking what is a fair standard for the “eligibility” of preference data and whether expectations for stated-preference (SP) data, as opposed to revealed-preference (RP) data, may represent a double standard. Would RP data be disqualified from use if it were found to be affected by heuristics? The entire field of behavioral finance and a growing number of influential field experiments suggest that answer is clearly no.

Given this, the next important question is whether SP data is more likely to be effected by heuristics that would comparable RP data. First, our respondent probably see more information, more comparably presented, than they would be shown in any real choice situation with respect to opportunities to reduce risks to their lives and health. Moreover, we probably spent more time, and provided more learning strategies (with risk measures and graphics) to prepare them for their decision making than would many medical office visits where patients at their annual check-ups must consider their doctor’s recommendations to elect (and subsequently pay for) a variety of diagnostic tests.

The next concern is whether respondents selectively discard or recode the information they are given and report their preferred alternative from each set in a way that renders these choice data unusable for the purposes of recovering an informative estimate of their WTP for health risk reductions. As experimentalists, we must first to acknowledge that it is not possible to observe directly any individual’s mental decision process, so we likewise cannot observe the presence or absence of heuristic processes. What we can do is to look for evidence that such heuristics, to the extent that they exist, have damaging consequences for our data and the inferences we draw from them.

Do we see any blatantly obvious evidence of the use of damaging heuristics? No. First, the attributes of the illnesses, and the characteristics of the respondents, are strongly statistically significant. Second, the nature of the systematic variability that we identify is consistent with what intuition and general economic theory would predict. Third, the fitted WTP amounts based on our final SP estimates of consumer preferences are generally consistent with the available benchmarks for RP data that exist within the literature.
11 Concerns about choice inconsistency

One concern is that due to the complex nature of the choices individuals may not correctly and consistently evaluated the risk-tradeoff questions. If this is true, we would observe respondent failing internal consistency checks such as those for transitivity of preferences. If the complexity overwhelmed respondents such that their choices did not preserve the properties of transitivity then a degree of randomness would characterize the choice data. In the extreme, if this happened, the observed choices would appear predominantly random. Our results would not be statistically significant. We conclude that choice inconsistency does not appear to be happening in the extreme and certainly not so much as to prevent us from getting fairly precise measurements of the central tendency of preferences.

11.1 Reed Johnson's VALIDTST program

In some simpler choice contexts, where there are no more than ten different levels for each attribute, and where utility can be assumed to be either strictly increasing or strictly decreasing in each distinct attribute, it is possible to assess choice consistency systematically, using a Gauss program called VALIDTST.PRG, prepared by Reed Johnson. This program, written in 2004 (subsequent to the fielding of our survey), is designed to take conjoint choice data and test for “stability, monotonicity, transitivity, and dominance relations in SP designs.”

The VALIDTST program allows the researcher to specify the number of attributes and the number of alternatives as well as the number of choice task repetitions. The researcher must specify whether each attribute has levels with are decreasing (-1), increasing (+1) or unordered (0). Identifiers must be provided for each respondent and for each choice. Unfortunately, the program appears to allows for no more than ten possible levels for each attribute, since each attribute level must be represented by a single digit, to be concatenated into a string. Many of the attributes in our choice sets have far more than ten possible levels.

The different available tests in the VALIDTST program are as follows (with discussion concerning the appropriateness to our study appended in each case):

1. Look for stability relations in repetitions of the same choice: A~B~C, A~B~C: "If A1=A2, B1=B2, and C1=C2 then Choice1 must equal Choice2"
   - Our survey design involved random draws of attribute levels for each alternative in each choice set (as described in Appendix C). Given the number of attributes and the number of possible levels for each attribute, it is highly unlikely that there is ANY pair of identical choices anywhere in our 7520 choice sets, let alone among the five choice sets posed to any one individual.

2. Look for within-pair monotonicity: A~B/A~C, where all elements of B <= C: “If A1<=B1, then Choice1 should not be A”
   - Our survey design rejected any pair of program alternatives for which one program both cost less and produced a greater risk reduction. Since we are confident that preferences are probably monotonic in cost and in the sizes of the risk reductions, there are no cases in our design where we could test for within-pair monotonicity.

3. Look for across-set monotonicity: A~B1~C1, A~B2~C2: "If B1<=B2, then Choice1 = B1 does not allow Choice2 = A"
• Each choice set presented to an individual involves a different disease label. For each gender, there were eleven possible disease labels, and we used ten labels randomly selected from this set, sorted randomly into pairs, for each of the five choice sets. In no case would the same individual see the same alternative paired with two different alternatives in two choice sets.

4. Look for consistency relations: A~B~C, A~B~D: "Choice sets A B1 C1 and A B2 D2. If B1=B2, then Choice1 = B1 does not allow Choice2 = A, and Choice1 = A does not allow Choice2 = B2"
   • Again, there is no instance in our survey design where the identical alternative A is paired with two different alternatives B1 and B2 for the same individual. Across all of the potentially 15040 unique hypothetical programs used in the 7520 different choice sets faced by our respondents, each alternative is a random combination of ten or eleven different attributes, most with a very large number of possible levels. It is likely that there is no pair of identical alternatives A proposed to any pair of individuals, either.

5. Look for transitivity relations: X~Y, Y~Z, X~Z: "For unique profiles X, Y, and Z, if Choice1=Y, Choice2=Z, then Choice3 can't equal X"
   • In our survey, the same person never sees the same alternative X in more than one choice set. All ten of the alternatives seen by each respondent are different. In particular, there are different disease labels associated with each of the ten non-status-quo programs described in the five choice sets presented to each individual. Thus it seems we have no real opportunity to test for transitivity for any given individual. Heterogeneity in subjective risks of different illnesses or injuries will mean that we cannot expect identical preferences across individuals, so even if we could find some set of identical alternatives posed to two individuals, there can be no expectations that they would make choices that imply transitivity. We estimate average preferences in this context.

6. Look for dominance (no-tradeoff situations): A~B/A~C, A~B~C, where preferred attribute is z: "Choice sets A1 B1 C1 and A2 B2 C2. If A2(x,y) > A1(x,y), B2(x,y) > B1(x,y), and C2(z) < C1(z), then Y = 1."
   • As mentioned above, we rejected all choice sets where the two substantive alternatives exhibited strict dominance in terms of program cost and the size of the risk reduction. Thus, by design, no substantive pair of alternatives involves strict dominance and both of the substantive pairs involve a greater (e.g. non-zero) risk reduction but a higher cost (lower net income) than the status quo. Thus there seems to be no opportunity to find any such case in our data.

Thus, unfortunately, we seem to have too many attribute levels to make use of this helpful utility, and likely no more than a very tiny fraction of cases, if any, where it would even be possible to assess these problematic relationships. Also, we know that preferences with respect to lost life-years, for example, depend upon the number of prior sick-years, so that lost life-years, if an illness is bad enough and long enough, could be viewed as a good thing (from the perspective of the present). Thus it is not even possible to assert, unambiguously, that increasing
levels of some attributes imply increasing utility. Preferences are not strictly monotonic in every attribute (and in any case, utility, empirically, is not linear and additively separable in each attribute).

Perhaps we should have seeded our choice sets with cases that would have permitted tests of choice consistency. However, ex ante, we were concerned that we should have enough independent variation in attribute levels across choice sets and across individuals so that we could identify robustly statistically significant marginal effects of illness profile attributes on indirect utility levels. Ex post, it appears that we would have had ample “room” to build in some redundancies among the choice sets, to permit such testing, but it is obviously too late for that now.

12 Unobserved recoding and scope insensitivity
A second concern is that even if choices are not random, respondents may recode the risk information from the metric in which it is given to them into one that is simpler and unknown to the researcher. It was this type of recoding and the resulting insensitivity to the scope of good that led to the recommendation for mandatory “external scope tests” in early CV formats. The proper way to conduct this external scope test is to “split the sample” and to administer one amount of the good some respondents and a different amount to others. The objective is to demonstrate that different people are willing to pay different amounts for different quantities of a non-market good. If respondents are prompted, in the context of the same questionnaire, to evaluate both a larger and a smaller amount, the test is merely an “internal scope test.”

We undertook a variant of the required scope test that is analogous to splitting the sample in the way we designed our choice sets. Importantly, the choice scenarios vary across respondents, as well as across choice sets. For each of the 1801 individuals whose choices were used to produce our estimates—indeed, for all 7520 choices analyzed in our data, the illness profiles were essentially unique.

Had all respondents seen the same five choice sets, we would not have met the external scope test requirement. We would only have been able to show that WTP was larger when benefits were larger “between scenarios.” This would have been a weaker internal scope test. However, instead of splitting our sample of 1801 individuals into two groups, and showing everyone within the same group the same set of choice scenarios, we effectively had 1801 different groups. We contend that this strategy actually vastly outdoes the usual “external scope test” because every respondent considered different illness profiles and different risk-reduction program costs. We did not simply split the sample into two groups, each of which saw a different level of benefits and different costs.
13 References


DeShazo, J.R. and Trudy Ann Cameron (2004a) “Mortality and Morbidity Risk Reduction: An Empirical Life-cycle Model of Demand with Two Types of Age Effects,” University of Oregon, manuscript


