

REVISITING THE DATA AND ESTIMATION IN LIST AND GALLET (2000)

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

December 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.

Revisiting the List and Gallet Data and Results ¹

List 2001 (hereafter LG), conducted a meta-analysis of hypothetical bias in stated values. As a corollary to our meta-analysis (Murphy, *et al.* 2004), we present a summary of their study and a sensitivity analysis of their findings. Table I in LG presents a summary of their data. Because there are a few typos and coding errors in this table, and because variations of this table appear in four separate journal articles (List and Shogren 1998; List and Gallet 2001; List and Shogren 2002; List 2003), we also identify and correct these errors.

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.² 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.

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

2 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.

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.³

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.

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

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

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.

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.

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.⁴ 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

4 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.

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.

In our paper (Murphy et al. 2004), 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. 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. Finally, we note that discussions that focus solely on the mean calibration factor could be misleading because of the skewed distribution, yielding a median calibration factor that is about half the mean.

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 Schulze, W., ed, *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, G. V., ed, *Economic Valuation of Natural Resources*. Boulder, CO: Westview Press, pp. 81-204.
- Bohm, P. (1972), 'Estimating the Demand for Public Goods: An Experiment,' *European Economic Review*, **3**, 111-130.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Murphy, J. J., T. H. Stevens, P. G. Allen and D. Weatherhead (2004), 'A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation,' *Environmental and Resource Economics*, **in review**.
- 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.
- 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.

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) *
<i>Type of elicitation:</i>				
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, <i>et al.</i> 1988	1.00 - 2.00	ERE typo	Study	1988	1990	Typo in ERE paper
Irwin, <i>et al.</i> 1992	1.00; 2.50	ERE typo	all			Study is in LG regression data, but missing from ERE paper
Kealy, <i>et al.</i> 1988	1.40	ERE typo	all			Study is in LG regression data, but missing from ERE paper
Kealy, <i>et al.</i> 1988	1.30	ERE typo	all			Study is in LG regression data, but missing from ERE paper. Observation from Kealy, <i>et al.</i> 1990.
Loomis, <i>et al.</i> 1996	2.00 - 3.60	ERE typo	elicitation	dc	open-ended	Correct in LG data, but typo in Table V in ERE paper
Boyce, <i>et al.</i> 1992	0.90	could not find			not included	Could not find this observation in the paper.
Kealy, <i>et al.</i> 1990	1.30	could not find			not included	Could not find this observation in the paper.
Balistreri, <i>et al.</i> 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, <i>et al.</i> 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, <i>et al.</i> 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, <i>et al.</i> 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, <i>et al.</i> 1993	2.20	error	comparison	between	within	
McClelland, <i>et al.</i> 1993	0.80	error	comparison	between	within	
Navrud 1992	3.20	error	good	private	public	
Neill, <i>et al.</i> 1994	3.10-25.10	error	elicitation	open-ended	Vickrey	For these two CFs, both hyp and actual use Vickrey
Spencer, <i>et al.</i> 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	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, <i>et al.</i> 2001	1.25	interpret	elicitation	open-ended	not included	Hypothetical and actual elicitation mechanisms differ.
Balistreri, <i>et al.</i> 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, <i>et al.</i> 1998	1.20, 1.50	interpret	elicitation	open-ended	not included	Hypothetical and actual elicitation mechanisms differ.
Irwin, <i>et al.</i> 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, <i>et al.</i> 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.