University of Massachusetts Amherst
Department of Resource Economics
Working Paper No. 2003-8

A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation

James J. Murphy[1], P. Geoffrey Allen[2], Thomas H. Stevens[3], and Darryl Weatherhead[4]

Abstract:

Individuals are widely believed to overstate their economic valuation of a good by a factor of two or three. This paper reports the results of a meta-analysis of hypothetical bias in 28 stated preference valuation studies that report monetary willingness-to-pay and that used the same mechanism for eliciting both hypothetical and actual values. The papers generated 83 observations with a median value of the ratio of hypothetical to actual value of 1.35, and the distribution has severe positive skewness. Since a comprehensive theory of hypothetical bias has not been developed, we use a set of explanatory variables based on issues that have been investigated in previous research. We find that a choice-based elicitation mechanism is important in reducing bias, though an insufficient number of studies and confounding with other variables prevents us from characterizing individual mechanisms. We provide some evidence that the use of student subjects may be a source of bias, but this variable is highly correlated with group experimental settings and no firm conclusions can be drawn. There is some weak evidence that bias increases when public goods are being valued, and that some calibration methods are effective at reducing bias. Results are quite sensitive to model specification, which will remain a problem until a comprehensive theory of hypothetical bias is developed

Keywords: contingent valuation, experiments, hypothetical bias, meta-analysis, stated preference

JEL Classification: C9, Q26, Q28, H41

[1] James J. Murphy, Department of Resource Economics
University of Massachusetts, Stockbridge Hall, 80 Campus Center Way
Amherst, MA 01003
E: murphy@resecon.umass.edu P: 413-545-5716 F: 413-545-5853

[2] P. Geoffrey Allen, Department of Resource Economics
University of Massachusetts, Stockbridge Hall, 80 Campus Center Way
Amherst, MA 01003
E: allen@resecon.umass.edu P: 413-545-5715 F: 413-545-5853

[3] Thomas H. Stevens, Department of Resource Economics
University of Massachusetts, Stockbridge Hall, 80 Campus Center Way
Amherst, MA 01003
E: tstevens@resecon.umass.edu P: 413-545-5714 F: 413-545-5853

[4] Darryl Weatherhead
U.S. Environmental Protection Agency
Office of Inspector General
Research Triangle Park, NC
E: Weatherhead.Darryl@epamail.epa.gov

A META-ANALYSIS OF HYPOTHETICAL BIAS IN STATED PREFERENCE VALUATION

James J. Murphy
Department of Resource Economics, and Center for Public Policy and Administration,
University of Massachusetts, Amherst.

P. Geoffrey Allen
Department of Resource Economics, University of Massachusetts, Amherst.

Thomas H. Stevens
Department of Resource Economics, University of Massachusetts, Amherst.

Darryl Weatherhead
U.S. Environmental Protection Agency, Office of Inspector General, Research Triangle Park, NC

June 2003

Please direct correspondence to:
James J. Murphy
Dept. of Resource Economics
Stockbridge Hall
80 Campus Center Way
University of Massachusetts
Amherst, MA 01003
phone: (413) 545-5716
fax: (413) 545-5853
email: murphy@resecon.umass.edu

Keywords: contingent valuation, experiments, hypothetical bias, meta-analysis, stated preference

JEL Classification: C9, Q26, Q28, H41

Acknowledgments

Funding was provided by the Center for Public Policy and Administration at the University of Massachusetts-Amherst, and by the Cooperative State Research Extension, Education Service, U. S. Department of Agriculture, Massachusetts Agricultural Experiment Station, under Project No. W-133. Ira Athale provided valuable research assistance. We take full responsibility for any errors.

A META-ANALYSIS OF HYPOTHETICAL BIAS IN STATED PREFERENCE VALUATION

Abstract

Individuals are widely believed to overstate their economic valuation of a good by a factor of two or three. This paper reports the results of a meta-analysis of hypothetical bias in 28 stated preference valuation studies that report monetary willingness-to-pay and that used the same mechanism for eliciting both hypothetical and actual values. The papers generated 83 observations with a median value of the ratio of hypothetical to actual value of 1.35, and the distribution has severe positive skewness. Since a comprehensive theory of hypothetical bias has not been developed, we use a set of explanatory variables based on issues that have been investigated in previous research. We find that a choice-based elicitation mechanism is important in reducing bias, though an insufficient number of studies and confounding with other variables prevents us from characterizing individual mechanisms. We provide some evidence that the use of student subjects may be a source of bias, but this variable is highly correlated with group experimental settings and no firm conclusions can be drawn. There is some weak evidence that bias increases when public goods are being valued, and that some calibration methods are effective at reducing bias. Results are quite sensitive to model specification, which will remain a problem until a comprehensive theory of hypothetical bias is developed.

I. Introduction

Recent empirical evidence suggests that stated preference (SP) valuation methods frequently overstate economic value, often by a large amount. Harrison and Rutström (1999), for example, found that 34 of 39 SP observations had an average hypothetical bias of about 338 percent. The well-known NOAA panel recommendations suggest that hypothetical values be divided by two, and List and Gallet (2001) found that on average, subjects responding to hypothetical situations overstated their preferences by a factor of about three. [1]

At this juncture, basic questions about hypothetical bias in SP valuation continue to be debated. First, what is the actual magnitude of hypothetical bias associated with the SP valuation approach? Second, what factors are responsible for this bias? This paper uses a meta-analysis to focus primarily on a reassessment of the magnitude of bias present in SP studies. We also attempt to evaluate the effect of several alternative SP formats and other factors on the degree of hypothetical bias. However, as noted by Carson*, et al.* (1996), due to the lack of theory about the causes of hypothetical bias, missing data, and the need to use a large set of dummy variables, our ability to determine the factors responsible for hypothetical bias is rather limited.

Our results differ from previous work in two important respects. First, we find that hypothetical bias in SP studies may not be as important as most previous studies suggest. Second, we question the prevailing wisdom about several of the factors responsible for this bias.

II. Revisiting the List and Gallet Results [2]

Since much of our analysis was conducted concurrent with the only other published meta-analysis of hypothetical bias in stated values (List and Gallet, 2001, hereafter LG), we begin with

a summary of their study and a sensitivity analysis of their findings. LG assume that actual cash-based estimates are unbiased measures of value and define hypothetical bias as a calibration factor (CF) that is the ratio of the hypothetical to actual expression of value. LG include 29 studies yielding 58 observations or calibration factors. Some studies derived several observations that LG report as a range, rather than as a single value. LG report the results from three different regression models, using the minimum, median, or maximum calibration factor values as the dependent variable.[3] The independent variables represent different experimental design parameters from the studies, including whether the calibration factor was based on an individual's willingness-to-pay or a willingness-to-accept, the type of experiment (lab or field), type of good (private or public), the type of comparison (within or between group), and eight different types of elicitation mechanism.

LG's estimation results using either the natural log of the calibration factor or the absolute value of the natural log of the calibration factor are qualitatively similar. LG mention that using a linear model, rather than semi-log, also yielded essentially the same conclusions. Since their results are not very sensitive to these differences, we focus on the natural log of the median value for ranges of the calibration factor.[4]

List and Gallet argue that hypothetical bias should be greater in WTA studies than in WTP studies, because most respondents are much more familiar with WTP situations. Using similar logic, bias associated with public goods is expected to exceed that of private goods since respondents are assumed to have more experience valuing private goods. And incentive compatible elicitation methods, such as dichotomous choice, are expected to result in less hypothetical bias, all else held constant.

Results of the LG analysis, summarized in the second column of Table I, indicate that the magnitude of hypothetical bias was statistically less for (a) WTP as compared to WTA applications, (b) private as compared to public goods, and (c) one elicitation method, the first price sealed bid, as compared to the Vickery second-price auction baseline. In the next section, we test the robustness of these conclusions.

<INSERT TABLE I>

Procedures and Results

Our sensitivity analysis of the LG results proceeded in two steps. We began by validating LG's coding of their data, and then tested the sensitivity of their results to particular observations and assumptions. We disagreed with LG's coding of several observations included in their analysis and grouped these disagreements into three "types of differences," summarized in Table II. *ERE typo* refers to observations that were reported incorrectly in their paper, but were correct in the actual data used in their regressions. Making these changes to the data reported in the paper, we were able to duplicate their original results as shown in Table I. Next, there were two observations that we could not find in the papers, so we did not include them. There were also 16 observations that appeared to be coding errors. For example, LG recorded the Bohm (1972) study as making comparisons within groups, whereas this study actually compared results between groups for three of the four observations. After making these changes, listed in Table II, we re-estimated the LG model. The results are in Table I under the Revision 1 heading. Although these changes affected the coefficient values, the results are qualitatively similar. This indicates that updating the LG data for typos and errors has a quantitative, but no qualitative, effect on their conclusions. However, it is possible that their conclusions are not driven by the

experiment protocol variables, but rather the results from one or two studies. We elaborate on this point below.

<INSERT TABLE II>

LG's sample size is relatively small with insufficient variation for the model they estimated. Using the revised LG data, there are 29 studies yielding 55 observations. Table III contains a frequency distribution of the LG data for each of their independent variables. Most of the elicitation mechanisms have just one study using that format, and there are only eight WTA observations from six studies. Moreover, two of these WTA observations are from a single study (Brookshire and Coursey (1987)) with calibration factors that are at least 17 times greater than the mean of the other six. Given the paucity of WTA observations, it is possible that the significance of the WTP coefficient is entirely due to this study and has nothing to do with a fundamental difference between responses to WTP and WTA questions. More importantly, Brookshire and Coursey (1987) use different mechanisms to elicit hypothetical and actual values (open-ended and Smith auction, respectively). It is possible that their calibration factors confound hypothetical bias with free-rider bias due to changing from a demand revealing mechanism to one that is not.

<INSERT TABLE III>

We tested the sensitivity of the LG results to the two large WTA calibration factors (28.20 and 25.79) from Brookshire and Coursey by dropping these observations; the Revision 2 results are reported in Table I.[5] Consistent with the original LG results, private goods still produce a lower and statistically significant hypothetical bias than public goods. However, the WTP coefficient is no longer statistically significant. It would be premature to conclude this suggests that there is no difference between WTP and WTA studies. Rather, we interpret this to

mean that there are an insufficient number of observations to say anything about their relative impacts on hypothetical bias.

We also did a similar analysis for the five elicitation mechanisms with just a single study. We ran a series of regressions in which we omitted, one at a time with replacement, the study and independent variable for first price sealed bid, provision point, Smith auction, random price auction and BDM. The LG results were quite robust with respect to these changes. In particular, the significance of the WTP dummy variable was consistently driven by Brookshire and Coursey (1987) and the coefficient on the private good dummy variable was consistently negative and significant. The dummy variable for a within group comparison was never significant.

In the Revision 3 regressions, we made another set of adjustments to the LG data for what we call differences in interpretation. For some observations, we disagreed with LG about how to code the observation. For example, the Bishop and Heberlein (1979) study does not report any actual WTP values. It appears that the LG calibration factor is the ratio of a hypothetical WTP and an actual WTA. Since this could confound hypothetical bias with differences in WTP/WTA, we decided not to include this observation. Also, to avoid confounding hypothetical bias with changes in the elicitation mechanism, we only included studies that used the same mechanism for both the hypothetical and the actual valuation. The interpretation differences are listed in Table IV. These changes leave us with 32 observations from 21 studies. The results of using all the changes for Revision 1, plus the interpretation differences, are reported in Table I, Revision 3. After updating the LG data for coding differences and testing for the sensitivity of the results to particular observations, two key conclusions emerge: (1) the statistically significant difference between WTP and WTA in the original LG results is sensitive to two extreme values that use different elicitation mechanisms

for actual and hypothetical valuation, and (2) private goods continue to have a lower bias than public goods. The negative coefficients for lab experiments and within group comparisons are now weakly significant at the 10 percent level. A few elicitation mechanisms are also significant, but since most of these variables are based on just a single study, we hesitate to interpret this.

<INSERT TABLE IV>

In the next section, we present our meta-analysis using an expanded data set with a different set of criteria for including observations. We estimate a different model than LG and arrive at somewhat different conclusions.

III. Description of Data

We were able to identify 59 studies that reported both hypothetical and actual values (there were an additional four studies that reported *ratios* of hypothetical and actual values, but not the respective values). In order to include an observation from a paper, we used the following criteria:

- The hypothetical and actual values had to be elicited using the same mechanism (for example, this would exclude Brookshire and Coursey (1987), because the hypothetical values were elicited using an open-ended format but the actual values were elicited using a Smith auction). We imposed this requirement to avoid confounding any affects from the different elicitation mechanisms with hypothetical bias. For nine studies, all the observations reported used different elicitation mechanisms so there are no observations from those papers in our sample.

- We only included WTP observations because, although it is possible that there are important differences between WTP and WTA responses, unfortunately there are not enough WTA studies to truly capture any such effects. With only a small number of studies, a dummy variable might simply reflect the influence of a study, rather than that of WTA, on hypothetical bias. This requirement removed five studies from the sample.
- The hypothetical and actual values had to be WTP measured in currency, not, for example, as a percent of people responding "yes" to a dichotomous choice question. All non-US currencies were converted to nominal US dollars. Since our regression models use hypothetical and actual values as variables, this requirement keeps the units consistent. We included dichotomous choice studies if the authors provided an estimate of WTP. However, since many of these studies do not report monetary estimates of WTP, this group of studies may be under-represented in our sample. We were able to locate 13 such studies that provided hypothetical and actual percent "yes" responses, but were excluded because no cash-based WTP estimates were provided.

After imposing these restrictions, our data set includes 28 studies yielding 83 observations (see Appendix A). The variables used in the analysis are defined in Table V, and summary statistics are provided in Tables VI and VII. *LnHypValue* is the natural log of the value elicited in a hypothetical setting, and *lnActValue* is natural log of the amount of cash actually paid by the respondent. We assume that these cash-based estimates are unbiased measures of the true WTP. For each observation, we also calculate the calibration factor, *CF*, which is the ratio of hypothetical value to actual value; *CF* exceeds one in the presence of hypothetical bias. Consistent with LG and Harrison and Rutström (1999), the mean *CF* in Table VI is 2.60. This

comes from a skewed distribution as indicated by a 1.35 median *CF*. Figure 1 presents the distribution of *CF*s.

<INSERT TABLE V>

<INSERT TABLE VI>

<INSERT TABLE VII>

<INSERT FIGURE 1>

The variables *Private* (=1 for private goods, =0 for public goods), and *Within* (=1 for within group comparison, =0 for between group comparison) are defined the same as in LG. We chose not to use the LG variable *Lab* because of challenges with precisely defining a laboratory experiment. Clearly, the typical experiment run on a college campus using the student body in either a classroom or computer lab would be coded as *Lab*. But what about a study such as Cummings*, et al.* (1995) in which members of a church group were asked about their WTP for an electric juicer? Procedurally, these experiments were similar to the "typical" on-campus lab experiment, the differences were in the location (church vs. campus) and the subject pool (students vs. adults). We created two new dummy variables, *Student* and *Group*, that are intended to capture essentially the same effects as LG's *Lab* variable. We coded an observation as *Student* = 1 if the subject pool was college students; *Student* = 0 if the subject pool was adults or adult students. *Group* = 1 if values were elicited in a group setting such as a classroom, computer lab or church hall; *Group* = 0 if values were elicited in an individual setting such as a phone or mail survey. We should note that the *Group* variable refers to the setting, not the nature of the decision. If an individual completed a survey in the classroom, then *Group* = 1, and if there was group interaction, e.g. through a Vickrey auction, but values were elicited individually (such as the baseball card auctions in List (2003)) then *Group* = 0. There is a high degree of

correlation between the *Student* and *Group* variables (Pearson correlation coefficient equals 0.77), therefore we do not use both variables in the same model.

LG included dummy variables for each of the elicitation mechanisms in their sample. However, there is not much variability in the elicitation mechanisms used. In our data, the Vickrey auction accounts for 19% of the observations, dichotomous choice 25% and open-ended 35%. As shown in Table VII, the other elicitation mechanisms are typically represented by one or two papers and provide between one and four observations. Moreover, some elicitation mechanisms are typically associated with a particular type of good, e.g., a Smith auction or a referendum is normally associated with a public good, and a Vickrey or first-price sealed bid auction is usually for private goods. This correlation makes it difficult to isolate the effects of the elicitation mechanism from the type of good. Because of this, we refrain from using dummy variables for each mechanism. Instead, we create a new dummy variable that aggregates the elicitation mechanisms into two groups. The dummy variable *Choice* equals one for studies that use a choice-based elicitation (dichotomous choice, polychotomous choice, payment card, referendum), and *Choice* equals zero for the remaining elicitation mechanisms.

Some studies report descriptive statistics such as mean WTP (e.g., Bohm (1972)). However, there has been a recent growth in the number of studies that attempt to calibrate responses either by getting unbiased responses from individuals *ex ante* (also referred to as instrument calibration, e.g. cheap talk) or by adjusting for the biased responses *ex post* (statistical calibration, e.g. uncertainty adjustments). The variable *Calibrate* equals one if the observation is based on any type of calibration technique.

IV. Estimation Procedures and Results

There is no theory explaining hypothetical bias that could provide guidance as to the appropriate model specification. Therefore, we limit our choice of variables to research protocol and study characteristics for which data were readily available. We begin with a simple double log regression model (Model 1a) that explains actual value as a function of the hypothetical value:

$$lnActValue = \beta_0 + \beta_1 \cdot lnHypValue + \beta_2 \cdot \left(lnHypValue\right)^2 + \varepsilon\,, \tag{1}$$

where *lnActValue* and *lnHypValue* denote the natural log of the actual and hypothetical values.[6] Because White's test indicates the presence of heteroskadasticity (p-value 0.0002), Table VIII reports the results from a weighted regression, using the square root of *lnHypValue* to transform the data.[7] This simple specification fits the data quite well, with an adjusted R^2 of 0.83. All the coefficients are positive and significant at the 10% level. The results indicate that for the range of hypothetical values in our sample, the bias increases as the hypothetical value increases. When evaluated at the mean hypothetical value (26.55), the predicted actual value is 10.24 which yields a calibration factor of 2.59. When the model is evaluated at the median hypothetical value (7.18), we get a predicted actual value of 3.89 and a 1.84 calibration factor. Interestingly, these estimates are roughly consistent with NOAA's calibration factor of two.

<INSERT TABLE VIII>

To determine whether there are some factors that may help explain the cause of this bias, we estimated the following model (Model 2a):

$$\begin{aligned} lnActValue = {} & \beta_0 + \beta_1 \cdot lnHypValue + \beta_2 \cdot \left(lnHypValue\right)^2 + \beta_3 \cdot Student + \beta_4 \cdot Private \\ & + \beta_5 \cdot Within + \beta_6 \cdot Choice + \beta_7 \cdot Calibrate + \varepsilon. \end{aligned} \tag{2}$$

The results for Model 2a are in Table VIII. When all independent variables are evaluated at their means, the resulting predicted actual value is 8.83 and the CF is 3.01. Evaluating the model at

the median of the independent variables yields a CF of 2.47. Variables with positive coefficients are associated with larger actual values and, therefore, lower hypothetical bias; negative coefficients have the opposite interpretation. The intercept and the coefficient on the quadratic term for *lnHypValue* continue to be positive and significant. The coefficient for *Within* is also positive and significant; this would be consistent with the possibility that in a within-group study, participants might try to maintain some consistency between their hypothetical and actual values. *Private* was significant in LG's results, but not in our Model 2a. Calibration techniques appear to be effective at reducing hypothetical bias.

The positive and significant coefficient for *Choice* indicates that the choice elicitation mechanisms (dichotomous and polychotomous choice, referendum, payment card and conjoint) are associated with less hypothetical bias. There may be several reasons for this finding. First, substitutes are made explicit in the choice format and this may encourage respondents to explore their preferences and tradeoffs in more detail. Neoclassical theory indicates that if few substitutes are considered, respondents will likely express a higher WTP than if many are considered, all else equal. From a psychological perspective, the process of making choices is quite different from that of pricing, as in open ended CV (Brown (1984; Irwin*, et al.* (1993; McKenzie (1993)). Another factor is that some choice formats, like conjoint, allow respondents to express ambivalence, indifference or uncertainty directly. Since a high level of uncertainty is often associated with significant hypothetical bias, choice formats may produce less bias (Champ*, et al.* (1997)).

The negative coefficient on *Student* suggests that there may also be a subject pool effect. However, since all the studies in our sample that use students are laboratory experiments, it is unclear whether the cause of hypothetical bias is the subject pool or the setting. We replaced the

Student variable in equation 2 with a *Group* dummy variable that equals one if values were elicited in a group setting such as a lab experiment, rather than an individual setting such as a phone or mail survey. The correlation between *Student* and *Group* variables is 0.77. The results of this regression are in Table VIII, Model 2b. The coefficient for *Group* is negative and significant, therefore, although there is clearly an effect, we cannot distinguish whether the cause is the subject pool or the setting.

In Model 2b, *Calibrate* is no longer significant, and *Private* is now significant at the 5% level possibly suggesting some sensitivity to model specification. In the absence of a theory that explains the relationship between hypothetical and actual values, we hesitate to place much emphasis on the significance of particular dummy variables. Moreover, there may simply not be sufficient variability in the data to capture some of these effects. For example, all but one of the observations for which *Calibrate* equals one use a between-group comparison. Instead, we note that most of the variation is explained by the simple Model 1a and make the primary conclusion that hypothetical bias increases with larger hypothetical values. For smaller hypothetical values that are common in CV studies, our results suggest that hypothetical bias may not be a major problem. For example, the predicted CF from a $10 hypothetical value is essentially one, a $21.50 hypothetical value produces a 1.50 CF, and a CF of 2 results from a $32.50 hypothetical value. The *Group*/*Student* and the *Choice* dummy variables are consistently significant and are therefore likely to have some impact on hypothetical bias. We also tested the sensitivity of our results to extreme values by dropping the five largest CFs and re-estimating equation 2. The results of this trimmed model (Model 3), provided in Table VIII, are generally consistent with those of Model 2.

There are a few studies that provide a relatively large number of observations. To control for the possibility that our results could be unduly influenced by such studies, we calculated the mean hypothetical and actual values from each study for a given set of independent variables. With this approach, it is still possible for a study to provide more than one observation. In the case of Sinden (1988), for example, 17 observations were reduced to two: the mean of the 16 observations that use students, and the single observation that uses adults. The resulting data set has 45 observations. The mean CF is 3.26 and the median is 1.50. Table IX summarizes the regression results. Consistent with the results in Table VIII (which uses the full data set), the hypothetical value seems to be the best predictor of actual value (for every regression in Tables VIII and IX, an F-test of the null hypothesis that $\beta_1 = \beta_2 = 0$ in equations 1 and 2 is rejected at the 1% level of significance). In Models 3c and 3d, none of the coefficients are individually significant and an F-test for the joint significance of all the dummy variables is also rejected. However, β_1 and β_2 are jointly significant, and in a separate linear model that omits the quadratic term, β_1 becomes highly significant reinforcing the conclusion about the significance of hypothetical values. In the Expanded Models (2c and 2d), both *Group*/*Student* and *Choice* are again significant, but the significance of other dummy variables appears sensitive to model choice.

<INSERT TABLE IX>

Because conclusions about the significance of most of the dummy variables is rather sensitive, another way to gauge whether a variable has an effect on hypothetical bias is to ask whether the CF changes as the variable changes within a particular study. Some studies report multiple observations because they are testing the effects of a particular variable. For example, nine of the ten studies that use a calibration technique report observations for which *Calibrate*=1

and *Calibrate*=0.[8] The authors then compare the hypothetical bias with and without calibration to test its effectiveness. In each of these nine studies, the mean CF using a calibration technique is less than the mean CF for the uncalibrated observations, suggesting that calibration techniques are effective at reducing hypothetical bias. When the observations from these nine studies are combined, the mean CF for the 15 observations that do not use a calibration is 5.42 with a standard deviation of 6.32, and the median is 2.66. There were another 15 observations that used a calibration technique; the mean was 1.59, standard deviation 1.02 and median 1.18. As one might expect, the mean and median CF are lower for those observations that use a calibration technique. A Wilcoxon rank sum test confirms that this difference is highly significant at the 1% level.

V. Conclusions

This paper presents a meta-analysis of hypothetical bias in WTP contingent valuation studies. We find that the primary factor that explains this bias is the magnitude of the hypothetical value. Attempts to identify other factors that may be associated with hypothetical bias yielded mixed results. In all the models we estimated, the coefficients for the *Group/Student* and *Choice* dummy variables were consistently significant and of large magnitude. We, therefore, cautiously note that these factors may be associated with hypothetical bias. In addition, a comparison of calibration factors within particular studies indicates that calibration techniques are effective at reducing hypothetical bias.

We are reluctant to over-emphasize the significance of the dummy variables because a meta-analysis of hypothetical bias appears to be very sensitive to model specification, a lack of variability in the data, and treatment of extreme values. In addition, some of our findings differ

from those in LG. For example, a consistent result in LG was that private goods had a lower and statistically significant CF than public goods, but our results on this conclusion are mixed, depending upon model specification. One variable that we found to consistently be statistically significant (*Student/Group*) was not significant in LG (their *Lab* variable).

We believe that this is a consequence of several factors. First, half of the calibration factors are between 0.85 and 1.50, and 70% of the calibration factors are below 2. However, as shown in Figure 1, the sample has severe positive skewness (value equals 2.44). The mean CF for the top 10 observations is 10.3, compared with 1.54 for the other 73 observations. This means that econometric estimates of hypothetical bias can often be driven by a few observations. Second, the need to use large sets of dummy variables and the multicollinearity associated with them can make it difficult to isolate the impact of factors that might be responsible for hypothetical bias. For example, in the LG data, provision point mechanisms and Smith auctions are only associated with public goods and Vickrey auctions only with private goods. In our data, only one of the studies that uses a calibration technique also uses a between group comparison. And, since a comprehensive theory of hypothetical bias has not been developed, model specification is generally based on intuition. As a result, the sensitivity of hypothetical bias meta-analyses should not be surprising. This means that our ability to determine the factors responsible for this bias is quite limited, and that estimates of statistical significance associated with several potentially important determinants of bias should be viewed with caution. However, the evidence is quite strong that there is a positive quadratic relationship between hypothetical values and hypothetical bias, and the results of our Model 1 may provide some insights into the potential magnitude of this bias. As shown in Figure 1, 70% of the studies examined here report CFs less than two, and only 30% report CFs that exceed 2. Consequently,

hypothetical bias may not be as significant a problem in stated preference analyses as is often thought.

Table I. Original and Revised LG Results

Variable	Estimated coefficients (standard errors)			
	Original	**Revision 1**[a]	**Revision 2**[b]	**Revision 3**[c]
Constant	1.98 (0.49) ***	2.27 (0.50) ***	1.66 (0.45) ***	2.21 (0.68) ***
Laboratory (X1)	-0.32 (0.23)	-0.17 (0.23)	-0.31 (0.20)	-0.47 (0.28)
WTP (X2)	-0.65 (0.33) *	-0.61 (0.33) *	0.10 (0.33)	0.38 (0.47)
Private good (X3)	-0.64 (0.30) **	-0.85 (0.32) **	-0.74 (0.28) **	-1.04 (0.36) ***
Within group (X4)	-0.01 (0.22)	-0.11 (0.23)	-0.20 (0.20)	-0.49 (0.28) *
Type of elicitation:				
Open-ended (X5)	0.15 (0.28)	-0.32 (0.28)	-0.39 (0.24)	-1.17 (0.42) **
First price sealed bid (X6)	-1.70 (0.75) **	-1.78 (0.75) **	-1.28 (0.65) *	-1.52 (0.78) *
Provision point (X7)	0.54 (0.61)	0.05 (0.79)	0.09 (0.67)	-0.58 (0.83)
Smith auction (X8)	0.32 (0.53)	0.01 (0.54)	-1.11 (0.53) **	—[d]
Random price auction (X9)	-0.76 (0.63)	-1.13 (0.77)	-0.41 (0.68)	-0.20 (0.85)
BDM (X10)	-0.34 (0.47)	-0.24 (0.55)	-0.44 (0.47)	-0.97 (0.55) *
Dichotomous choice (X11)	-0.30 (0.25)	-0.43 (0.26) *	-0.40 (0.22) *	-0.67 (0.33) *
Sample size	58	54	52	32
Adjusted R^2	0.33	0.32	0.17	0.30
F	3.55	3.30	1.97	2.36
p-value	0.001	0.003	0.058	0.047

Dependent variable is the natural log of the median calibration factor in List and Gallet (2001).
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level.

[a] Corrects LG data for errors listed in Table II.

[b] Corrects LG data for errors listed in Table II (Revision 1) and drops two WTA observations with a calibration factor greater than 20 from Brookshire and Coursey (1987).

[c] Corrects LG data for errors listed in Table II (Revision 1) and interpretation differences listed in Table IV.

[d] Variable dropped because no observations with a Smith auction.

Table II. Typos and Coding Errors in the LG Data [a]

LG Study	**LG CF**	**Type of difference**	**Variable**	**LG Coding**	**Our Coding**	*Comments*
Bishop and Heberlein (1986)	1.30-2.30; 0.80	ERE typo	Study	B&H 1986	H&B 1986	Values are from Heberlein and Bishop, 1986, not Bishop 1986.
Kealy, *et al.* (1988)	1.00 - 2.00	ERE typo	Study	1988	1990	Typo in ERE paper
Irwin, *et al.* (1992)	1.00; 2.50	ERE typo	all			Study is in LG regression data, but missing from ERE paper
Kealy, *et al.* (1988)	1.40	ERE typo	all			Study is in LG regression data, but missing from ERE paper
Kealy, *et al.* (1988)	1.30	ERE typo	all			Study is in LG regression data, but missing from ERE paper. Observation from Kealy, *et al.* (1990).
Loomis, *et al.* (1996)	2.00 - 3.60	ERE typo	elicitation	dc	open-ended	Correct in LG data, but typo in Table V in ERE paper
Boyce, *et al.* (1992)	0.90	could not find			not included	Could not find this observation in the paper.
Kealy, *et al.* (1990)	1.30	could not find			not included	Could not find this observation in the paper.
Balistreri, *et al.* (2001)	0.58	error	CF	0.58	1.58	Typo in LG data and ERE paper
Bohm (1972)	1.16; 1.16; 1.34	error	comparison	within	between	

[a] These are the changes made for the Revision 1 regression in Table I.

Table II (cont.). Typos and Coding Errors in the LG Data [a]

LG Study	**LG CF**	**Type of difference**	**Variable**	**LG Coding**	**Our Coding**	*Comments*
Dickie, *et al.* (1987)	1.00	error	elicitation	dichot. choice	open-ended	Actually a posted offer. Experimenter names a price, subjects chooses any quantity. It is not dichotomous choice because subject can choose any quantity.
Fox, *et al.* (1998)	1.20	error	comparison	between	within	LG CF appears to be ratio of survey/final bid for irradiated pork. This is a within group comparison.
Heberlein and Bishop (1986)	1.30 - 2.30	error	elicitation	open-ended	CF=2.26, open-ended. CF=1.33, first-price.	LG present as a range, we split into two observations. The 1.30 CF uses a 1st price sealed bid (error in LG), and the 2.30 CF is open-ended (OK in LG).
Heberlein and Bishop (1986)	0.80	error	CF	0.80	1.13	Appears that LG CF 0.80 is inverse CF (actual/hypothetical) using the 1.24 CF reported in H&B 1986. (0.80=1/1.24). Our CF=1.13 is from Bishop and Heberlein (1990).
Kealy, *et al.* (1988)	1.40	error	good	public	private	

[a] These are the changes made for the Revision 1 regression in Table I.

Table II (cont.). Typos and Coding Errors in the LG Data [a]

LG Study	**LG CF**	**Type of difference**	**Variable**	**LG Coding**	**Our Coding**	*Comments*
List and Shogren (1998)	1.42	error	CF	1.42	not included	Two issues 1. LG used CF=actual/hyp 2. This result is repeated in List Shogren 2002, so we deleted to avoid double counting.
List and Shogren (2002) [b]	0.70 - 1.66	error	CF	0.70 - 1.66	0.60 - 1.41	LG used CF=actual/hyp. Should be hyp/actual.
McClelland, *et al.* (1993)	2.20	error	comparison	between	within	
McClelland, *et al.* (1993)	0.80	error	comparison	between	within	
Navrud (1992)	3.20	error	good	private	public	
Neill, *et al.* (1994)	3.10-25.10	error	elicitation	open-ended	Vickrey	For these two CFs, both hyp and actual use Vickrey
Spencer, *et al.* (1998)	4.66	error	CF	4.66	not included	There is only one CF for both Pond A and Pond B (4.67). LG appear to double count.

[a] These are the changes made for the Revision 1 regression in Table I.

[b] This is listed as List and Shogren (1999) in LG because at the time the paper was forthcoming.

Table III. Frequency Distribution for LG Data after Correcting Typos and Errors[a]

Variable	IX. Value	Number of observations	Number of Studies [b]
Type of Experiment	Laboratory	33	17
	Field or field/lab	22	12
WTP / WTA	WTP	47	25
	WTA	8	6
Type of Good	Private	42	22
	Public	13	7
Type of comparison	Within	18	12
	Between	37	21
Type of elicitation	Open-ended	12	8
	First price sealed bid	1	1
	Provision point	2	1
	Smith auction	4	1
	Random price auction	1	1
	BDM	2	1
	Dichotomous choice	20	14
	Vickrey	13	7
	TOTALS	55	29

[a] Corrections for typos and errors are listed in Table II.

[b] For each variable, the sum could exceed the total number of studies because some studies generate multiple types of observations. For example, Brookshire and Coursey (1987) have two WTP observations and two WTA observations, so this study is counted as providing both a WTP and a WTA observation.

Table IV. Differences in Interpretation about How to Code Data [a]

LG Study	**LG CF**	**Type of difference**	**Variable**	**LG Coding**	**Our Coding**	**Comments**
Balistreri, *et al.* (2001)	1.25	interpret	elicitation	open-ended	not included	Hypothetical and actual elicitation mechanisms differ.
Balistreri, *et al.* (2001)	1.54, 0.58	interpret	elicitation	dichot. choice	not included	Hypothetical and actual elicitation mechanisms differ.
Bishop and Heberlein (1979)	0.30 – 1.60	interpret	WTP/WTA		not included	Study does not have actual WTP. LG appear to use hyp WTP / actual WTA.
Bohm (1972)	1.00, 1.16	interpret	elicitation	open-ended	not included	Hypothetical and actual elicitation mechanisms differ.
Bohm (1972)	1.34	interpret	elicitation	open-ended	not included	The hypothetical elicitation uses a provision point, but the actual does not. Hypothetical and actual elicitation mechanisms differ.
Fox, *et al.* (1998)	1.20, 1.50	interpret	elicitation	open-ended	not included	Hypothetical and actual elicitation mechanisms differ.
Irwin, *et al.* (1992)	1.00, 2.50	interpret	CF		not included	Cannot get CFs. Would have to infer from the charts.
Navrud (1992)	3.20, 1.60 – 2.10	interpret	elicitation	dichot. choice	not included	Hypothetical elicitation was a newspaper ad that did not mention contributions.
Brookshire and Coursey (1987)	2.00, 1.85, 28.20, 25.79	interpret	elicitation	Smith	not included	Hypothetical and actual elicitation mechanisms differ.
Coursey, *et al.* (1987)	1.00, 2.00	interpret	elicitation	Vickrey	not included	Hypothetical and actual elicitation mechanisms differ.

[a] These are the changes made for the Revision 3 regression in Table I.

Table V. Variable Definitions

Variable Name	Description
Study	Name of study
LGStudy	1 = Study included in List and Gallet analysis, 0 = otherwise.
lnHypValue	Natural log of the hypothetical value
lnActValue	Natural log of the actual value (assuming real cash payments represent unbiased estimates of actual value)
CF	Calibration factor (= Hypothetical Value / Actual Value)
Student	1 = subjects were college students, 0 = non-students
Group	1 = values elicited in group setting (e.g., lab) 0 = values elicited in individual setting (e.g., phone or mail survey)
Private	1 = Private good, 0 = Public good
Within	1 = Within group comparison, 0 = Between group comparison
Choice	1 = Choice-based elicitation mechanism • Dichotomous and polychotomous choice, referendum, payment card 0 = Market-based elicitation mechanism • First-price sealed bid, open-ended, BDM, random *n*-th price, Vickrey
Calibrate	1 = ex ante or ex post calibration applied 0 = no calibration applied

Table VI. Descriptive Statistics

	Hypothetical Value	*Actual Value*	*CF*
Mean	26.55	11.69	2.60
Median	7.18	3.67	1.35
Standard deviation	47.33	18.05	3.52
Minimum	0.08	0.07	0.76
Maximum	301	95.5	25.08
N	83	83	83

Table VII. Dummy Variable Descriptive Statistics

Variable	**Value**	**Number of observations**	**Number of Studies** [a]
Subject Pool	College students	35	11
	Non-college students	48	21
Type of Setting	Group	46	15
	Individual	37	13
Type of Good	Private	41	14
	Public	42	14
Type of comparison	Within	28	8
	Between	55	24
Non-choice-based elicitation	Open-ended	29	8
	Vickrey	16	6
	First price sealed bid	1	1
	Random price auction	4	1
	BDM	2	1
Choice-based elicitation	Dichotomous choice	23	12
	Referendum	1	1
	Payment card	4	1
	Polychot. choice	3	2
Calibrate	Ex ante or ex post calibration	17	10
	No calibration used	66	27
	TOTALS	83	28

[a] For each variable, the sum could exceed the total number of studies because some studies generate multiple types of observations.

Table VIII. Regression Results Using All Observations [a]

	Base model		Expanded model				Trimmed model [b]			
	Model 1a		Model 2a		Model 2b		Model 3a		Model 3b	
Variable	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error
Intercept	0.199 ***	0.035	0.357 **	0.163	0.528 ***	0.189	0.230	0.146	0.322 *	0.169
lnHypValue	0.498 ***	0.096	0.171	0.139	0.152	0.139	0.284 **	0.129	0.273 **	0.129
$lnHypValue^2$	0.046 *	0.026	0.096 ***	0.029	0.091 ***	0.028	0.092 ***	0.027	0.089 ***	0.027
Student			-0.470 ***	0.14			-0.244 *	0.130		
Group					-0.539 ***	0.151			-0.292 **	0.142
Private			0.105	0.124	0.293 **	0.118	0.122	0.111	0.227 **	0.107
Within			0.326 **	0.144	0.233 *	0.134	0.222 *	0.129	0.183	0.121
Choice			0.508 ***	0.154	0.465 ***	0.149	0.365 **	0.139	0.351 **	0.135
Calibrate			0.296 **	0.135	0.122	0.137	0.217 *	0.117	0.126	0.119
n	77		77		77		72		72	
Adj R^2	0.83		0.86		0.87		0.90		0.91	
F	188.72		70.50		71.99		97.28		98.37	
p-value	<.0001		<.0001		<.0001		<.0001		<.0001	

[a] Weighted least squares estimates. Dependent variable is the natural log of the actual value (*lnActValue*).
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level.

[b] Trimmed regression – dropped highest five calibration factors.

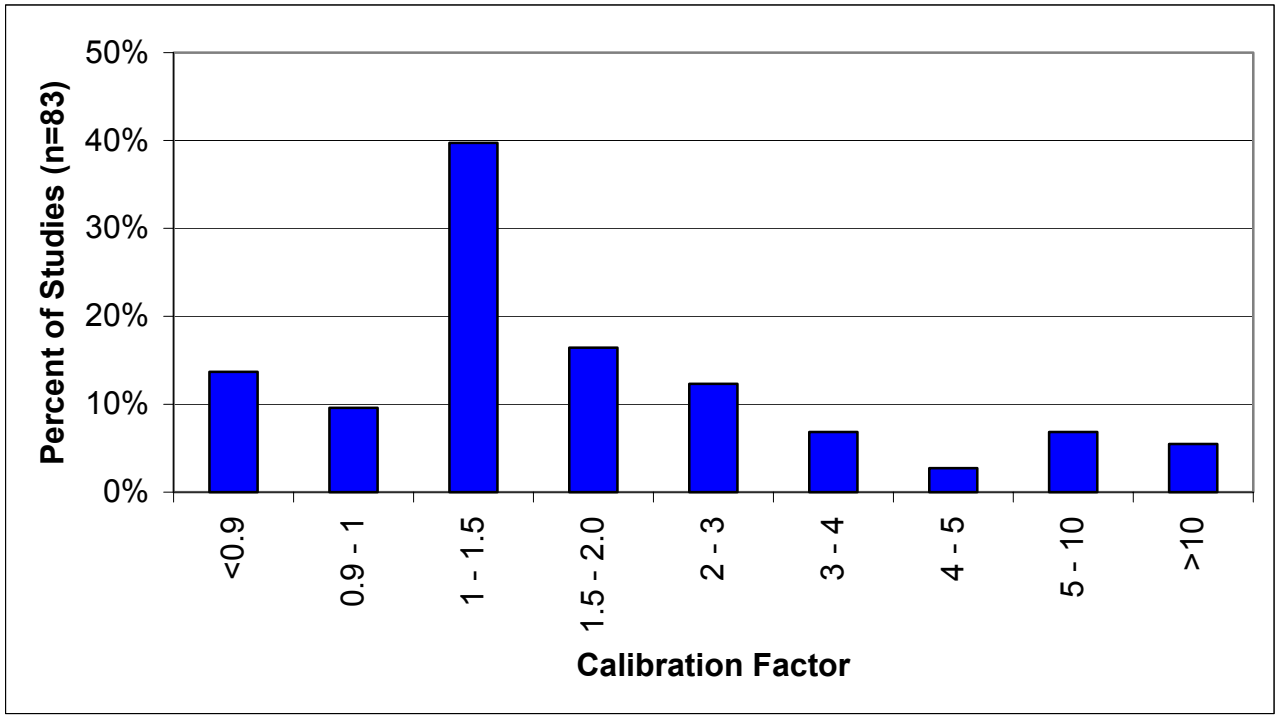
Table IX. Regression Results Using Average Values Per Study [a]

	Base model		Expanded model				Trimmed model [b]			
	Model 1b		Model 2c		Model 2d		Model 3c		Model 3d	
Variable	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error	Coefficient	Std error
Intercept	0.215	0.204	0.408	0.285	0.752 **	0.338	0.188	0.241	0.315	0.295
lnHypValue	0.507 **	0.238	0.112	0.276	-0.010	0.282	0.359	0.261	0.305	0.274
$lnHypValue^2$	0.035	0.052	0.100 *	0.054	0.111 **	0.054	0.081	0.052	0.086	0.053
Student			-0.506 ***	0.177			-0.200	0.151		
Group					-0.662 ***	0.216			-0.258	0.195
Private			0.243	0.172	0.409 **	0.175	0.177	0.142	0.255	0.153
Within			0.189	0.207	0.225	0.205	0.183	0.166	0.204	0.168
Choice			0.471 **	0.212	0.418 **	0.204	0.221	0.182	0.207	0.178
Calibrate			0.286	0.194	0.118	0.192	0.175	0.153	0.114	0.152
n	42		42		42		37		37	
Adj R^2	0.68		0.74		0.75		0.86		0.86	
F	45.08		18.02		18.72		31.45		31.46	
p-value	<.0001		<.0001		<.0001		<.0001		<.0001	

[a] Weighted least squares estimates. Dependent variable is the natural log of the actual value (*lnActValue*).
*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level.

[b] Trimmed regression – dropped highest five calibration factors.

Figure 1. Distribution of Calibration Factors



Appendix A. Data Used in the Analysis

Study	LGStudy	HypValue	ActValue	CF	Student	Group	Private	Within	Elicitation	Choice	Calibrate
Blumenschein, *et al.* (1997)	0	11.97	3.24	3.69	1	1	1	0	Vickrey	0	0
Blumenschein, *et al.* (1997)	0	11.97	1.02	11.74	1	1	1	1	Vickrey	0	0
Blumenschein, *et al.* (2001)	0	10.59	8.97	1.18	0	0	1	0	dichot. choice	1	1
Blumenschein, *et al.* (2001)	0	16.20	8.97	1.81	0	0	1	0	dichot. choice	1	1
Blumenschein, *et al.* (2001)	0	29.23	8.97	3.26	0	0	1	0	dichot. choice	1	0
Bohm (1972)	1	1.76	1.77	0.99	0	1	1	0	open-ended	0	0
Bohm (1972)	1	2.04	2.07	0.99	0	1	1	1	open-ended	0	0
Bohm (1972)	1	1.76	1.52	1.15	0	1	1	0	open-ended	0	0
Bohm (1972)	1	1.76	1.46	1.20	0	1	1	0	open-ended	0	0
Botelho and Costa Pinto (2002)	0	1.39	1.40	0.99	1	1	0	0	open-ended	0	1
Botelho and Costa Pinto (2002)	0	16.14	1.40	11.51	1	1	0	0	open-ended	0	0
Boyce, *et al.* (1989)	1	6.06	4.81	1.26	0	1	1	1	BDM	0	0
Boyce, *et al.* (1989)	1	16.80	7.81	2.15	0	1	1	1	BDM	0	0
Brown, *et al.* (1996)	1	18.98	4.62	4.11	0	0	0	0	open-ended	0	0
Brown, *et al.* (1996)	1	46.55	7.22	6.45	0	0	0	0	dichot. choice	1	0
Brown and Taylor (2000)	0	27.97	3.23	8.66	0	1	0	0	open-ended	0	0
Brown and Taylor (2000)	0	72.22	6.14	11.76	0	1	0	0	open-ended	0	0
Cameron, *et al.* (2002)	0	6.13	4.08	1.50	0	0	0	0	dichot. choice	1	0
Carlsson and Martinsson (2001)	0	0.08	0.07	1.13	1	1	0	1	choice	1	0
Champ, *et al.* (1997)	0	12.00	9.00	1.33	0	0	0	0	dichot. choice	1	1
Champ, *et al.* (1997)	0	52.00	9.00	5.78	0	0	0	0	dichot. choice	1	0
Champ, *et al.* (1997)	0	79.00	9.00	8.78	0	0	0	0	dichot. choice	1	0
Champ and C. (2001)	0	101.00	59.00	1.71	0	0	0	0	dichot. choice	1	0
Duffield and Patterson (1992)	0	14.92	17.69	0.84	0	0	0	0	payment card	1	0
Duffield and Patterson (1992)	0	15.26	17.69	0.86	0	0	0	0	payment card	1	0
Duffield and Patterson (1992)	0	31.18	28.43	1.10	0	0	0	0	payment card	1	0
Duffield and Patterson (1992)	0	31.85	28.43	1.12	0	0	0	0	payment card	1	0
Frykblom (1997)	1	17.69	11.79	1.50	1	1	1	0	dichot. choice	1	0
Frykblom (2000)	1	11.80	8.88	1.33	1	1	1	0	Vickrey	0	0

Appendix A (cont.). Data Used in the Analysis

Study	LGStudy	HypValue	ActValue	CF	Student	Group	Private	Within	Elicitation	Choice	Calibrate
Heberlein and Bishop (1986)	1	35.00	31.00	1.13	0	0	1	0	dichot. choice	1	0
Heberlein and Bishop (1986)	1	32.00	24.00	1.33	0	0	1	0	1st price sealed-bid	0	0
Heberlein and Bishop (1986)	1	43.00	19.00	2.26	0	0	1	0	open-ended	0	0
Johannesson, *et al.* (1998)	1	3.39	4.22	0.80	1	1	1	0	dichot. choice	1	1
Johannesson, *et al.* (1998)	1	3.39	3.85	0.88	1	1	1	1	dichot. choice	1	1
Johannesson, *et al.* (1998)	1	4.97	4.22	1.18	1	1	1	0	dichot. choice	1	0
Johannesson, *et al.* (1998)	1	4.97	3.85	1.29	1	1	1	1	dichot. choice	1	0
Kealy, *et al.* (1988)	1	0.79	0.56	1.41	1	1	1	0	dichot. choice	1	0
Kealy, *et al.* (1988)	1	0.81	0.56	1.45	1	1	1	1	dichot. choice	1	0
List (2001)	0	26.15	25.60	1.02	0	0	1	0	Vickrey	0	1
List (2001)	0	107.89	59.56	1.81	0	0	1	0	Vickrey	0	1
List (2001)	0	49.03	25.60	1.92	0	0	1	0	Vickrey	0	0
List (2001)	0	116.09	59.56	1.95	0	0	1	0	Vickrey	0	0
List (2003)	0	2.58	2.78	0.93	0	0	1	0	Vickrey	0	1
List (2003)	0	3.54	3.42	1.04	0	0	1	0	random *n*-th price	0	1
List (2003)	0	4.97	3.67	1.35	0	0	1	0	random *n*-th price	0	1
List (2003)	0	7.18	3.67	1.96	0	0	1	0	random *n*-th price	0	0
List (2003)	0	5.05	2.28	2.21	0	0	1	0	Vickrey	0	1
List (2003)	0	8.65	3.42	2.53	0	0	1	0	random *n*-th price	0	0
List (2003)	0	7.40	2.78	2.66	0	0	1	0	Vickrey	0	0
List (2003)	0	6.67	2.28	2.93	0	0	1	0	Vickrey	0	0
List and Shogren (1998)	1	208.80	95.50	2.19	0	0	1	1	Vickrey	0	0
List and Shogren (1998)	1	142.02	55.87	2.54	0	0	1	1	Vickrey	0	0
List and Shogren (1998)	1	91.71	26.40	3.47	0	0	1	1	Vickrey	0	0
Loomis, *et al.* (1997)	1	28.00	11.00	2.55	0	1	1	0	dichot. choice	1	0
Loomis, *et al.* (1997)	1	33.00	11.00	3.00	0	1	1	0	dichot. choice	1	0
MacMillan, *et al.* (1999)	0	2.18	2.37	0.92	0	0	0	0	open-ended	0	1
MacMillan, *et al.* (1999)	0	3.97	2.37	1.67	0	0	0	0	open-ended	0	0
Murphy, *et al.* (2002)	0	4.32	0.96	4.50	1	1	0	0	dichot. choice	1	1
Murphy, *et al.* (2002)	0	8.51	0.96	8.86	1	1	0	0	dichot. choice	1	0
Neill, *et al.* (1994)	1	31.00	10.00	3.10	1	1	1	0	Vickrey	0	1
Neill, *et al.* (1994)	1	301.00	12.00	25.08	1	1	1	0	Vickrey	0	0

Appendix A (cont.). Data Used in the Analysis

Study	LGStudy	HypValue	ActValue	CF	Student	Group	Private	Within	Elicitation	Choice	Calibrate
Sinden (1988)	1	0.70	0.92	0.76	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.43	1.86	0.77	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.01	1.28	0.79	1	1	0	1	open-ended	0	0
Sinden (1988)	1	0.79	0.92	0.86	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.10	1.28	0.86	1	1	0	1	open-ended	0	0
Sinden (1988)	1	2.40	2.76	0.87	0	1	0	1	open-ended	0	0
Sinden (1988)	1	0.84	0.92	0.91	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.06	1.12	0.95	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.30	1.28	1.02	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.36	1.28	1.06	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.22	1.12	1.09	1	1	0	1	open-ended	0	0
Sinden (1988)	1	2.06	1.86	1.11	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.60	1.40	1.14	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.60	1.40	1.14	1	1	0	1	open-ended	0	0
Sinden (1988)	1	2.38	1.87	1.27	1	1	0	1	open-ended	0	0
Sinden (1988)	1	1.27	0.92	1.38	1	1	0	1	open-ended	0	0
Sinden (1988)	1	2.10	1.40	1.50	1	1	0	1	open-ended	0	0
Spencer, *et al.* (1998)	1	4.70	3.16	1.49	1	1	0	0	tri-chot. choice	1	0
Spencer, *et al.* (1998)	1	3.24	2.10	1.54	1	1	0	0	tri-chot. choice	1	0
Vossler, *et al.* (2003)	0	49.67	48.89	1.02	0	0	0	0	referendum	1	1
Vossler, *et al.* (2003)	0	75.43	48.89	1.54	0	0	0	0	referendum	1	1
Vossler and Kerkvliet (2003)	0	52.27	51.75	1.01	0	0	0	0	referendum	1	0

References

Balistreri, E., G. McClelland, G. Poe and W. Schulze (2001), 'Can Hypothetical Questions Reveal True Values? A Laboratory Comparison of Dichotomous Choice and Open-Ended Contingent Values with Auction Values,' *Environmental and Resource Economics*, **18**, 275-292.

Bishop, R. C. and T. A. Heberlein (1979), 'Measuring Values of Extramarket Goods: Are Indirect Measures Biased? ' *American Journal of Agricultural Economics*, **61**, 926-930.

Bishop, R. C. and T. A. Heberlein (1986), 'Does Contingent Valuation Work?' in Cummings, R., D. Brookshire and W. Schulze, eds., *Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Valuation Method*. Totowa, NJ: Rowman and Allenheld.

Bishop, R. C. and T. A. Heberlein (1990), 'The Contingent Valuation Method,' in Johnson, R. L. and G. V. Johnson, eds., *Economic Valuation of Natural Resources*. Boulder, CO: Westview Press, pp. 81-204.

Blumenschein, K., M. Johannesson, G. C. Blomquist, B. Liljas and R. M. O'Connor (1997), 'Hypothetical versus Real Payments in Vickery Auctions,' *Economics Letters*, **56**, 177-180.

Blumenschein, K., M. Johannesson, K. K. Yohoyama and P. R. Freeman (2001), 'Hypothetical versus Real Willingness to Pay in the Health Care Sector: Results from a Field Experiment,' *Journal of Health Economics*, **20**, 441-457.

Bohm, P. (1972), 'Estimating the Demand for Public Goods: An Experiment,' *European Economic Review*, **3**, 111-130.

Botelho, A. and L. Costa Pinto (2002), 'Hypothetical, Real, and Predicted Willingness to Pay in Open-Ended Surveys: Experimental Results,' *Applied Economics Letters*, **9**, 993-996.

Boyce, R. R., T. C. Brown, G. D. McClelland, G. L. Peterson and W. D. Schulze (1989), *Experimental Evidence of Existence Value in Payment and Compensation Contexts*. Presented at the Joint Meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W-133) and the Western Regional Science Association, San Diego, CA.

Boyce, R. R., T. C. Brown, G. D. McClelland, G. L. Peterson and W. D. Schulze (1992), 'An Experimental Examination of intrinsic Values as a Source of the WTA-WTP Disparity,' *American Economic Review*, **82**, 1366-1373.

Brookshire, D. S. and D. L. Coursey (1987), 'Measuring the Value of a Public Good: An Empirical Comparison of Elicitation Procedures,' *The American Economic Review*, **77**, 554-566.

Brown, K. M. and L. O. Taylor (2000), 'Do As You Say, Say As You Do: Evidence on Gender Differences in Actual and Stated Contributions to Public Goods,' *Journal of Economic Behavior and Organization*, **43**, 127-139.

Brown, T. (1984), 'The Concept of Value in Resource Allocation,' *Land Economics*, **60**, 231-246.

Brown, T. C., P. Champ, R. Bishop and D. McCollum (1996), 'Which Response Format Reveals the Truth About Donations to a Public Good?' *Land Economics*, **72**, 152-166.

Cameron, T. A., G. L. Poe, R. G. Ethier and W. D. Schulze (2002), 'Alternative Non-market Value Elicitation Methods: Are the Underlying Preferences the Same?' *Journal of Environmental Economics and Management*, **44**, 391-425.

Carlsson, F. and P. Martinsson (2001), 'Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? Application to the Valuation of the Environment,' *Journal of Environmental Economics and Management*, **41**, 179-192.

Carson, R. T., N. E. Flores, K. M. Martin and J. L. Wright (1996), 'Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods,' *Land Economics*, **72**, 80-99.

Champ, P. A., R. C. Bishop, T. C. Brown and D. W. McCollum (1997), 'Using Donation Mechanisms to Value Nonuse Benefits from Public Goods,' *Journal of Environmental Economics and Management*, **33**, 151-162.

Champ, P. A. and B. R. C. (2001), 'Donation Payment Mechanisms and Contingent Valuation: An Empirical Study of Hypothetical Bias,' *Environmental and Resource Economics*, **19**, 383-402.

Coursey, D. L., J. L. Hovis and W. D. Schulze (1987), 'The Disparity between Willingness to Accept and Willingness to Pay Measures of Value,' *The Quarterly Journal of Economics*, **102**, 679-690.

Cummings, R. G., G. W. Harrison and E. E. Rutström (1995), 'Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible?' *American Economic Review*, **85**, 260-266.

Dickie, M., A. Fisher and S. Gerking (1987), 'Market Transactions and Hypothetical Demand Data: A Comparative Study,' *Journal of the American Statistical Association*, **82**, 69-75.

Duffield, J. W. and D. A. Patterson (1992), *Field Testing Existence Values: Comparison of Hypothetical and Cash Transaction Values*. Presented at the Joint meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W-133) and the Western Regional Science Association, South Lake Tahoe, Nevada.

Fox, J. A., J. F. Shogren, D. J. Hayes and J. B. Kliebenstein (1998), 'CVM-X: Calibrating Contingent Values with Experimental Auction Markets,' *American Journal of Agricultural Economics*, **80**, 455-465.

Frykblom, P. (1997), 'Hypothetical Question Models and Real Willingness to Pay,' *Journal of Environmental Economics and Management*, **34**, 275-287.

Frykblom, P. (2000), 'Willingness to Pay and Choice of Question Format: Experimental Results,' *Applied Economics Letters*, **7**, 665-667.

Harrison, G. W. and E. E. Rutström (1999), 'Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods,' in Smith, V. L., ed, *Handbook of Results in Experimental Economics*. New York: Elsevier Science.

Harrison, G. W. and E. E. Rutström (2002), 'Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods,' in Plott, C. and V. L. Smith, eds., *Handbook of Results in Experimental Economics*. New York: Elsevier Science, forthcoming.

Heberlein, T. A. and R. Bishop (1986), 'Assessing the Validity of Contingent Valuations: Three Field Experiments,' *Science of the Total Environment*, **56**, 434-479.

Irwin, J. R., G. H. McClelland and W. D. Schulze (1992), 'Hypothetical and Real Consequences in Experimental Auctions for Insurance Against Low-Probability Risks,' *Journal of Behavioral Decision Making*, **5**, 107-116.

Irwin, J. R., P. Slovic, S. Lickenstein and G. McClelland (1993), 'Preference Reversals and the Measurement of Environmental Values,' *Journal of Risk and Uncertainty*, **6**, 5-18.

Johannesson, M., B. Liljas and P. Johansson (1998), 'An Experimental Comparison of Dichotomous Choice Contingent Valuation Questions and Real Purchase Decisions,' *Applied Economics*, **30**, 643-647.

Kealy, M., J. Dovidio and M. Rockel (1988), 'Accuracy in Valuation is a Matter of Degree,' *Land Economics*, **64**, 158-171.

Kealy, M. J., M. Montgomery and J. F. Dovidio (1990), 'Reliability and Predictive Validity of Contingent Values: Does the Nature of the Good Matter?' *Journal of Environmental Economics and Management*, **19**, 244-263.

List, J. A. (2001), 'Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards,' *American Economic Review*, **91**, 1498-1507.

List, J. A. (2003), 'Using Random nth Price Auctions to Value Non-Market Goods and Services,' *Journal of Regulatory Economics*, **23**, 193-205.

List, J. A. and C. Gallet (2001), 'What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values?' *Environmental and Resource Economics*, **20**, 241-254.

List, J. A. and J. F. Shogren (1998), 'Calibration of the Difference between Actual and Hypothetical Valuations in a Field Experiment,' *Journal of Economic Behavior and Organization*, **37**, 193-205.

List, J. A. and J. F. Shogren (2002), 'Calibration of Willingness to Accept,' *Journal of Environmental Economics and Management*, **43**, 219-233.

Loomis, J., T. Brown, B. Lucero and G. Peterson (1996), 'Improving Validity Experiments of Contingent Valuation Methods: Results of Efforts to Reduce the Disparity of Hypothetical and Actual Willingness to Pay,' *Land Economics*, **72**, 4450-4461.

Loomis, J., T. Brown, B. Lucero and G. Peterson (1997), 'Evaluating the Validity of the Dichotomous Choice Question Format in Contingent Valuation,' *Environmental and Resource Economics*, **10**, 109-123.

MacMillan, D. C., T. Smart and A. Thorburn (1999), 'A Field Experiment Involving Cash and Hypothetical Charitable Donations,' *Environmental and Resource Economics*, **14**, 399-412.

McClelland, G. H., W. D. Schulze and D. L. Coursey (1993), 'Insurance for Low-Probability Hazards: A Bimodal Response to Unlikely Events,' *Journal of Risk and Uncertainty*, **7**, 95-116.

McKenzie, J. (1993), 'A Comparison of Contingent Preference Models,' *American Journal of Agricultural Economics*, **75**, 593-603.

Murphy, J., T. Stevens and D. Weatherhead (2002), 'An Empirical Study of Hypothetical Bias in Voluntary Contribution Contingent Valuation: Does Cheap Talk Matter?' Department of Resource Economics Working Paper, 2003-2, University of Massachusetts-Amherst.

Navrud, S. (1992), 'Willingness to Pay for Preservation of Species - An Experiment with Actual Payments,' in Navrud, S., ed, *Pricing the European Environment*. Oslo: Scandinavian University Press, pp. 231-246.

Neill, H. R., R. G. Cummings, P. T. Ganderton, G. W. Harrison and T. McGuckin (1994), 'Hypothetical Surveys and Real Economic Commitments,' *Land Economics*, **70**, 145-154.

Sinden, J. A. (1988), 'Empirical Tests of Hypothetical Biases in Consumers' Surplus Surveys,' *Australian Journal of Agricultural Economics*, **32**, 98-112.

Spencer, M. A., S. K. Swallow and C. J. Miller (1998), 'Valuing Water Quality Monitoring: A Contingent Valuation Experiment Involving Hypothetical and Real Payments,' *Agricultural and Resource Economics Review*, **27**, 28-41.

Vossler, C. A. and J. Kerkvliet (2003), 'A Criterion Validity Test of the Contingent Valuation Method: Comparing Hypothetical and Actual Voting Behavior for a Public Referendum,' *Journal of Environmental Economics and Management*, forthcoming.

Vossler, C. A., J. Kerkvliet, S. Polasky and O. Gainutdinova (2003), 'Externally Validating Contingent Valuation: An Open-Space Survey and Referendum in Corvallis, Oregon,' *Journal of Economic Behavior and Organization*, forthcoming.

NOTES

1 The Carson*, et al.* (1996) comparison of revealed and SP studies indicated a strong correlation (0.89) between hypothetical and market behavior, but since revealed preference measures, like estimates derived from travel cost studies, contain substantial unexplained variation, Carson et al. test SP convergent validity. List and Gallet (2001) and Harrison and Rutström (2002) test SP criterion validity because a 'true' measure of value is obtained from actual payments for the good being valued.

2 Thanks to John List and Craig Gallet for sharing their original data files for this analysis of their results.

3 Note that the term "median" calibration factor refers to the midpoint between the minimum and maximum values within a range, not the median of all the calibration factors in a single study.

4 In our reconsideration of the LG results, we also tested sensitivity to the functional form and got similar results.

5 Neill et al. (1994) also report a very high calibration factor (25.1), but since this was part of a range of values for which the median calibration factor was used, this value was not omitted.

6 LG use the natural log of calibration factor as the dependent variable in their model. It is straightforward to show that our equation (1) can also be specified using the log of the inverse of the calibration factor as the dependent variable: $\ln(CF^{-1}) = \beta_0 + \beta_1' \cdot lnHypValue + \beta_2 \cdot lnHypValue^2 + \varepsilon$ where $\beta_1' = \beta_1 - 1$. LG note that they also estimated a model using $\ln(CF^{-1})$ and found that this did not affect their conclusions.

7 This transformation required that six of the 83 observations be dropped due to negative lnHypValue.

8 We only did this simple comparison for Calibrate because none of the other dummy variables had a sufficient number of studies to conduct a within-study analysis of its effects.