Environ Resource Econ (2009) 43:473–502
DOI 10.1007/s10640-008-9242-8

Eliciting Willingness to Pay without Bias using Follow-up Certainty Statements: Comparisons between Probably/Definitely and a 10-point Certainty Scale

Glenn C. Blomquist · Karen Blumenschein · Magnus Johannesson

Received: 10 March 2008 / Accepted: 3 October 2008 / Published online: 26 October 2008

Abstract Correction for hypothetical bias using follow up certainty questions often takes one of two forms: (1) two options, "definitely sure" and "probably sure", or (2) a 10-point scale with 10 very certain. While both have been successful in eliminating hypothetical bias from estimates of WTP by calibrating based on the certainty of yes responses, little is known about the relationship between the two. The purpose of this paper is to compare the two using data from three field experiments in a private good, dichotomous choice format. We compare four types of yes responses that differ in the criterion used to determine if there is sufficient certainty for a hypothetical yes response to be considered a true yes response. We make several comparisons, but focus on determining which values on the 10-point scale give the same estimates of WTP as "definitely sure" hypothetical yeses and real yeses (actual purchases). Values that produce equivalence are near 10 on the certainty scale.

Prepared for the 6th iHEA World Congress held July 8–11, 2007 in Copenhagen, Denmark and the Conference "Validating Values of Safety and Travel Time" held August 16–17, 2007 at Örebro University in Örebro, Sweden. For comments we thank conference participants, participants in a seminar at the Baltic International Centre for Policy Studies in Riga, Latvia, as well as John Horowitz, John Whitehead, and two anonymous referees. Work on this paper was done while Glenn Blomquist was a Visiting Fulbright Scholar at the Stockholm School of Economics in Riga. He gratefully acknowledges this support. Brandon Koford provided exemplary research assistance.

G. C. Blomquist (✉)
Department of Economics and Martin School of Public Policy and Administration, University of Kentucky,
Gatton Business and Economics Building 335,
Lexington, KY 40506-0034, USA
e-mail: gcblom@uky.edu

K. Blumenschein
College of Pharmacy and Martin School of Public Policy and Administration, University of Kentucky,
Lexington, KY 40536-0082, USA
e-mail: kbluml@uky.edu

M. Johannesson
Department of Economics, Stockholm School of Economics, Box 6501, SE-113 83, Stockholm, Sweden
e-mail: magnus.johannesson@hhs.se

Keywords Certainty statements · Contingent valuation · Field experiments · Hypothetical bias · Willingness to pay

JEL Classification C93 · D61 · I10

1 Introduction

Valuation of non-market goods is essential to making efficient individual and public decisions. Undervaluation leads to missed opportunities for worthwhile investment or consumption. Overvaluation leads to investment or consumption which costs too much in terms of other valuable options. Behavior in implicit markets for non-market goods can yield useful information such as compensating wage or housing price differences. Stated preferences in constructed markets can yield useful information also in the form of contingent values. Contingent valuation is useful for health, safety, and environmental goods for which markets, implicit or explicit, do not exist. While a great deal of progress has been made with various valuation approaches and collectively they offer much to decision makers, each valuation approach has limitations. One limitation of contingent valuation, perhaps the most important, is that hypothetical responses tend to overestimate real responses. Meta-analyses by List and Gallet (2001), Little and Berrens (2004) and reviews by Harrison (2006) and Harrison and Rutström (2008) suggest that contingent valuation tends to produce hypothetical bias in the form of overestimation of actual (real) value.

Eliciting willingness to pay (WTP) values with confidence and without bias requires mitigation of potential hypothetical bias.[1] Mitigation efforts include ex ante countermeasures taken before elicitation, ex post countermeasures taken after elicitation, and in medias res countermeasures taken during elicitation. Ex ante mitigation includes state of the art survey design that incorporates reminders of closely-related goods, especially substitutes, and reminders of the individual or household budget constraint, see Loomis et al. (1996) and Whitehead and Cherry (2007). Ex ante cheap talk explicitly informs respondents that in similar hypothetical situations people tend to say yes more than they would in real situations and exhorts respondents to state what they would actually do. Cummings and Taylor (1999) find that cheap talk works well for their four environmental goods, but Little and Berrens (2004) meta analysis showed mixed success for cheap talk.

Ex post mitigation primarily has taken the form of determining how certain respondents are that they would actually do what they say they would do. All respondents are asked a contingent valuation question. After they state what they would do, they are asked a follow up question to determine how certain they are. Only respondents who are sufficiently certain that yes they would actually pay are counted as giving a yes response. This calibration tends to remove any hypothetical bias in the initial elicitation. One way to determine if individuals are sufficiently sure is to follow up and ask if they are "probably sure" or "definitely sure." Based on comparisons between hypothetical and real purchases decisions Blumenschein et al. (1998, 2001, 2008) find that willingness to pay can be elicited without bias if only yes responses by individuals who are "definitely sure" are considered true yes responses. Another way to determine if individuals are sufficiently sure is to follow up and ask respondents to

[1] Evidence that hypothetical bias is a problem in contingent valuation does not imply that all estimates of WTP are biased upwards. Farmer and Lipscomb (2008), for example, illustrate how estimates can be biased downwards by conservative responses. Their analysis depends on identifying different types of respondents who have different incentives. To get unbiased estimates using follow up certainty statements what matters is identifying respondents who will actually pay out of the group that says that they will pay.

indicate how certain they are using a 10-point scale where ten is very certain. Champp et al. (1997) and Champ and Bishop (2001) find that average WTP can be estimated without bias if only yes responses with a certainty value greater than a critical value are considered true yes responses.

Calibration using either type of follow up certainty questions is based on the idea that the individual has a value for the good and compares the value to the price. For prices below the value, the lower is the price the more certain the individual is about paying. For prices above the value, the higher is the price the more certain the individual is about not paying. For prices close to the value, the individual is less certain the closer is the price to the value. This idea can be expanded to allow for the individual to have a range or distribution of values for the good, see for example, Ready et al. (1995). When the price falls below or at the lower end of the distribution of values, the individual is likely to be definitely sure or very certain about paying. When the price falls above or at the upper end of the distribution, the individual is likely to be definitely sure or very certain about not paying. When the price falls in between, the respondent is less certain about the decision with certainty of payment varying inversely with price.[2]

In medias res mitigation incorporates the degree of certainty into elicitation. The "don't know" option recommended by the NOAA panel (Arrow et al. 1993) can be interpreted as incorporating certainty into elicitation. Wang (1997) uses don't know responses to estimate a value distribution function. Johannesson et al. (1993) and Ready et al. (1995) incorporate several levels of certainty with polychotomous choice contingent valuation. Multiple-bounded discrete choice (MBDC) incorporates the uncertainty extensively by integrating certainty into a payment card. For example, Vossler et al. (2003) present to the individual a two-dimensional matrix with several dollar amounts (prices or bids) and several levels of certainty at each price. The five levels of certainty are definitely no, probably no, not sure, probably yes, and definitely yes. It is polychotomous choice at various prices in the format of a payment card. They use multiple bounded logit to estimate separate willingness to pay (WTP) distributions for each level of certainty. WTP for the definitely yes is lower than WTP which includes the probably yes which, in turn, is lower than the WTP which includes the not sure. They find the best match between hypothetical and real behavior when counting probably yes as the threshold level of certainty.[3]

Evans et al. (2003) draw on cognitive research that indicates that verbal probability statements convey subjective probabilities. They use a payment card that has dollar amounts and the five qualitative categories and assign probabilities of yes to each category. Estimates from their base case (definitely no 0, 0.15, 0.50, 0.75, and 1 definitely yes) are compared to estimates from a Definitely Yes Model (0, 0, 0, 0, 1) and estimates from models based on other assignments of probabilities. Berrens et al. (2002) use a 0–10 scale to elicit how likely the individual would be to pay the stated amount and treat the response as a subjective probability; the scale value divided by 10 is the probability. They compare estimates based on

[2] The idea that individuals who are uncertain will be conservative has been applied to explaining part of the difference between willingness to pay and willingness to accept regarding risks to health and safety. See Dubourg et al. (1994).

[3] Vossler and McKee (2006) compare elicitation formats including dichotomous choice with follow up certainty statements and MBDC payment cards for hypothetical bias. Their study is related but different from the current study. One difference is that they allow "probably" yes responses to count as true yes responses sometimes instead of only "definitely." A more fundamental difference is that in the induced-value experiments, values are assigned to subjects as part of the experimental design. The current study allows for preference uncertainty and estimates individuals' values for the goods.

incorporating the 0–10 into the value elicitation question directly and using it as a follow-up question to dichotomous choice.[4]

Regardless of whether the elicited certainty is interpreted as a subjective probability or not, individuals who are more certain of their stated responses have a better match between stated intentions in contingent valuation and real behavior. In addition, and perhaps not surprisingly, individuals who are more certain of their stated responses give more internally valid responses. For example, Blumenschein et al. (2008) compare logit regressions of hypothetical purchase decisions for a diabetes management program for two subsamples: (1) when the definitely sure yes responses are excluded and only the probably sure yes responses and no responses are included and (2) when the probably sure yes responses are excluded and only the definitely sure and no responses are included. The second subsample with the definitely sure yes responses is explained better in that the Chi-squared value, percentage of correct predictions, and McFadden's R-squared are all much higher than for the subsample with the probably sure yes responses included. In addition, the coefficient on price is negative and highly significant in the subsample with the definitely sure responses while the coefficient on price is not statistically significant at conventional levels in the subsample with the probably sure yes responses.

Another example of more internally valid responses is the Watson and Ryan (2007) study of willingness to pay for air ambulance services in which values are elicited using double-bounded dichotomous choice. Respondent certainty is determined by a follow up certainty scale.[5] Their analysis focuses on anomalies in terms of internal validity. A key finding is that individuals who are very certain exhibit few anomalies such as starting point bias compared to individuals who are less certain.

Ex post mitigation of hypothetical bias using follow up certainty questions has produced promising results in which stated hypothetical intentions match real behavior. Two oft-used ways in which follow up certainty questions have been asked are: (1) using two options, definitely sure and probably sure, and (2) using a 10-point scale with 10 very certain. While both have been successful in eliminating hypothetical bias from estimates of WTP, little is known about the relationship between the two.[6] The purpose of this paper is to compare these two ways of asking follow-up certainty questions. The data are from three field experiments, the first offered a diabetes management program, the second offered an asthma management program, and the third offered a lipid management program. In each experiment the good was offered hypothetically in contingent valuation and for real, i.e., actual purchases. By (split sample) design in the diabetes and asthma experiments, individuals were offered the good either hypothetically or for real purchase. In the lipid experiment, approximately half of the individuals were offered the good hypothetically in contingent valuation and then for real purchase and the other half was only offered the good for real purchase. We make several comparisons, but the focus is on determining which values on the 10-point scale give the same estimates of WTP as definitely sure. We

[4] Li and Mattsson (1995) ask a follow up question using a 0–100% certain scale and interpret it as a probability.

[5] Watson and Ryan (2007) ask a follow-up certainty question using a 1–5 scale where 5 is very certain. As mentioned above, the 0–100 scale has also been used to elicit the probability of paying.

[6] Svensson (2000) notes that rating scales have been used for many years in many contexts. Her results from nonparametric tests for consistency for a verbal descriptor scale, a graphic rating scale, and a visual analog scale can be interpreted as relevant to our comparison. She finds that the verbal descriptor scale similar to probably sure/definitely sure is best and slightly better than the graphic rating scale similar to the 10-point certainty scale. Both are superior to the visual analog scale. However, the matches between her scales and the two certainty scales compared in this paper are loose enough and the subject matter is different enough to make closer comparisons worthwhile.

also determine which values on the 10-point scale give the same estimates of WTP as the real purchases. We find that the values that produce equivalence are always near 10 on the certainty scale.

2 Types of Hypothetical Yes Responses in Contingent Valuation—What is True?

We make comparisons for four types of yes responses in contingent valuation. We choose four types that have generated interest in eliciting WTP without bias. They differ in how they determine the subset of hypothetical yes responses that are presumably true yes responses.

2.1 Definitely Sure

The first type is based on the follow up certainty question that offers two options: "definitely sure" and "probably sure." In contingent valuation if the respondent answers yes and is definitely sure, then the response is considered a true yes. If the respondent answers yes and probably sure or answers no, then the response is considered a true no. It is possible that certainty matters for no responses, but in two previous experiments and the lipid experiment reported in this paper a respondent who says no and then actually makes a purchase when offered the real choice has not been observed.[7]

2.2 Comparison to a Critical Value Based on Our Estimated Statistical Bias Function

The second type is based on calibration using a statistical bias function. Johannesson et al. (1999) estimate a statistical bias function based on experiments for two goods. Individuals were first offered the good hypothetically and then the same individuals were offered the goods for real. This sequence allows for within sample comparisons, i.e., for the same individuals. For all individuals who said yes in the contingent market, the probability of a hypothetical yes matching a real yes was estimated using the individual's self-assessed certainty as measured on a ten point scale and a variable representing the price of the good. The probability that a hypothetical yes was followed by a real yes was estimated. It was found that the probability increased with certainty and with the proportion of yes responses (representing lower price.) The statistical bias function was then used to calibrate hypothetical yes responses. If the stated certainty value was greater than the critical value based on the calibration function, then the hypothetical yes was counted as a true yes. If the stated certainty value was less than the critical value, then the hypothetical yes was considered a true no. Without calibration hypothetical bias was found, but after calibration there was no statistically significant difference between hypothetical and real responses.

[7] Johannesson et al. (1999) report results from two experiments in which subjects were first offered a good hypothetically in contingent valuation and then offered it for real purchase. Fifty-nine subjects from one experiment and 114 from the other experiment said no in contingent valuation. None of those subjects made a real purchase when it was offered. In the lipid management field experiment that will be described below, 39 subjects said no in contingent valuation and then were offered the real opportunity to purchase. None made the real purchase when offered. In total, none of the 212 hypothetical no responses were followed by a real purchase. In our experiments a hypothetical no, regardless of certainty, means a real no. This result is broadly consistent with Loomis and Ekstrand (1998) who find that calibration of only yes responses for certainty produces better fit of WTP logit regressions than recoding both yes and no responses. (Berrens et al. 2002, p. 158) also find a "yes means maybe and no means no" pattern.

In this paper we update the statistical bias function from Johannesson et al. (1999) by adding the 19 observations available from the lipid management program field experiment reported in this paper to the 99 observations from the chocolates and sunglasses experiments used in Johannesson et al. (1999) and reestimating the statistical bias function. Using the same parsimonious specification for the sample of 118 respondents who stated yes in the contingent valuation part of the experiment, we estimated the following probit regression:

$$\text{real yes/no} = \underset{(0.818)}{-4.635} + \underset{(0.091)}{0.530}\ \text{certainty scale value} + \underset{(0.617)}{1.869}\ \text{proportion yes} \qquad (1)$$

where standard errors are shown in parentheses and the constant and both coefficients are statistically significant at the 1% level.[8] The Chi-squared value equals 78.76 and is statistically significant at the 1% level. The McFadden's R-squared is 0.5211.

This updated statistical bias function is used to determine if a hypothetical yes is a true yes by comparing the calculated value for an individual to the critical value for the price that the individual faced. For example, for the diabetes management program, the critical value for an individual who was offered the program at a price of $80 (which had a proportion yes of 0.17), the critical value on the certainty scale is 8.15. A respondent who gave a certainty scale value of 10 would be considered to have given a true yes. If that respondent had been less certain and given a value of 5, then the answer would be considered a no. For all respondents who say yes in contingent valuation, the statistical bias function is used to determine if the yes should be considered a true yes.

2.3 Comparison to a Critical Value Based on Representative Studies: Eight or Greater

The third type of yes is based on a follow-up certainty question that offers a 10-point certainty scale. If the respondent answers yes and indicates a certainty value that is high enough on the 10-point scale, then the response is considered a true yes. The crucial decision concerns the critical value. With the statistical bias function, the second type of yes, the critical certainty value was determined by within sample comparisons between hypothetical yes responses and real decisions. For this third type of yes, the critical value is determined by the value that produces a good match between hypothetical and real decisions in studies by others. Field experiments for environmental goods by Champ et al. (1997) and Champ and Bishop (2001) find a good match between hypothetical and real donations for respondents who report that they are certain with value of at least 8 on the 10-point scale. Not all studies find exactly this result, but we believe it characterizes the studies that use the 10-point scale.[9] For now, a respondent who answers yes in contingent valuation and gives a value of 8 or greater will be considered to have given a true yes response. While we believe a value of at least 8 represents the spirit of using the 10-point scale for calibration, a critical value other than eight can be used. In fact, in Sect. 8 below we make comparisons for *all* values on the 10-point scale.

[8] The field experiment for the lipid management program was the only one of the three experiments that we consider in this paper that had respondents who sequentially were offered the good hypothetically and then for real. Therefore only the additional observations from the lipid experiment allow for within sample comparison.

We tried several specifications for the statistical bias function, but we kept the same specification as in our earlier study for ease of comparison and because no other specifications were obviously better than the simple specification based on standard criteria.

[9] For example, Poe et al. (2002) find the best match between hypothetical and real for certainty scale values greater than or equal to 7 or 8 depending on the criteria.

2.4 All Yeses

In contrast to the first three types of yes responses which are subsamples of hypothetical yes responses, the last type of yes response includes all hypothetical yes responses. This type of yes response is influenced by the quality of the contingent valuation study. It is influenced by ex ante mitigation measures such as reminders of related goods and the personal budget constraint and elicitation format. However, unlike the first three types of yes, it is not calibrated in an attempt to identify true yes responses.[10]

Before comparing results using these four types of hypothetical yes responses, we first describe briefly the three field experiments.

3 Three Field Experiments and Certainty of the Hypothetical Responses

We conducted three separate field experiments with three different goods. In each experiment a good was offered for hypothetical purchase and for real purchase. In the first two experiments, subjects were offered the good either hypothetically or for real, i.e., a split-sample design. In the third experiment, approximately half of the subjects were offered the good hypothetically and then for real, i.e., within-sample design, and the other half were offered the good for real only. Face-to-face interviews were used, and the subjects were asked to value a non-trivial, private good that was not available on the market. The good was described in detail and the description was read aloud by the interviewer while the subject followed along on a written description. The elicitation format was dichotomous choice with two options, "yes" and "no". The dichotomous choice contingent valuation question was followed by a question in which the subjects were asked to state how certain they were of their answers. This question appeared on the page following the willingness to pay question and was worded as follows:[11]

> If you answered YES, are you "probably sure" or "definitely sure" that you would buy the diabetes management service here and now at a price of $40? Please circle your answer below.

(A similar question was asked for subjects who answered no.) This definitely sure/probably sure question was followed by a question that asked subjects to state how certain they were on a 10-point visual analog scale. This question was worded as follows:

> If you answered YES, mark with an "*x*" on the line below how sure you are that you would buy the diabetes management service here and now at a price of $40.

[10] Other types of hypothetical yes responses could be defined with more data. In elicitation Evans et al. (2003) include "not sure" and Ready et al. (1995) include "maybe yes" as less certain than "probably yes" and "maybe no" as being less certain of no than "probably no". Follow up questions could be made using the same categories. Ready et al. (2001) find that responses from dichotomous choice and payment card formats for valuing changes in health converge for respondents who are certain. *If* incorporating certainty in the value elicitation and asking certainty in a follow up question have the same effect, then based on our previous studies we would expect that "definitely yes" is a true yes and all others are true no responses, see for example Blumenschein et al. (2008). Because we included only definitely sure and probably sure in follow up questions we leave tests for other types of yeses for future work.

[11] The price of $40 is shown as an example. Each subject was offered the good at only one price as is the practice in dichotomous choice contingent valuation. The price was varied across individuals so as to be able to estimate demand curves for the goods.

0	1	2	3	4	5	6	7	8	9	10
Very unsure									Very sure	

A similar question was asked for subjects who answered no. Both the definitely sure/probably sure and 10-point follow up certainty questions were asked of the same subjects. Because certainty was elicited in these two ways, we are able to compare the four different types of hypothetical yes responses. We can compare the three types of calibrated, "true" responses and compare them to all hypothetical (uncalibrated) yes responses.[12]

Three pharmacist-provided health management programs were offered in the field experiments. Diabetics were offered a diabetes management program through their local pharmacy in one experiment. Asthmatics were offered an asthma management program in a second, separate experiment. Individuals with heart and blood pressure problems were offered a lipid management program in a third, separate experiment. In each experiment, subjects were recruited from prescription patient lists at the participating pharmacies and the disease management programs were offered through a number of pharmacies in Kentucky.

All three studies used focus groups in development of the survey instrument and experimental protocol and all three were approved by the University of Kentucky Medical Institutional Review Board. Subjects were paid $25 for participating in the diabetes and asthma experiments and $20 for the lipid experiment. For the diabetes experiment, we use a sample of 181 subjects with 91 offered the program hypothetically and 90 for real at prices of $15, $40, and $80. They were interviewed during the period May 1 to July 23, 2003. For the asthma experiment, we use the entire sample of 172 subjects who were offered the program at prices of $15, $40, and $80 during the period October 1–November 19, 1999. For the lipid experiment, we use the entire sample of 114 who were offered the program at prices of $15 and $60 and were interviewed during the period November 5–December 21, 2000.[13]

Before making comparisons among the various types of yes responses, it is worth reporting how certain the subjects are about their responses. We focus only on the hypothetical yes responses because, as described above, a subject who responded with a hypothetical no and then made a real purchase has not been observed in the three experiments.

For the diabetes experiment, 53.7% (46.3%) of the hypothetical yes respondents are definitely (probably) sure. Figure 1a shows the distribution of certainty scale values for hypothetical yes respondents. More than half (53.7%) of the hypothetical yes respondents have a certainty scale value of 10. Figure 1b shows the distributions of certainty scale values for hypothetical yes respondents who are probably sure and definitely sure. As expected, definitely sure yes responses have certainty scale values closer to 10 than the probably sure. The mean certainty scale value for definitely sure is 9.73 (standard deviation = 0.77) while the mean for the probably sure is only 7.47 (standard deviation = 2.37). The Spearman correlation between definitely sure (=1) and the certainty scale value is 0.683.

[12] Champ et al. (1997) and Champ and Bishop (2001) use a 1–10 scale. We have been using a 0–10 scale because it is easy to mark the midpoint of 5. Solving the equation $y = -(10/9) + (10/9)x$ for when $y = 8$ yields that the value of 8 on the 0–10 scale is comparable to 8.2 on the 1–10 scale. For comparison of this calibration to others, we think the difference seems inconsequential.

[13] In this study we use 181 of the 267 subjects used in Blumenschein et al. (2008) which contains a full description of the diabetes experiment. In this study we do not include the 86 subjects who were read a cheap talk script in contingent valuation and included in the earlier study. Additional information about the asthma experiment is reported in Blumenschein et al. (2001). Additional information about the lipid experiment can be found in the Blumenschein and Johannesson (2001) report that is available upon request.

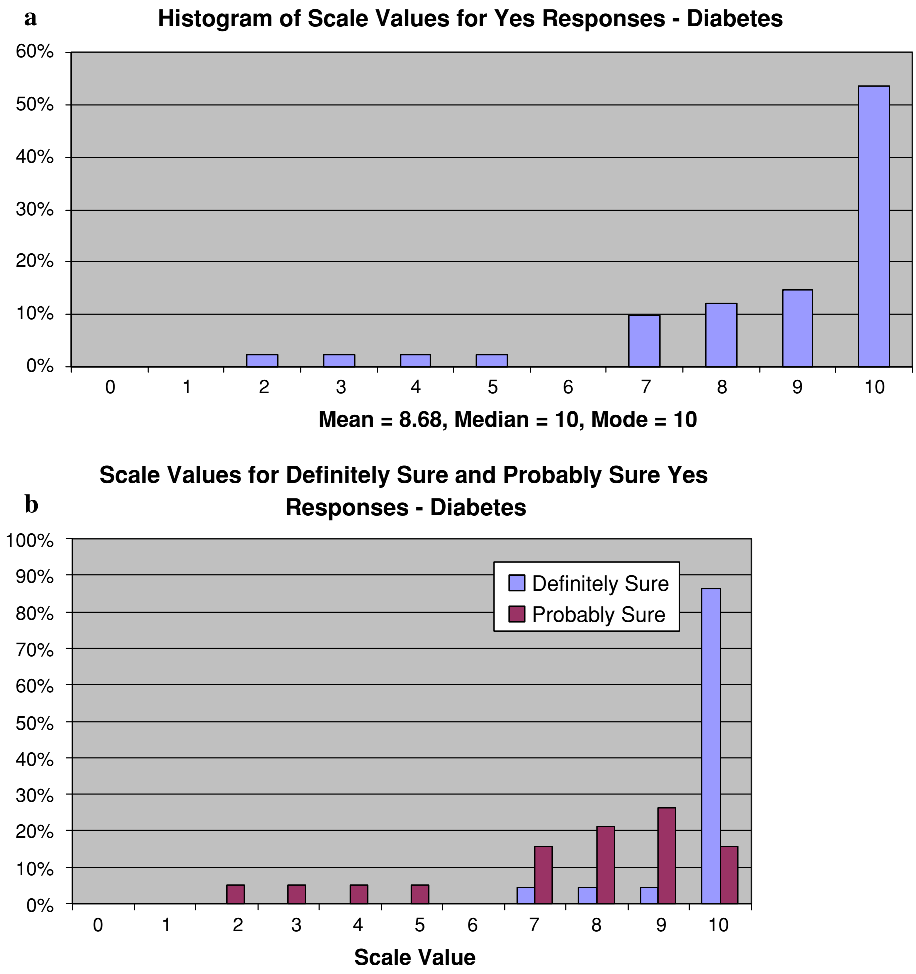


Fig. 1 **a** Distribution of certainty scale values for hypothetical yes responses, diabetes. **b** Distribution of certainty scale values for definitely sure and probably sure hypothetical yes responses, diabetes

For the asthma experiment, 35.5% (62.5%) of the hypothetical yes respondents are definitely (probably) sure. Figure 2a shows the distribution of certainty scale values for hypothetical yes respondents. Many (75.0%) of the hypothetical yes respondents have a certainty scale value of 10. Figure 2b shows the distributions of certainty scale values for hypothetical yes respondents who are probably sure and definitely sure. While only 35.5% were definitely sure, the mean certainty scale value for the definitely sure is 9.50 (standard deviation = 0.77) which is greater than the 6.50 (standard deviation = 2.54) for the probably sure. The Spearman correlation between definitely sure (=1) and the certainty scale value is 0.667.

For the lipid experiment, 79.0% (21.0%) of the hypothetical yes respondents are definitely (probably) sure. Figure 3a shows the distribution of certainty scale values for hypothetical yes respondents. Nearly all (93.3%) of the hypothetical yes respondents have a certainty scale value of 10. Figure 3b shows the distributions of certainty scale values for hypothetical yes respondents who are probably sure and definitely sure. The mean certainty scale value for

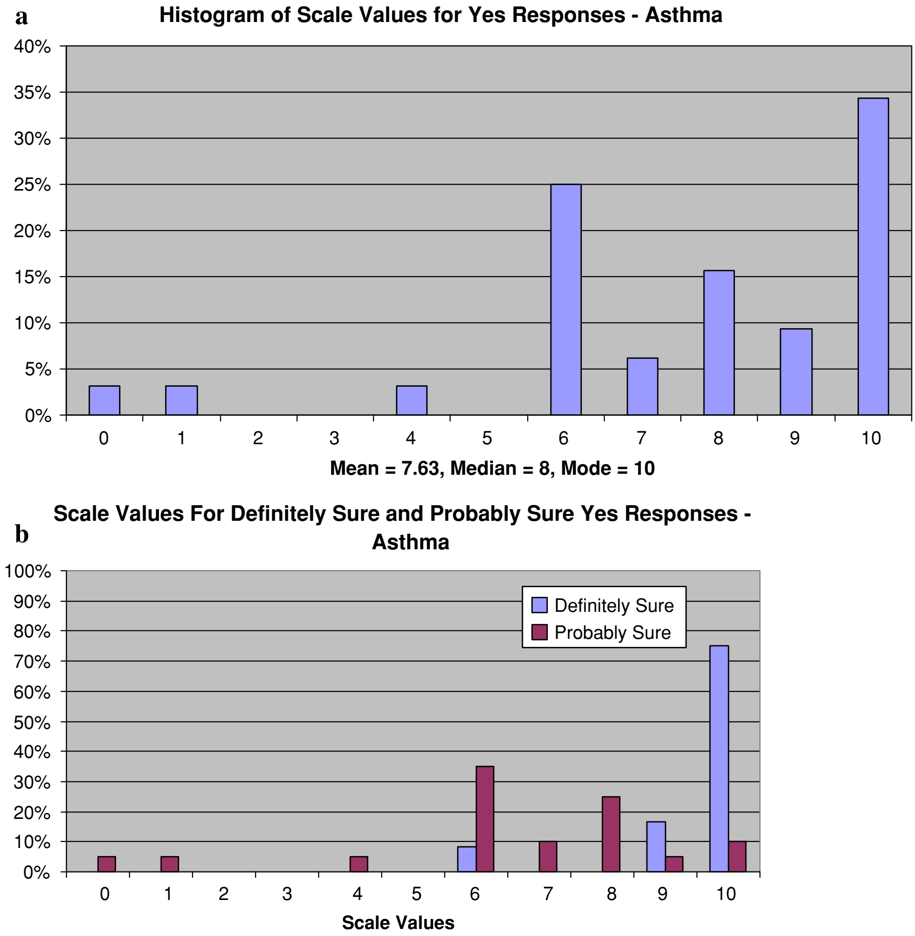


Fig. 2 **a** Distribution of certainty scale values for hypothetical yes responses, asthma. **b** Distribution of certainty scale values for definitely sure and probably sure hypothetical yes responses, asthma

definitely sure of 9.93 (standard deviation = 0.26) is much greater than the 4.75 (standard deviation = 3.30) for the probably sure. The Spearman correlation between definitely sure (=1) and the certainty scale value is 0.912.

For all three disease management programs there is variation in the degree of certainty respondents express regarding how sure they are they would actually make the purchase. For all three programs at least half of the respondents who said yes in contingent valuation give a certainty scale value of 10. The mean scale value for respondents who are definitely sure is greater than the mean for those who are only probably sure, and the correlation between definitely sure and the certainty scale value is at least 0.6.

Given the variation in certainty of hypothetical yes responses, data from these three field experiments facilitate comparisons of the three types of hypothetical yes responses that are calibrated to be true yes responses and also comparison to all hypothetical yes responses. We make a number of comparisons for each of the three field experiments. First, we compare observable subject characteristics. Second, we compare the percentage of yes respondents, the

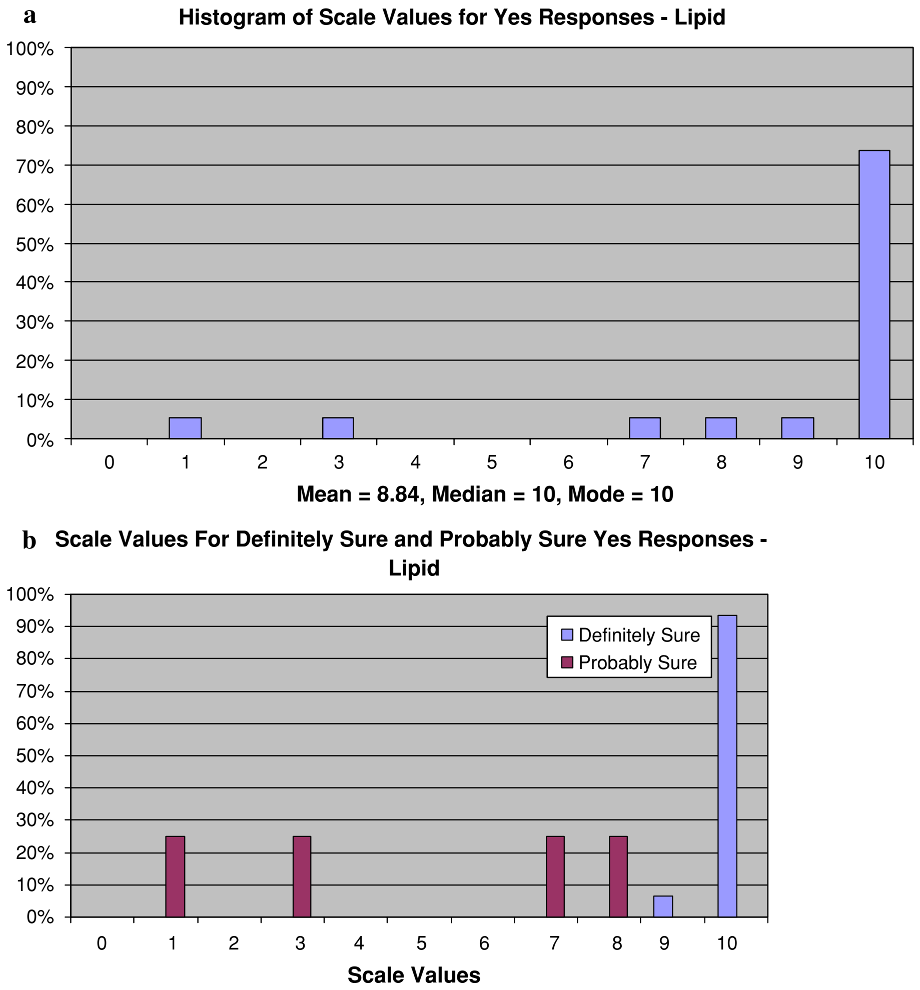


Fig. 3 **a** Distribution of certainty scale values for hypothetical yes responses, lipid. **b** Distribution of certainty scale values for definitely sure and probably sure hypothetical yes responses, lipid

performance of the dummy variable Hypothetical in logit regressions of all yes responses, and the mean WTP. Third, we report our estimate of the value on the 10-point scale that produces the same calibrated mean WTP as definitely sure, and we report our estimate of the value on the 10-point scale that produces the same calibrated mean WTP as mean WTP from real purchases.

4 Differences in Observable Characteristics of Subjects

Champ and Bishop (2001) find that when hypothetical and real donations to generating electricity using wind power are compared, a cutoff value of 8 on the 10-point scale yielded a mean WTP that was indistinguishable from the mean of real payments. They report that, in addition, respondents who are willing to pay and have certainty scale value of at least 8 are similar across a range of measures including attitudes, experience, and demographics. They attach importance to this result because it is an indication that the follow up certainty calibration works to separate individuals who will really pay from individuals who just say they will.

We compare observable characteristics of subjects for each of the three health goods. Table 1 shows the means of background characteristics for the diabetes management program. Means are shown for the four types of hypothetical yes responses: (1) calibrated with definitely sure, (2) calibrated with the statistical bias function, (3) calibrated with 8 or greater on the 10-point certainty scale, and (4) all hypothetical yes responses. Also shown are the means for subjects who actually made real purchases. This last group is not the same as the experimental group that was offered the program for real because some of those subjects declined. Table 2 shows the means of background characteristics for subjects who said yes for the asthma management program and Table 3 gives similar information for subjects who said yes for the lipid management program. While there are some differences in the means of some of these characteristics, what is not clear is any pattern among the four types of hypothetical yeses. One might expect the all yes means to be different from the calibrated means, but differences are not striking. Champ and Bishop (2001) have several variables that measure attitudes towards the environment. We do not have similar attitudinal variables for health and that may account for the difference in our findings about respondent characteristics.

5 Differences in Percentages of Yes Responses

Another way to compare the four types of yes responses to hypothetical purchase questions is to compare the percentages who say yes. This comparison can show how much difference the calibrations make. Yes responses for the diabetes management program are shown in Table 4. For the hypothetical yeses, the percentage yes tends to increase as we move from left to right. For example, at the price of $15 the percentage rises from 35% for definitely sure, to 58% for the statistical bias function and 55% for 8 or greater on the certainty scale, to 71% for all hypothetical yeses. For all three prices combined, the yeses rise from 24% for definitely sure, to 36% for the statistical bias function and for 8 or greater on the certainty scale, to 45% for all yeses. There is no indication of statistically significant bias for any of the three calibrated yeses, but there is hypothetical bias for all (uncalibrated) yeses at the 5% level.

Yes responses for the asthma management program shown in Table 4. For the hypothetical responses, the percentage yes again tends to rise as we move from left to right. For example, at the price of $40 the percentage rises from 9% for the three calibrated yeses to 29% for all yeses. For all three prices combined, the yeses rise from 14% for the definitely sure, to 23% for both the statistical bias function and 8 or greater on the certainty scale, to 38% for all yeses. There is no statistically significant hypothetical bias for the yeses calibrated by definitely sure. There is weak evidence of hypothetical bias for calibration by the statistical bias function and 8 or greater on the certainty scale at the 10% level. There is strong evidence of hypothetical bias for all yeses at the 1% level.

Yes responses for the lipid management program are shown in Table 4. Again for the hypothetical responses, the percentage tends to rise as we move from left to right. For both prices combined, the yeses rise from 26% for definitely sure, to 28% for the statistical bias function and 8 or greater on the certainty scale, to 32% for all yeses. While the pattern is

Table 1 Comparison of background characteristics of respondents who stated yes to purchase[a]-Diabetes management program

	Real	Hypothetical[b]			
	Yeses ($n = 23$)	Definitely sure yeses ($n = 22$)	Function yeses ($n = 33$)	Certain 8 yeses ($n = 33$)	All yeses ($n = 41$)
Income and wealth					
Annual household income ($1,000)	32.62 (23.67)	40.12 (25.68)	39.06 (26.13)	38.75 (26.36)	35.94 (26.19)
Household size	2.04 (1.22)	2.55 (1.22)	2.63* (1.32)	2.67* (1.29)	2.61* (1.26)
Owns residence (%)	86.96	95.45	90.91	90.91	90.24
Health and health behavior					
Previous participation in disease management (%)	17.39	9.10	9.10	12.12	9.77
Member of diabetes support group (%)	8.70	4.55	6.10	6.06	4.88
Time with diabetes (years)	7.87 (7.09)	7.25 (6.40)	7.54 (6.39)	7.70 (6.27)	7.63 (5.81)
Diabetes severity					
Mild (%)	21.74	18.18	18.18	15.15	19.51
Moderate (%)	69.56	63.63	66.67	66.67	63.41
Severe (%)	8.70	18.18	15.15	18.18	17.07
Cardiovascular disease (%)	86.70	86.36	81.82	84.85	85.37
Renal disease (%)	8.70	4.55	3.03	3.03	2.44
Vision problems (%)	30.43	22.73	27.27	30.30	29.27
Neuropathies (%)	60.87	68.18	57.58	57.58	58.54
Complications of diabetes in family (%)	43.48	63.64	60.61	60.61	60.98
Smoking (%)	30.43	13.64	12.12*	15.15	14.63
Body mass index	36.63 (8.35)	31.54** (7.83)	31.66** (7.05)	31.91** (7.08)	32.45** (6.97)
Know their hemoglobin A_1C level (%)	30.43	22.73	21.21	24.24	21.95
General health					
Excellent (%)	4.35	0	0	0	0
Very good (%)	4.35	18.18	15.15	15.15	12.20
Good (%)	21.74	36.36	36.36	36.36	39.02
Fair (%)	39.13	40.91	45.45	45.45	41.46
Poor (%)	30.43	4.55**	3.03***	3.03***	7.32**
Socioeconomics					
Age (years)	57.87 (10.91)	60.32 (12.70)	60.12 (13.15)	60.72 (11.88)	59.39 (12.63)
Women (%)	65.22	63.64	69.70	69.70	70.73
Education (years)	10.70 (4.26)	12.25 (2.75)	12.26 (2.92)	12.09 (2.77)	12.26* (2.77)
Ethnic background white (%)	86.96	90.91	93.94	93.94	90.24

Table 1 continued

	Real	Hypothetical[b]			
	Yeses ($n = 23$)	Definitely sure yeses ($n = 22$)	Function yeses ($n = 33$)	Certain 8 yeses ($n = 33$)	All yeses ($n = 41$)
Time cost					
Travel time to pharmacy (min)	13.33 (10.05)	12.63 (4.61)	13.18 (5.29)	13.55 (4.97)	13.31 (5.61)

[a] Means (standard deviations in parentheses)

[b] The asterisks *, **, *** indicate different from responses in the real group at the 10%, 5%, and 1% level of significance, respectively

similar to that for diabetes and asthma programs, there is no statistically significant indication of hypothetical bias.

For the three experiments, the overall pattern of yes responses is clear. For the hypothetical responses, the percentage yes tends to increase as we move from calibration by definitely sure to the statistical bias function and 8 or greater on the certainty scale to all yeses. There is evidence of hypothetical bias for the diabetes management and asthma management programs but not the lipid program. All three calibrations remove the hypothetical bias for both the diabetes and asthma management programs if the 5% level of statistical significance is used. There is weak evidence that the statistical bias function and the 8 or greater on the certainty scale calibrations do not eliminate hypothetical bias for the asthma management program if the 10% level of significance is used. Using definitely sure to identify true yes responses produces a set of yeses that give no indication of hypothetical bias at any of the usual levels of statistical significance.

6 Differences in How the Hypothetical Variable Performs in Logit Regressions

In Sect. 4 above the differences in observable characteristics of subjects in the field experiments were examined by type of hypothetical yes response. Although no striking differences or obvious patterns emerged, subtle differences may be influencing the comparisons of percentages of yes responses just described in the previous section, Sect. 5. If any differences are important, we should be able to detect them with logit regressions of all hypothetical and real responses in which we use the observable characteristics as explanatory variables. The variable of interest is a dummy variable (Hypothetical) that takes on a value of 1 if the individual is in the experimental group offered the disease management program hypothetically in contingent valuation or the value of 0 if the individual is in the experimental group that was offered the program for real purchase.

Table 5 reports logit regression results for the diabetes management program. In addition to the Hypothetical variable, and variables for price and household income, there are 23 control variables. They include characteristics such as education, age, time with diabetes, and measures of health status. The Hypothetical variable always represents the hypothetical group, but what is counted as a true hypothetical yes changes as we move from left to right across the table. All (uncalibrated) hypothetical yeses are counted as yes in the first column of results. The next column to the right considers only definitely sure yeses as true yeses. The next column to the right calibrates yeses by the statistical bias function and the right-most column calibrates yeses by 8 or greater on the certainty scale. The positive

Table 2 Comparison of background characteristics of respondents who stated yes to purchase[a]—Asthma management program

	Real	Hypothetical[b]			
	Yeses ($n = 11$)	Definitely sure yeses ($n = 12$)	Function yeses ($n = 19$)	Certain 8 yeses ($n = 19$)	All yeses ($n = 32$)
Annual household income ($1,000)	15.00 (13.96)	19.17 (22.24)	13.42 (18.86)	14.47 (18.92)	17.34 (25.34)
Time with asthma (years)	18.59 (18.66)	15.10 (16.31)	19.01 (18.13)	19.28 (18.15)	18.45 (15.72)
Asthma severity					
Mild (%)	27.27	8.33	10.53	10.53	12.50
Moderate (%)	45.46	50.00	42.10	42.11	50.00
Severe (%)	27.27	41.67	47.37	47.37	37.50
Age (years)	49.18 (16.20)	58.83 (10.67)	54.00 (13.80)	54.37 (12.85)	52.72 (17.26)
Women (%)	90.91	75.00	89.47	89.47	75.00
Education (years)	11.86 (3.18)	9.58 (4.32)	9.11* (3.80)	9.26* (3.94)	10.56 (3.92)

[a] Means (standard deviations in parentheses)
[b] The asterisk * indicates different from responses in the real group at the 10% level of significance

Table 3 Comparison of background characteristics of respondents who stated yes to purchase[a]—Lipid management program

	Real	Hypothetical[b]			
	Yeses (n = 14)	Definitely sure yeses (n = 15)	Function yeses (n = 16)	Certain 8 yeses (n = 16)	All yeses (n = 19)
Annual household income ($1,000)	32.14 (25.17)	38.33 (34.31)	38.44 (33.15)	38.44 (33.15)	37.37 (32.89)
Severity					
Mild (%)	14.29	6.67	6.25	6.25	10.53
Moderate (%)	50.00	73.33	75.00	75.00	73.68
Severe (%)	35.71	20.00	18.75	18.75	15.79
Smoke	14.29	20.00	18.75	18.75	15.79
Know Cholesterol (%)	42.86	46.67	43.75	43.75	36.84
Age (years)	59.86 (12.51)	60.67 (12.33)	61.38 (12.24)	61.38 (12.24)	61.63 (11.19)
Women (%)	64.29	60.00	62.50	62.50	63.15
Education (years)	12.98 (2.94)	14.13 (2.23)	14.00 (2.22)	14.00 (2.22)	14.16 (2.52)

[a] Means (standard deviations in parentheses)

[b] No means are significantly different from responses in the real group at the 10% level of significance

Table 4 Number (%) of yes responses. Four definitions of yes for the hypothetical group

	Real	Hypothetical							
Price	Real yeses	Yes only if definitely sure yes		Yes only if yes and predicted yes by statistical bias function		Yes only if yes and 8 or greater on Certainty Scale		All yeses	
	Number (%)	Number (%)	p-value[a]	Number (%)	p-value[a]	Number (%)	p-value[a]	Number (%)	p-value[a]
Diabetes management program									
$15	13/29 (45)	11/31 (35)	0.460	18/31 (58)	0.305	17/31 (55)	0.438	22/31 (71)	0.040
$40	7/30 (23)	11/34 (32)	0.423	13/34 (38)	0.199	13/34 (38)	0.199	14/34 (41)	0.129
$80	3/31 (10)	0/26 (0)	0.103	2/26 (8)	0.792	3/26 (12)	0.820	5/26 (19)	0.301
All	23/90 (26)	22/91 (24)	0.830	33/91 (36)	0.119	33/91 (36)	0.119	41/91 (45)	0.006
Asthma management group									
$15	6/37 (16)	9/32 (28)	0.232	14/32 (44)	0.012	13/32 (40)	0.024	19/32 (59)	0.000
$40	5/36 (14)	3/34 (9)	0.506	3/34 (9)	0.506	3/34 (9)	0.506	10/34 (29)	0.114
$80	0/16 (0)	0/18 (0)	–	2/18 (11)	0.169	3/18 (17)	0.087	3/18 (17)	0.087
All	11/89 (12)	12/84 (14)	0.709	19/84 (23)	0.075	19/84 (23)	0.075	32/84 (38)	0.000
Lipid management program									
$15	8/28 (29)	12/27 (44)	0.221	13/27 (48)	0.135	13/27 (48)	0.135	15/27 (56)	0.043
$60	6/28 (21)	3/31 (10)	0.210	3/31 (10)	0.210	3/31 (10)	0.210	4/31 (13)	0.383
All	14/56 (25)	15/58 (26)	0.916	16/58 (28)	0.754	16/58 (28)	0.754	19/58 (32)	0.361

[a] p-value of the difference compared to the yes responses in the real group

and significant coefficient on Hypothetical indicates hypothetical bias in uncalibrated yeses. All three calibrations remove indications of statistically significant hypothetical bias. The coefficient on price is negative and statistically significant in all regressions.

The logit regression results for the asthma management program are reported in Table 5. In addition to the Hypothetical variable and variables for price and household income, there are six control variables. The positive and statistically significant coefficient on Hypothetical indicates hypothetical bias in uncalibrated yeses. All three calibrations remove evidence of hypothetical bias if the criterion is the 5% level of significance. There is weak evidence that the statistical bias function and 8 or greater on the certainty scale calibrations do not remove hypothetical bias if the criterion is the 10% level of significance. The coefficient on price is negative and statistically significant in all regressions.

The logit regressions for the lipid management program are shown in Table 5. In addition to the Hypothetical variable and variables for price, and household income, there are seven individual characteristics that are used as control variables. The Hypothetical variable is positive, but not statistically different from zero for all yeses and all three calibrated yeses. The point estimate is greatest for all yeses and smallest for definitely sure, but none is statistically significant.

For the three experiments, overall comparisons of the Hypothetical dummy variable in logit regressions that control for characteristics of the subjects mimic the comparisons of the percentage yes responses. There is evidence of hypothetical bias for the diabetes management and asthma management programs. Definitely sure, statistical bias function, and 8 or greater on the certainty scale calibrations remove the hypothetical bias in both diabetes and asthma

Table 5 Results of logistic regression analysis of the effect of experimental group on the probability of a yes response

	All yes responses in hypothetical group (Equation 1)		Definitely sure yes responses in hypothetical group (Equation 2)		Calibrated yes responses in hypothetical group (Equation 3)		Yes with certainty 8 or greater in hypothetical group (Equation 4)	
	β	SE	β	SE	β	SE	β	SE
Diabetes management study[a]								
Constant	−3.354	2.710	−0.365	2.999	−2.536	2.752	−3.810	2.776
Hypothetical group	1.215	0.454***	−0.095	0.472	0.543	0.452	0.538	0.456
Price	−0.046	0.010***	−0.051	0.012***	−0.049	0.010***	−0.046	0.010***
Household income	0.022	0.009**	0.028	0.010***	0.024	0.009***	0.026	0.009***
Household size	−0.337	0.193*	−0.507	0.224**	−0.282	0.193	−0.230	0.195
Owns residence	1.004	0.642	1.594	0.789**	0.972	0.686	0.932	0.690
Disease management	0.982	0.711	1.329	0.818	0.936	0.743	1.392	0.749*
Diabetes support group	3.044	1.558*	2.121	1.294	3.504	1.620**	3.175	1.575**
Time with diabetes	−0.057	0.037	−0.083	0.041**	−0.069	0.038*	−0.074	0.038*
Moderate diabetes[b]	1.087	0.558*	1.198	0.631*	1.238	0.569**	1.397	0.577**
Severe diabetes[b]	1.904	0.834	2.017	0.961**	1.862	0.870**	2.312	0.868***
Cardiovascular disease	−0.161	0.600	−0.400	0.662	−0.639	0.598	−0.400	0.610
Renal disease	−0.189	0.948	0.676	1.008	0.301	0.941	0.011	0.937
Vision problems	−0.844	0.508*	−1.304	0.574**	−1.010	0.523*	−0.853	0.520
Neuropathies	0.058	0.494	0.850	0.542	0.035	0.497	−0.081	0.502
Complications of diabetes in family	−0.076	0.437	−0.094	0.456	−0.072	0.438	−0.100	0.440
Smoking	0.283	0.518	0.251	0.568	0.029	0.536	0.333	0.531
Body mass index	0.059	0.030**	0.051	0.032	0.046	0.030	0.053	0.030*
Know hemoglobin A_1C	0.779	0.568	0.679	0.596	0.706	0.571	0.930	0.568

Table 5 continued

	All yes responses in hypothetical group (Equation 1)		Definitely sure yes responses in hypothetical group (Equation 2)		Calibrated yes responses in hypothetical group (Equation 3)		Yes with certainty 8 or greater in hypothetical group (Equation 4)	
	β	SE	β	SE	β	SE	β	SE
Good health[c]	1.909	0.780**	1.250	0.774	1.473	0.755*	1.517	0.768*
Fair health[c]	1.366	0.791*	0.619	0.765	1.153	0.767	1.130	0.780
Poor health[c]	1.667	0.970*	0.610	1.022	1.004	0.971	0.867	0.982
Age	0.011	0.022	0.004	0.023	0.017	0.022	0.028	0.022
Woman	0.210	0.448	−0.020	0.478	0.200	0.448	0.259	0.448
Education	−0.621	0.083	−0.152	0.097	−0.095	0.087	−0.121	0.088
Ethnic: white	−0.944	0.723	−1.232	0.781	−0.319	0.749	−0.449	0.748
Travel time	0.181	0.027	−0.015	0.030	0.009	0.028	0.020	0.028
Number of obs.[d]	181		181		181		181	
$\chi^2 (p - \text{value})$	72.72	(<0.0001)	59.83	(0.0002)	65.55	(<0.0001)	66.23	(<0.0001)
Log-likelihood	−81.223		−71.59		−79.1923		−78.85	
McFadden pseudo-R^2	0.3092		0.2947		0.2927		0.2958	
Asthma management study[a]								
Constant	−0.616	1.292	−0.792	1.678	−0.425	1.564	−0.905	1.537
Hypothetical group	1.590	0.425***	0.009	0.499	0.763	0.454*	0.753	0.447*
Price	−0.032	0.010***	−0.045	0.016***	−0.035	0.012***	−0.026	0.011**
Household income	0.007	0.009	0.012	0.012	0.006	0.011	0.007	0.010
Time with asthma	−0.003	0.012	−0.012	0.015	−0.002	0.013	−0.002	0.013
Moderate asthma	0.327	0.561	0.525	0.690	0.164	0.632	0.098	0.625
Severe asthma	0.283	0.612	0.273	0.742	0.227	0.673	0.226	0.666

Table 5 continued

	All yes responses in hypothetical group (Equation 1)		Definitely sure yes responses in hypothetical group (Equation 2)		Calibrated yes responses in hypothetical group (Equation 3)		Yes with certainty 8 or greater in hypothetical group (Equation 4)	
	β	SE	β	SE	β	SE	β	SE
Age	0.001	0.013	0.022	0.017	0.003	0.015	0.005	0.015
Woman	0.045	0.489	0.068	0.634	0.984	0.683	0.990	0.678
Education	−0.061	0.065	−0.112	0.084	−0.147	0.079*	−0.133	0.076*
Number of obs.[e]	173		173		173		173	
$\chi^2(p - \text{value})$	30.28	(0.0004)	16.82	(0.0516)	22.44	(0.0076)	17.94	(0.0359)
Log-likelihood	−81.86		−59.40		−68.57		−70.83	
McFadden pseudo-R^2	0.1561		0.1240		0.1406		0.1124	
Lipid management study[a]								
Constant	−2.592	2.493	−2.397	2.483	−2.663	2.505	−2.663	2.505
Hypothetical group	0.429	0.471	0.016	0.477	0.102	0.476	0.102	0.476
Price	−0.028	0.011**	−0.026	0.011**	−0.027	0.012**	−0.027	0.012**
Household income	0.005	0.010	0.003	0.011	0.006	0.010	0.006	0.010
Moderate lipid	1.035	0.681	1.099	0.728	1.265	0.735*	1.265	0.735*
Severe lipid	1.645	0.822**	1.849	0.848**	1.946	0.859**	1.946	0.859**
Know cholesterol	−1.148	0.510**	−0.739	0.511	−0.868	0.512*	−0.868	0.512*
Smoke	−0.339	0.650	−0.005	0.644	−0.096	0.648	−0.096	0.648
Age	0.007	0.025	0.005	0.025	0.011	0.025	0.011	0.025
Woman	0.127	0.526	−0.062	0.531	0.058	0.530	0.058	0.530
Education	0.109	0.100	0.090	0.101	0.069	0.100	0.069	0.100
Number of obs.[f]	114		114		114		114	

Table 5 continued

	All yes responses in hypothetical group (Equation 1)		Definitely sure yes responses in hypothetical group (Equation 2)		Calibrated yes responses in hypothetical group (Equation 3)		Yes with certainty 8 or greater in hypothetical group (Equation 4)	
	β	SE	β	SE	β	SE	β	SE
$\chi^2(p-\text{value})$	21.49	(0.0179)	15.29	(0.1217)	17.40	(0.0659)	17.40	(0.0659)
Log-likelihood	−57.846		−57.003		−57.001		−57.001	
McFadden pseudo-R^2	0.1567		0.1183		0.1324		0.1324	

[a] Statistical significance is indicated by *** at the 1% level, ** at the 5% level, and * at the 10% level

[b] Baseline category: mild diabetes

[c] Baseline category: excellent or very good general health

[d] The number of subjects offered the diabetes management program for real is 90

[e] The number of subjects offered the asthma management program for real is 89

[f] The number of subjects offered the lipid management program for real is 56

management programs if the 5% level of statistical significance is used. There is weak evidence that the statistical bias function and the 8 or greater on the certainty scale calibrations do not eliminate hypothetical bias for the asthma management program if the 10% level of significance is used. Using definitely sure to identify true yes responses produces a set of yeses that give no indication of hypothetical bias at any of the usual levels of statistical significance.

7 Differences in Estimates of Average Willingness to Pay

Given the interest in using estimates of WTP in benefit-cost analysis, comparison of estimates of average WTP is the crux of the matter. We compare mean WTP for all hypothetical yeses and each of the three calibrated hypothetical yes responses. We also compare the estimates to the mean WTP of real purchases to check for hypothetical bias. We make these comparisons for each of the three disease management programs using both nonparametric and parametric methods.

For the nonparametric estimates of WTP we use a method developed by Kriström (1990) in which the area under the demand curve (price and fraction of buyers) is an estimate of the mean. For each experiment, we assume that the maximum WTP equals the highest price (either \$60 or \$80) used in the experiment and that the proportion of subjects with zero WTP equals the proportion of responses at the lowest price (\$15) used in the experiment. For the parametric estimates of WTP, we use a method developed by Johansson (1995) that restricts the value to be positive which is appropriate for these private disease management programs that do not have to be consumed. The estimation of mean willingness to pay is based on estimating the area below the demand curve using the formula: $-(1/\beta)\ln(1+e^{\alpha})$, where β is the price coefficient in the logistic regression equation and α is the constant in the logistic regression equation with the effect of all other covariates evaluated at their means and added to the constant. The results, while not identical, are similar for the nonparametric and parametric methods, and so our discussion will focus on the parametric estimates of mean WTP.

Estimates of mean WTP for the diabetes management program for the different types of yes responses are reported in Table 6. As we read down the column for the different types of yeses, the mean WTP tends to increase. For the parametric estimates, the mean increases from \$17.36 for the definitely sure yeses, to \$28.72 and \$28.64 for the yeses calibrated by the statistical bias function and 8 or greater on the certainty scale, to \$38.90 for all hypothetical (uncalibrated) yeses. There is no evidence of hypothetical bias for the calibrated yeses, but there is evidence at the 5% level for all hypothetical yeses.

Estimates of mean WTP for the asthma management program for the different types of yeses are shown in Table 6. Again as we read down the column for the different types of yeses, the mean WTP increases. For the parametric estimates, the mean increases from \$9.69 for the definitely sure, to \$17.37 for the statistical bias function, to \$19.03 for the 8 or greater on the certainty scale, to \$33.24 for all yeses. There is no evidence of hypothetical bias for the definitely sure yeses. There is evidence of hypothetical bias at the 10% level for yeses calibrated by the statistical bias function and it is close to the 10% level for 8 or greater. There is evidence at the 5% level for all yeses.

For the lipid management program estimates of mean WTP for the different types of yeses are reported in Table 6. As with the other two goods, as we read down the column for the different types of yeses, the estimate of mean WTP tends to increase. For the parametric estimates, the mean WTP increases from \$21.76 for definitely sure, to \$22.65 for the calibration

Table 6 Willingness to pay (dollars). Four definitions of yes for the hypothetical yes group[a]

		Non-parametric method[b]	Parametric method[c]
Diabetes management program			
Real	Mean	21.85	19.78
	(Standard error)	(3.77)	(4.30)
Definitely sure yes	Mean	20.27	17.36
	(Standard error)	(3.60)	(3.88)
	p-value	0.767	0.840
	(Diff. 95% conf. int.)	(−11.95 to 8.81)	(−11.94 to 9.71)
Statistical bias function yes	Mean	29.93	28.72
	(Standard error)	(3.73)	(4.66)
	p-value	0.134	0.229
	(Diff. 95% conf. int.)	(−2.53 to 18.70)	(−4.88 to 20.39)
Certainty scale ≥ 8 yes	Mean	29.81	28.64
	(Standard error)	(3.81)	(4.84)
	p-value	0.144	0.273
	(Diff. 95% conf. int.)	(−2.42 to 18.36)	(−5.19 to 20.96)
All yeses	Mean	36.74	38.90
	(Standard error)	(3.81)	(5.35)
	p-value	0.006	0.007
	(Diff. 95% conf. int.)	(4.27–25.52)	(5.33–32.92)
Asthma management program			
Real	Mean	8.97	10.27
	(Standard error)	(2.54)	(3.02)
Definitely sure yes	Mean	10.60	9.69
	(Standard error)	(2.62)	(2.90)
	p-value	0.662	0.985
	(Diff. 95% conf. int.).	(−5.58 to 8.84)	(−7.58 to 7.73)
Statistical bias function yes	Mean	17.08	17.37
	(Standard error)	(3.32)	(4.18)
	p-value	0.058	0.098
	(Diff. 95% conf. int.)	(0.10–16.10)	(−1.47 to 17.41)
Certain 8 yes	Mean	17.23	19.03
	(Standard error)	(3.41)	(5.23)
	p-value	0.056	0.106
	(Diff. 95% conf. int.)	(−0.23 to 16.74)	(−1.87 to 19.65)
All yeses	Mean	29.22	33.24
	(Standard error)	(4.04)	(6.14)
	p-value	0.000	0.000
	(Diff. 95% conf. int.)	(11.50–28.99)	(10.27–35.67)
Lipid management program			
Real	Mean	15.53	20.41
	(Standard error)	(3.54)	(5.88)

Table 6 continued

		Non-parametric method[b]	Parametric method[c]
Definitely sure yes	Mean	18.84	21.76
	(Standard error)	(3.74)	(6.68)
	p-value	0.538	0.973
	(Diff. 95% conf. int.)	(−7.22 to 13.83)	(−15.21 to 15.74)
Statistical bias function yes	Mean	20.23	22.65
	(Standard error)	(3.75)	(6.61)
	p-value	0.381	0.830
	(Diff. 95% conf. int.)	(−5.80 to 15.20)	(−13.67 to 17.03)
Certainty scale $\geq$ 8 yes	Mean	20.23	22.65
	(Standard error)	(3.75)	(6.61)
	p-value	0.381	0.830
	(Diff. 95% conf. int.)	(−5.80 to 15.20)	(−13.67 to 17.03)
All yeses	Mean	23.74	27.87
	(Standard error)	(3.83)	(7.34)
	p-value	0.124	0.368
	(Diff. 95% conf. int.)	(−2.24 to 18.65)	(−8.77 to 23.69)

[a] *p*-value of the difference compared to the mean willingness to pay in the real group and 95% confidence interval of the difference

[b] Non parametric standard errors and confidence intervals are based on 1000 bootstrap replications

[c] Parametric standard errors and confidence intervals are based on the delta method

by the statistical bias function and 8 or greater on the certainty scale, to $27.87 for all yeses. None of the estimates is statistically different from the estimate of mean WTP based on real purchases.

For the three experiments, overall comparison shows that estimates of mean WTP tend to increase from yeses calibrated by definitely sure, to estimates based on either the statistical bias function or 8 or greater on the certainty scale, to all (uncalibrated) yeses. For all yeses there is strong evidence of hypothetical bias for the diabetes and asthma management programs, but not for the lipid program. All three of the calibration methods remove evidence of statistically significant hypothetical bias at usual levels of confidence except for weak evidence for the statistical bias function for asthma. Calibration by definitely sure produces point estimates of mean WTP closest to the mean of real WTP for all three disease management programs.

8 What on the 10-point Certainty Scale is Equivalent to Definitely Sure and Real?

We have made several comparisons of the differences associated with calibrating hypothetical yes responses using three different ex post certainty techniques. However, the primary purpose of this paper is to compare two ways of asking follow-up certainty questions, definitely/probably sure and the 10-point certainty scale. The comparisons for the 10-point, follow up certainty scale were based on assuming that a true yes is one in which the subject indicated a certainty scale value of 8 or higher. While that may be representative, we now estimate which values on the 10-point scale are equivalent to definitely sure for each of the three disease management programs. By equivalent, we mean what value on the 10-point scale produces an estimate of mean WTP that equals the mean WTP using definitely sure yeses. We also determine which values on the 10-point scale give the same estimates of WTP as the real purchases.

Allowing all values on the 10-point certainty scale is the same as considering all yeses to be true yeses, i.e., no calibration. Allowing only yes responses for subjects who indicate 8 or greater on the certainty scale, for example, decreases the number of calibrated true yeses by the number of subjects with certainty less than 8 and reduces the estimate of mean WTP accordingly. The estimate of the mean WTP cannot increase as the critical scale value increases and will typically decrease. The estimate of the mean WTP will be lowest for a critical certainty scale value of 10. A figure that shows a plot of mean WTP on the vertical axis and critical certainty scale value that is used for calibration on the horizontal axis will show a downward-sloping curve from left to right. Because the estimates of mean WTP are similar for nonparametric and parametric estimations we will discuss just the parametric estimates.

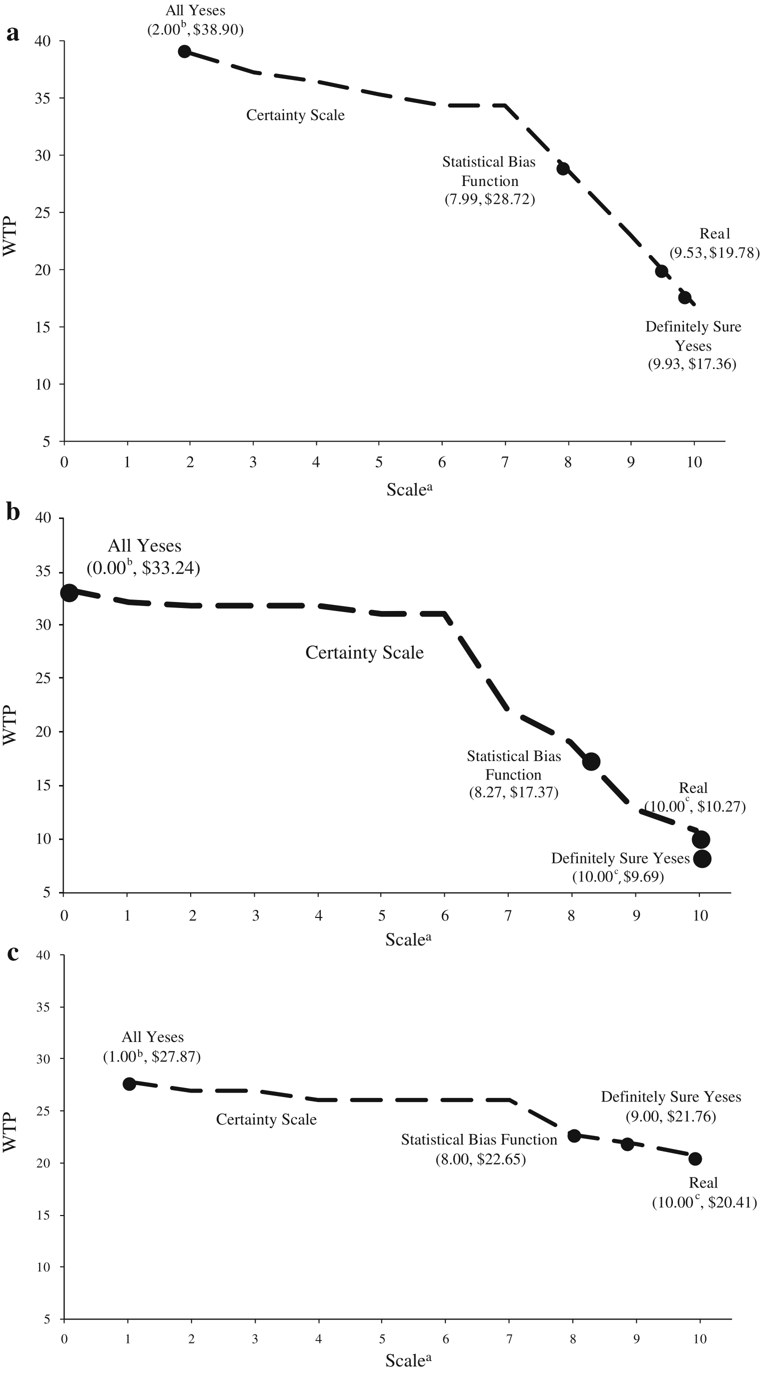
Figure 4a shows the downward-sloping curve for the diabetes management program. When all yeses are counted as true yeses, the estimate of the mean WTP is $38.90. The point at certainty scale value 2.00, the minimum value reported, and $38.90 is for all (uncalibrated) hypothetical yeses. When only yeses with certainty greater than the value required by the statistical bias function are considered true yeses, the estimate of the mean WTP is $28.72. The estimate of mean WTP for definitely sure calibration is $17.36, and the value on the 10-point certainty scale that produces the same estimate of mean WTP is 9.93. The estimate of mean WTP for real purchases is $19.78, and the value on the 10-point scale that produces the same estimate of mean WTP is 9.53.

The downward-sloping line in Fig. 4b shows the relationship between mean WTP and certainty scale value for the asthma management program. When all yeses are counted as true yeses, the estimate of mean WTP is $33.24. When only yeses with certainty value greater than the value required by the statistical bias function are considered true yeses, the estimate of the mean WTP is $17.37. The estimate of mean WTP for definitely sure calibration is $9.69, and the value on the 10-point certainty scale that produces the closest estimate of mean WTP is 10. The estimate of mean WTP for real purchases is $10.27, and the value on the 10-point scale that produces the closest estimate of mean WTP is 10.

The downward-sloping line in Fig. 4c depicts the relationship between mean WTP and certainty scale value for the lipid management program. When all yeses are counted as true yeses, the estimate of mean WTP is $27.87. When only yeses with certainty value greater than the value required by the statistical bias function are considered true yeses, the estimate of the mean WTP is $22.65. The estimate of mean WTP for definitely sure calibration is $21.76, and the value on the 10-point certainty scale that produces the same estimate of mean WTP is 9.00. The estimate of mean WTP for real purchases is $20.41, and the value on the 10-point scale that produces the closest estimate of mean WTP is 10.

9 Discussion and Conclusions

In striving to elicit willingness to pay without bias researchers use a variety of measures to mitigate hypothetical bias including ex ante, in medias res, and ex post efforts. This paper focuses on two ex post measures, follow up certainty statements. One asks respondents to indicate whether they are “definitely sure” of their decision in contingent valuation or “probably sure.” The other asks respondents to indicate their certainty on a 10-point scale.


a
All Yeses
(2.00b, $38.90)
Certainty Scale
Statistical Bias Function
(7.99, $28.72)
Real
(9.53, $19.78)
Definitely Sure Yeses
(9.93, $17.36)
WTP
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
Scalea
b
All Yeses
(0.00b, $33.24)
Certainty Scale
Statistical Bias Function
(8.27, $17.37)
Real
(10.00c, $10.27)
Definitely Sure Yeses
(10.00c, $9.69)
WTP
Scalea
c
All Yeses
(1.00b, $27.87)
Certainty Scale
Statistical Bias Function
(8.00, $22.65)
Definitely Sure Yeses
(9.00, $21.76)
Real
(10.00c, $20.41)
WTP
Scalea

◀ **Fig. 4** **a** Scale value producing WTP equal to definitely sure WTP. Diabetes. Mean WTP declines as the critical certainty scale value increases toward 10^{a}.
Superscript 'a' indicates the mean WTP based on certainty scale values of 7, 8, 9, and 10 are $34.32, $28.64, $22.99, and $16.96, respectively; Superscript 'b' indicates the minimum certainty scale value reported by a subject was 2.
b Scale value producing WTP equal to definitely sure WTP. Asthma. Mean WTP declines as the critical certainty scale value increases toward 10^{a}.
Superscript 'a' indicates the mean WTP based on certainty scale values of 7, 8, 9, and 10 are $22.04, $19.03, $12.86, and $10.75, respectively; Superscript 'b' indicates the minimum certainty scale value reported by a subject was 0; Superscript 'c' indicates the even a certainty scale value of 10 yields a point estimate of willingness to pay greater than this point estimate.
c Scale value producing WTP equal to definitely sure WTP. Lipid. Mean WTP declines as the critical certainty scale value increases toward 10^{a}.
Superscript 'a' indicates The mean WTP based on certainty scale values of 7, 8, 9, and 10 are $26.11, $22.65, $21.76, and $20.59, respectively; Superscript 'b' indicates the The minimum certainty scale value reported by a subject was 1; Superscript 'c' indicates even a certainty scale value of 10 yields a point estimate of willingness to pay greater than $20.41

Both types of follow up certainty statements have produced calibration of hypothetical yeses that match well with real purchases of environmental and health related goods.

We make comparisons using data from three separate field experiments in a private good, dichotomous choice setting. One was for a diabetes management program, a second was for an asthma management program, and a third was for a lipid management program. In each experiment the program was offered hypothetically in contingent valuation and for real purchase. The diabetes and asthma experiments allow for split sample comparisons because individuals were offered the good either hypothetically or for real. We compare four types of hypothetical yes responses. They are yeses calibrated by definitely sure, an updated statistical bias function, 8 or greater on the 10-point certainty scale, and lastly all (uncalibrated) yeses. We compared these hypothetical responses to real purchases also. We compared the means of observable characteristics of the respondents for the four types of yeses. While there are some differences in the means of some of these characteristics, it is not clear if there is any pattern.

We compared the percentage of yes responses in contingent valuation. The percentage of yeses tends to increase as we move from calibration by definitely sure to the statistical bias function and 8 or greater on the certainty scale, to all yeses. Comparisons to the real purchase decisions showed that all three calibrations remove indications of hypothetical bias at the 5% level for diabetes and asthma, the two programs for which bias was indicated. Using definitely sure to identify true yes responses produces a set of yeses that give no indication of hypothetical bias at any of the usual levels of statistical significance. To account for any influential differences in observable characteristics of subjects we compared the performance of a dummy variable for being a hypothetical yes in contingent valuation in logit regressions of all (hypothetical and real, yes and no) responses. We found that overall comparisons of the Hypothetical dummy variable in logit regressions that control for characteristics of the subjects mimic the comparisons of the percentage yes responses. Our interpretation is that any differences in the observable characteristics do not have much effect on differences between hypothetical and real responses for the three disease management programs.

We compared the estimates of mean WTP based on the three calibrations and all yeses and also based on real purchases. For the three experiments, estimates of mean WTP tend to increase from yeses calibrated by definitely sure, to estimates based on either the statistical bias function or 8 or greater on the certainty scale, to all hypothetical (uncalibrated) yeses. For all yeses there is strong evidence of hypothetical bias for the diabetes and asthma management programs. All three of the calibration methods remove evidence of statistically significant hypothetical bias at usual levels of confidence for diabetes and asthma programs except for weak evidence for the statistical bias function for asthma. For all three disease management programs, calibration by definitely sure produces point estimates of mean WTP closest to the mean of real WTP.

Finally, we estimated the values on the 10-point certainty scale that will produce the same estimate of mean WTP as calibration by “definitely sure” and the same estimate of mean WTP as real purchases. For the three disease management programs the certainty scale values that produce the same estimates of mean WTP as calibration by definitely sure are 9.9 for diabetes, 10 for asthma, and 9.0 for lipid. The certainty scale values that produce the same or closest estimates of mean WTP as real purchases are 9.5 for diabetes, 10 for asthma, and 10 for lipid. All of these values are close to 10.

Eliciting willingness to pay without bias means that the ratio of sample estimates of hypothetical values divided by real values equals one. Little and Berrens (2004) conduct a meta-analysis of ratios from 41 studies that, on average, have hypothetical values that are approximately triple the real values. Factors considered relevant to explaining hypothetical bias include laboratory setting or not, private or public good, elicitation method, and hypothetical bias mitigation measures such as cheap talk and certainty corrections. Their analysis (and that of Murphy et al. 2005) find little evidence that hypothetical bias differs between private and public goods. However, they also find that a referendum format for public goods reduces hypothetical bias consistent with Carson and Groves (2007) contention that a consequential, coercive dichotomous choice referendum can be incentive compatible. Their meta-analysis shows that correction for certainty reduces hypothetical bias.

Whether calibration by the simple definitely sure follow up question would perform as well and be equivalent to values near 10 on the certainty scale for environmental goods or other goods that might be provided publicly remains an open question. Recent work by Flachaire and Hollard (2007) based on “coherent arbitrariness” suggests that uncertain respondents will tend to say yes for environmental goods such as preventing another oil spill like the one by Exxon Valdez and that estimates of WTP are likely to be upper bounds unless a second dichotomous choice question is asked. Whether more reliable estimates are obtained by asking a follow up valuation question or a follow up certainty question is another open question. Håkansson (2008) develops a classic and interval open-ended elicitation (CIOE) format that permits respondents who know their value to report a single value and respondents who are less certain to report an interval of values. How the mean and lower and upper bound estimates from this format compare to estimates based on dichotomous choice with certainty follow up statements is worth exploring. For now, we do have some idea how identifying true hypothetical yes responses by definitely sure/probably sure follow up certainty questions compares to calibration using a 10-point certainty scale for private goods when values are elicited using dichotomous choice.

References

Arrow K, Solow R, Leamer E, Portney P, Radner R, Schuman H (1993) Report of the NOAA panel on contingent valuation. Fed Regist 58:4602–4614

Berrens RP, Jenkins-Smith H, Bohara AK, Silva CL (2002) Further investigation of voluntary contribution contingent valuation: fair share, time of contribution, and respondent uncertainty. J Environ Econ Manag 44(1):144–168

Blumenschein K, Johannesson M (2001) Patient willingness to pay for lipid management services provided by pharmacists: an application of the contingent valuation method. University of Kentucky, College of Pharmacy, Unpublished Report

Blumenschein K, Johannesson M, Blomquist GC, Liljas B, O'Conor RM (1998) Experimental results on expressed certainty and hypothetical bias in contingent valuation. South Econ J 65(1):169–177

Blumenschein K, Johannesson M, Yokoyama KK, Freeman PR (2001) Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. J Health Econ 20(3):441–457

Blumenschein K, Blomquist GC, Johannesson M, Horn N, Freeman P (2008) Eliciting willingness to pay without bias: evidence from a field experiment. Econ J 118(525):114–137

Carson RT, Groves T (2007) Incentive and informational properties of preference questions. Environ Resour Econ 37(1):181–210

Champ PA, Bishop RC (2001) Donation payment mechanisms and contingent valuation: an empirical study of hypothetical bias. Environ Resour Econ 19(4):383–402

Champ PA, Bishop RC, Brown TC, McCollum DW (1997) Using donation mechanisms to value nonuse benefits from public goods. J Environ Econ Manag 33(2):151–162

Cummings RG, Taylor LO (1999) Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. Am Econ Rev 89(3):649–665

Dubourg WR, Jones-Lee MW, Loomes G (1994) Imprecise preferences and the WTP-WTA disparity. J Risk Uncertain 9:115–133

Evans MF, Flores NE, Boyle KJ (2003) Multiple-bounded uncertainty choice data as probabilistic intentions. Land Econ 79:549–560

Farmer MC, Lipscomb CA (2008) Conservative dichotomous choice responses in the active policy setting: DC rejections below WTP. Environ Resour Econ 39:223–246

Flachaire E, Hollard G (2007) Starting point bias and respondent uncertainty in dichotomous choice contingent valuation surveys. Resour Energy Econ 29(3):183–194

Håkansson C (2008) A new valuation question: analysis and insights from interval open-ended data in contingent valuation. Environ Resour Econ 39:175–188

Harrison GW (2006) Experimental evidence on alternative environmental valuation methods. Environ Resour Econ 34(1):125–162

Harrison GW, Rutström E (2008) Experimental evidence on the existence of hypothetical bias in value elicitation methods. In Plott C, Smith VL (eds) Handbook of experimental economics results. Elsevier Science, New York

Johannesson M, Johansson P-O, Kriström B, Gerdtham U-G (1993) Willingness to pay for antihypertensive therapy: further results. J Health Econ 12(1):95–108

Johannesson M, Blomquist GC, Blumenschein K, Johansson P-O, Liljas B, O'Conor RM (1999) Calibrating hypothetical willingness to pay responses. J Risk Uncertain 18(1):21–32

Johansson P-O (1995) Evaluating health risks: an economic approach. Cambridge University Press, Cambridge

Kriström B (1990) A non-parametric approach to the estimation of welfare measures in discrete response valuation studies. Land Econ 66(2):135–139

Li C-Z, Mattsson L (1995) Discrete choice under preference uncertainty: an improved structural model for contingent valuation. J Environ Econ Manag 28:256–269

List JA, Gallet CA (2001) What experimental protocol influence disparities between actual and hypothetical stated values. Environ Resour Econ 20(3):241–254

Little J, Berrens R (2004) Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. Econ Bull 3(6):1–13

Loomis J, Ekstrand E (1998) Alternative approaches for incorporating respondent uncertainty when estimating willingness to pay: the case of the Mexican spotted owl. Ecol Econ 27(1):29–41

Loomis J, Brown T, Lucero B, Peterson G (1996) Improving validity of experiments of contingent valuation methods: results of efforts to reduce the disparity of hypothetical and actual willingness to pay. Land Econ 72(4):450–461

Murphy JJ, Allen PG, Stevens TH, Weatherhead D (2005) Meta-analysis of hypothetical bias in stated preference valuation. Environ Resour Econ 30(3):313–325

Poe GL, Clark JE, Rondeau D, Schulze WD (2002) Provision point mechanisms and field validity tests of contingent valuation. Environ Resour Econ 23(1):105–131

Ready RC, Whitehead JC, Blomquist GC (1995) Contingent valuation when respondents are ambivalent. J Environ Econ Manag 29(2):181–196

Ready RC, Navrud S, Dubourg WR (2001) How do respondents with uncertain willingness to pay answer contingent valuation questions. Land Econ 77(3):315–326

Svensson E (2000) Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometr J 42(4):417–434

Vossler CA, Ethier RG, Poe GL, Welsh MP (2003) Payment certainty in discrete choice contingent valuation responses: results from a field validity test. South Econ J 69(4):886–902

Vossler CA, McKee M (2006) Induced-value tests of contingent valuation elicitation mechanisms. Environ Resour Econ 35:137–168

Wang H (1997) Treatment of don't-know responses in contingent valuation surveys: a random valuation model. J Environ Econ Manag 32(2):219–232

Watson V, Ryan M (2007) Exploring preference anomalies in double-bounded contingent valuation. J Health Econ 26(3):463–482

Whitehead JC, Cherry T (2007) Willingness to pay for a green energy program: a comparison of ex-ante and ex-post hypothetical bias mitigation approaches. Resour Energy Econ 29(4):247–326