A meta-analysis of hypothetical bias in stated preference valuation

A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation

JAMES J. MURPHY[1,*], P. GEOFFREY ALLEN[1], THOMAS H. STEVENS[1] and DARRYL WEATHERHEAD[2]
[1]*Department of Resource Economics and Center for Public Policy and Administration, Stockbridge Hall, 80 Campus Center Way, University of Massachusetts, Amherst, MA 01003, USA;* [2]*U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, USA; *Author for correspondence (e-mail: murphy@resecon.umass.edu)*

Accepted 18 August 2004

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 used the same mechanism for eliciting both hypothetical and actual values. The papers generated 83 observations with a median ratio of hypothetical to actual value of only 1.35, and the distribution has severe positive skewness. We find that a choice-based elicitation mechanism is important in reducing bias. We provide some evidence that the use of student subjects may be a source of bias, but since this variable is highly correlated with group experimental settings, firm conclusions cannot be drawn. There is some weak evidence that bias increases when public goods are being valued, and that some calibration methods may be effective at reducing bias. However, results are quite sensitive to model specification, which will remain a problem until a comprehensive theory of hypothetical bias is developed.

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

JEL classifications: C9, H41, Q26, Q28

1. Introduction

Stated preference (SP) survey techniques, such as the contingent valuation method (CVM), typically ask participants questions about their value for some non-market good. The hypothetical nature of these surveys – in both the payment for and provision of the good in question – can result in responses that are significantly greater than actual payments. This difference between stated and revealed values is often referred to as hypothetical bias.[1] Despite an abundance of studies, there is no consensus about the underlying causes of hypothetical bias or ways to calibrate survey responses for it.

At this juncture, two basic questions about hypothetical bias in SP valuation have become paramount. First, what is the 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 reassess the magnitude of bias present in SP studies. We also attempt to evaluate the effect of several 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 somewhat limited.

Bohm's seminal paper comparing hypothetical and actual values was published in 1972, but it was not until nearly a decade later that this literature began to grow. In the 1980s, much of the experimental hypothetical bias literature tested the overall validity of contingent valuation (Harrison and Rutström, forthcoming). In a series of papers, Bishop and Heberlein found that hypothetical values for hunting permits consistently exceeded actual values (Bishop and Heberlein 1979, 1986; Heberlein and Bishop 1986).[2] On the other hand, Dickie et al. (1987) found that values for pints of strawberries elicited in a hypothetical survey were consistent with those observed when individuals were given an opportunity to actually purchase the good.[3] Subsequent research consistently suggested that values derived from surveys typically exceed actual values (e.g., Cummings et al. 1995; Fox et al. 1998; List and Shogren 1998), sometimes by a substantial margin (e.g., Neill et al. 1994). There are exceptions to the conclusion about the existence of hypothetical bias (e.g., Johannesson 1997; Sinden 1988; Smith and Mansfield 1998), but these studies appear to be in the minority: in a recent survey of the literature, Harrison and Rutström (forthcoming) found a positive bias in 34 of 39 observations. The mean bias in these 39 observations was about 300%, however this comes from a skewed distribution with a median closer to 67%.

In the last few years, there have been several attempts to synthesize the plethora of hypothetical bias studies in an attempt to find some common denominators. Foster et al. (1997) present a simple table summarizing 13 studies that highlights two main points: (1) although the primary purpose of contingent valuation is to value public goods, most of the studies focus on private goods, and (2) there are significant methodological differences between the studies, such as the type of comparison or the elicitation mechanism. Harrison and Rutström (forthcoming) present a review of the literature which clearly demonstrates that "the weight of the evidence supports the claim that hypothetical valuations exceed real valuations". However, they do not attempt to identify factors that might be associated with hypothetical bias, instead noting that they are unable to draw any broad conclusions. List and Gallet (2001) update Foster et al.'s table and then use a meta-analysis to explore whether there are any systematic relationships between these methodological differences and hypothetical bias.[4] Their results indicate that the magnitude of hypothetical bias was statistically less for (a) willingness-to-pay

(WTP) as compared to willingness-to-accept (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.

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

The remainder of this paper proceeds as follows. Section 2 presents our data and Section 3 describes the model and estimation results. Section 4 concludes with a summary of our findings. 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.

2. Description of Data

Meta-analysis can be very sensitive to outliers and a lack of variability in the data. For example, List and Gallet's (2001, hereafter LG) meta-analysis of hypothetical bias in stated values includes dummy variables for whether the study used a WTA or a WTP format, the type of experiment (lab or field), type of good (public or private), type of comparison (within or between subjects), and eight different elicitation mechanisms. Most of the elicitation mechanisms have just one study using that format, and there are only eight WTA observations. 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 others. 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 actual and hypothetical values (Smith auction and open-ended, 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.

After updating the LG data for coding differences (see endnote 4) and testing for the sensitivity of their results to particular observations, two of LG's main conclusions change: (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) a few elicitation mechanisms remain significant, but most of these variables are based on just a single study and, therefore, should be interpreted with caution. Their result that hypothetical bias is lower for private goods is robust throughout the sensitivity analysis.

In an attempt to avoid these and related data problems, we used the following criteria for determining whether to include an observation in our dataset:

- 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 elicited using the same mechanism. 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.
- 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.[5]

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). After imposing these restrictions, our data set includes 28 studies yielding 83 observations. The hypothetical values range between 0.08 and 301; the mean is 26.55, median 7.18, and standard deviation 47.33. The actual values range between 0.07 and 95.5; the mean is 11.69, median 3.67, and standard deviation 18.05. We assume that actual cash-based estimates are unbiased measures of the true WTP. Consistent with LG and Harrison and Rutström (forthcoming), the mean CF in our data is 2.60. However, as in the other datasets, this may be misleading as it comes from a highly skewed distribution with a 1.35 median CF. Figure 1 presents the distribution of CFs.

For our econometric analysis of hypothetical bias we defined the independent variables as follows. 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 *versus* campus) and the subject pool (students *versus* 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%. The other

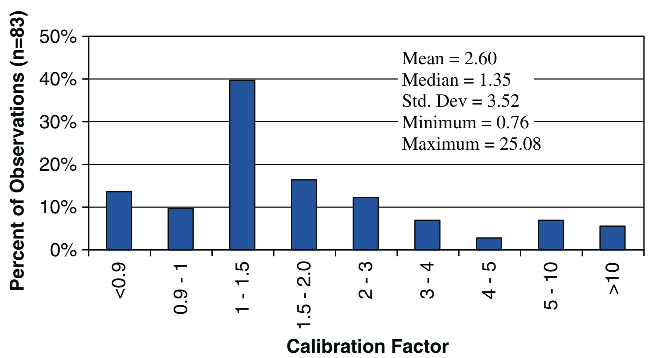

Figure 1. Distribution of calibration factors.

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 referendum is normally associated with a public good, and a Vickrey 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 mechanism (dichotomous and polychotomous choice, referendum, payment card and conjoint).

Some studies report simple descriptive statistics such as mean WTP (e.g., Bohm 1972). However, there has been a recent growth in the number of studies that utilize calibration techniques to control for hypothetical bias. Studies that employ *ex ante*, or instrument calibration, techniques, such as budget reminders (Loomis et al. 1996) or cheap talk scripts (Cummings and Taylor 1999; List 2001), attempt to get unbiased responses from participants. *Ex post*, or statistical calibration techniques, on the other hand, recognize that responses are biased and attempt to control for it using lab experiments to calibrate field data (Fox et al. 1998) or uncertainty adjustments (Champ et al. 1997; Poe et al. 2002). The variable Calibrate equals one if the observation is based on any type of calibration technique.

3. 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

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

where lnActValue and lnHypValue denote the natural log of the actual and hypothetical values.[6, 7]

Because White's test indicates the presence of heteroskadasticity (P-value = 0.0002), Table I reports the results from a weighted regression, using the square root of lnHypValue to transform the data.[8] 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 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

Table I. Regression results using all observations[a]

Variable	Base model		Expanded model				Trimmed model[b]			
	Model 1a		Model 2a		Model 2b		Model 3a		Model 3b	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard 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	< 0.0001		< 0.0001		< 0.0001		< 0.0001		< 0.0001	

*** Significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.

[a]Weighted least squares estimates. Dependent variable is the natural log of the actual value (lnActValue).

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

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.

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

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

The results for Model 2a are in Table I. 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-based elicitation mechanisms 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 directly express ambivalence, indifference or uncertainty. 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. The results of this regression are in Table I, 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. 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 I, 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. Regression results are available on the authors' web site. Consistent with the results in Table I (which uses the full data set), the hypothetical value seems to be the best predictor of actual value (for every regression, 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).

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.[9] 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.

4. 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 estimated, the coefficients for the Group/Student and Choice dummy variables were consistently significant and of relatively large magnitude. In addition, a comparison of calibration factors within particular studies indicates that calibration techniques are effective at reducing hypothetical bias. We also find that LG's conclusion that hypothetical bias is greater in WTA studies is based on only eight observations and is driven by a pair of extreme values from a single study that use different elicitation mechanisms for hypothetical and actual values. We exclude WTA studies from our data because there are insufficient studies to incorporate this variable in a meta-analysis.

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 key findings differ from those reported in previous research. 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. The mean CF for the top 10 observations is 10.3, compared with 1.54 for the other 73 observations. This suggests 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, provision point mechanisms and Smith auctions are only associated with public goods, and Vickrey auctions only with private goods. 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.

Finally, we note that discussions that focus solely on the mean calibration factor could be misleading because of the large disparity between the mean and median calibration factors. As shown in Figure 1, the overwhelming majority of observations have relatively low CFs, possibly suggesting that hypothetical bias may not be as significant a problem in stated preference analyses as is often thought. On the other hand, a small but non-trivial number of observations have rather large CFs. We were unable to identify any systematic patterns in these observations. Although our trimmed model excludes the largest CFs, it would be premature to simply dismiss these observations as outliers. Rather, it is important to develop a better understanding of the conditions under which these large CFs arise.

Acknowledgements

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, US 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.

Notes

1. The terms revealed, real and actual values are used interchangeably and refer to situations in which an individual makes a consequential economic commitment – in experimental studies, this typically involves payment for a good by the participant. Most studies of hypothetical bias assume that these cash-based estimates are unbiased. On the other hand, stated or hypothetical values refer to survey responses that lack any salient economic commitment.
2. Hanemann (1984) highlights the sensitivity of this conclusion.

3. Harrison and Rutström (forthcoming) argue that a more detailed examination their data yields mixed results, and that, on average, hypothetical values exceed actual values by 58%.
4. Because there are a few typos and coding errors in the List and Gallet table, and because variations of this table appear in four separate journal articles, the authors' web site contains a brief comment that identifies and corrects these. The URL is http://www.umass.edu/resec/faculty/murphy/meta/meta.html. This web site also contains the data and some supplemental tables.
5. The data in Harrison and Rutström (forthcoming) include both dollar-based estimates and values derived from yes–no studies. We refrain from combining these two types of responses because calibration factors derived from each type of response have different interpretations.
6. The quadratic term in this equation allows for the possibility that there is a non-linear relationship. A simple linear relationship is a special case in which $\beta_2 = 0$.
7. 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: $\ln(CF^{-1}) = \beta_0 + \beta_1' \cdot \text{lnHypValue} + \beta_2 \cdot \text{lnHypValue}^2 + \varepsilon$ 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.
8. This transformation required that six of the 83 observations be dropped due to negative lnHypValue.
9. 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.

References

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 R. Cummings, D. Brookshire and W. Schulze, eds., *Valuing Environmental Goods: A State of the Art Assessment of the Contingent Valuation Method*. Totowa, New Jersey: Rowman and Allenheld.

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

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

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.

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.

Cummings, R. G. and L. O. Taylor (1999), 'Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method', *The American Economic Review* **89**, 649–665.

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.

Foster, V., I. J. Bateman and D. Harley (1997), 'Real and Hypothetical Willingness to Pay for Environmental Preservation: A Non-Experimental Comparison', *Journal of Agricultural Economics* **48**, 123–138.

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.

Hanemann, W. M. (1984), 'Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses', *American Journal of Agricultural Economics* **66**, 332–341.

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

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., 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. (1997), 'Some Further Experimental Results on Hypothetical versus Real Willingness to Pay', *Applied Economics Letters* **4**, 535–536.

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.

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.

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

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.

Poe, G. L., J. E. Clark, D. Rondeau and W. D. Schulze (2002), 'Provision Point Mechanisms and Field Validity Tests of Contingent Valuation', *Environmental and Resource Economics* **23**, 105–131.

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

Smith, V. K. and C. Mansfield (1998), 'Buying Time: Real and Hypothetical Offers', *Journal of Environmental Economics and Management* **36**, 209–224.