10. Hypothetical bias: a new meta-analysis

Harry Foster and James Burrows[1]

INTRODUCTION

Participants in hypothetical surveys or referenda typically express higher values for goods than do participants faced with similar choices in which the stakes involve real money. Previous meta-analyses have confirmed the widespread presence of hypothetical bias in stated preference studies and have identified certain factors associated with higher or lesser degrees of bias. These studies, and indeed the broader stated preference valuation literature, have not offered any definitive insights that can reliably be used to eliminate these biases. The earlier meta-analyses are now dated and were based on a limited number of studies.

In this chapter we assess the evidence from the literature up to the present time on hypothetical bias. We include many more papers touching on hypothetical bias than were available to or used by the authors of the prior meta-analyses. We also add two variables (not analyzed in the existing literature) to our meta-analysis: one that is designed to capture whether the good in question is likely to be perceived as familiar or unfamiliar to the study's survey participants and a second that indicates whether or not the valuation of the good in question is largely or exclusively generated by non-use considerations.

In the remainder of this chapter, we first discuss how our meta-data were created. We then identify and briefly discuss some of the issues in survey design that have been hypothesized to contribute to the presence or extent of the hypothetical bias exhibited in various studies. We then present results from a regression analysis of the meta-data, and follow with some concluding remarks.

[1] Respectively, Principal, and Vice Chairman, Charles River Associates, Boston.

DATA

Our initial sample for our meta-analysis includes all of the relevant comparisons of mean willingness to pays (WTPs) between hypothetical and real survey treatments that we were able to draw from the papers cited in one or more of three previous meta-analyses: those of List and Gallet (2001), Little and Berrens (2004), and Murphy et al. (2005).[2] Similar to the practice adopted in the first two of these studies and in one of the analyses done in Murphy et al. (2005), we analyze only those comparisons from studies that included explicit calculations of mean WTPs across both hypothetical and real treatments. We thus eliminated from further analysis those studies that merely provide percentage yes/no results drawn from dichotomous choice referendum questions and that did not go on to calculate population mean WTPs. To this sample of results drawn from previously cited papers, we added data comparing WTPs drawn from additional papers not cited in prior meta-analyses.[3] The unit of observation is a comparison between a hypothetical WTP and a corresponding real WTP for the same good drawn from the same paper. Any particular paper could contribute one or multiple observations to the dataset, with the number depending on the number of survey variants conducted as a part of the paper's survey design. All told, our sample includes 432 comparisons between hypothetical and real results drawn from 77 studies. These studies are listed in the Bibliography with an asterisk.

For each of the comparisons of inferred WTP from CV surveys to inferred WTP from real transactions involving money, we calculate a "bias ratio" (BR) defined as the ratio of the mean WTP drawn from the hypothetical treatment to the mean WTP drawn from the real treatment. A histogram of the BRs found in Figure 1 provides summary data on the bias ratios we derived from the observations contained in our meta-data. The median value in the distribution is 1.39, while the mean value is 2.33.

[2] In addition to these three papers, we consulted two more recently published meta-analyses as potential sources of papers to our database. Schläpfer and Fischhoff (2010) relies upon the same sample of papers as was used in Murphy et al. (2005), while Little et al. (2012) does not provide a list of the studies it relied upon in creating its dataset.

[3] To assemble our database of studies related to measuring hypothetical bias ratios, we supplemented the articles cited in the prior meta-analyses and in another published overview of the hypothetical bias literature (Harrison and Rutström, 2008) by searching the EVRI (Environmental Valuation Reference Inventory) and NOEP (National Ocean Economics Program – Middlebury College) databases, government websites and publication sources (including NOAA, EPA, and the US Fisheries and Wildlife Agency, among others), academic websites (including Richard Carson's invaluable website for collected studies, accessed December 9, 2016 at https://ideas.repec.org/i/p.html), the mammoth bibliography in Carson, 2012, EBSCO, Econlit, and Google Scholar. In addition to these sources, we also cross-referenced citations in all the articles we identified.

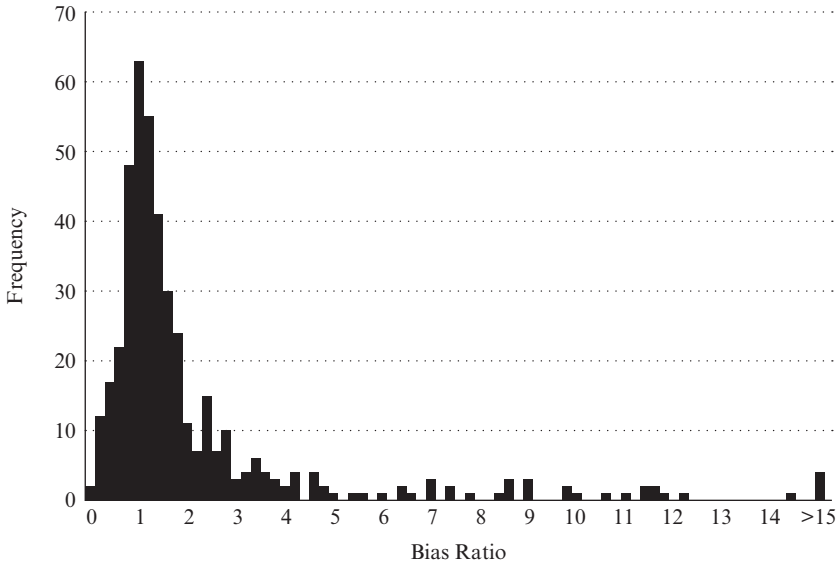

Figure 1 Bias ratio frequency distribution

The range of BRs exhibited in the distribution is relatively large, with a 5th percentile BR of 0.50 and a 95th percentile value of 8.66. Notably, the shape of the distribution of BRs provided in Figure 1 suggests that the values found in the dataset follow something like a log-normal distribution. Figure 2, which displays a histogram of the BRs arrayed on a logarithmic scale, confirms that the BRs are indeed consistent with a log-normal distribution.

We calculate a bias ratio (BR) for each observation retained in our data and assign a series of indicator variables to each comparison reflecting factors present or absent in the study's design that have been hypothesized in the literature to influence the extent of hypothetical bias. These factors are:

- whether or not the BR was calculated through use of an ex-post certainty correction;
- the presence or absence of a cheap talk script in the survey instrument;
- whether the hypothetical and real observations are drawn from a survey in which a single group of participants are asked to respond to both hypothetical and real treatments (same) or are drawn from two separate survey panels (different);
- whether or not the study uses a conjoint/choice experiment framework rather than any other type of contingent valuation;

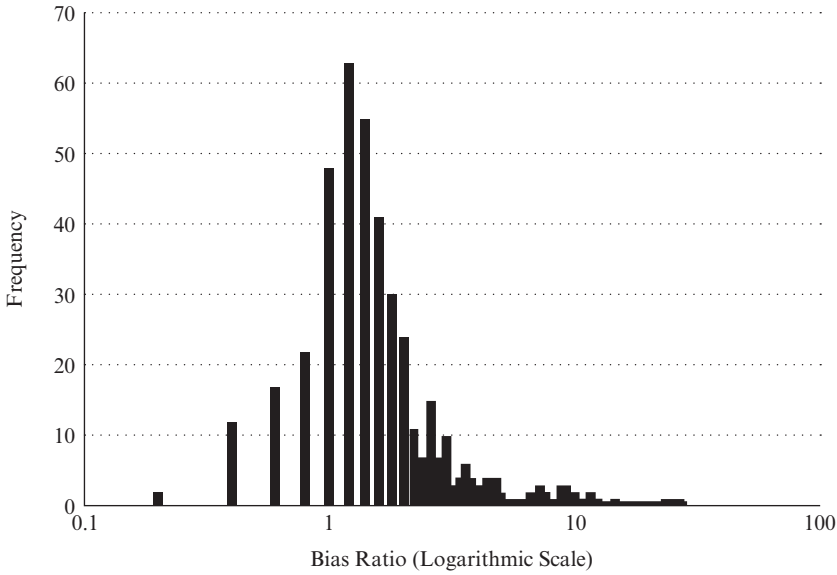

Figure 2 Bias ratio frequency distribution

- whether or not the survey group consists entirely of students;
- whether or not the survey was administered in a laboratory setting;
- whether the good is a public good or a private good;
- whether the good is likely to be perceived as a familiar or an unfamiliar one by the survey's participants;
- whether the perceived benefits to the survey participants of providing the good are generated primarily by non-use considerations.

Each of these nine factors can be used by itself to divide the full dataset into a pair of non-overlapping and fully inclusive subsamples. Table 1 provides summary statistics on the median, mean, and standard deviations for all of the 18 subsamples that can be created in this way. In addition to the median, Table 1 also provides the 5th percentile and 95th percentile BRs for each subsample and *p*-values associated with an equality of the means test across each relevant pairing.

Certainty correction

Certainty correction takes a value of one if the observation is derived from a hypothetical treatment employing an ex post certainty correction and is set to zero otherwise. This variable is meant to control for the

Table 1 Summary statistics for bias ratio by observation type

	Full Dataset
Mean	2.329
Median	1.388
Standard deviation	3.138
5–95%	0.495–8.659
Number of observations	432
P–value for two independent samples *t*–test	N/A

	Cheap Talk	*No Cheap Talk*	*Same*	*Not Same*	*Lab*	*Non-lab*
Mean	1.620	2.422	2.088	2.428	1.779	2.722
Median	1.410	1.381	1.214	1.471	1.275	1.424
Standard deviation	1.306	3.294	2.374	3.399	1.609	3.833
5–95%	0.205–4.156	0.531–8.778	0.698–8.500	0.422–8.659	0.507–4.203	0.422–10.188
Number of observations	50	382	125	307	180	252
P–value	0.0016		0.2385		0.0005	

	Private	*Public*	*Conjoint*	*Non-conjoint*	*Student*	*Non-student*
Mean	2.458	2.240	1.798	2.456	2.410	2.286
Median	1.630	1.260	1.419	1.378	1.233	1.456
Standard deviation	3.176	3.115	1.493	3.404	3.714	2.787
5–95%	0.700–7.912	0.402–8.778	0.333–4.156	0.545–9.083	0.495–9.167	0.554–7.067
Number of observations	177	255	83	349	151	281
P–value	0.4790		0.0077		0.7212	

Table 1 (continued)

	Familiar	*Unfamiliar*	*Non-use*	*Use*	*Certainty correction*	*Non-certainty correction*
Mean	2.213	2.423	2.631	2.076	0.875	2.578
Median	1.362	1.411	1.412	1.359	0.769	1.499
Standard deviation	3.030	3.226	3.501	2.781	0.513	3.327
5–95%	0.698–7.067	0.409–9.471	0.417–10.667	0.513–5.417	0.299–1.573	0.789–9.053
Number of observations	192	240	197	235	63	369
P–value	0.4872		0.0726		0.0000	

use in several studies of certainty correction techniques in an attempt to reconcile the differences between mean WTPs exhibited in paired hypothetical and real survey treatments by using data drawn from follow-up questions asking survey recipients to rate how sure they are in the answer they gave to the valuation question. Most commonly, this certainty question asks recipients to rate their degree of certainty in their answers to a hypothetical valuation question on a numerical scale, typically running from one to ten, with one representing "very uncertain" and ten representing "very certain." Alternatively, some studies dispense with creating a numerical scale, instead asking survey participants to indicate the degree of certainty in their responses by choosing the phrase that best describes their level of certainty from a set of qualitative options (for example, "very uncertain," "somewhat uncertain," "somewhat certain" or "highly certain") presented to them in the survey instrument. Such studies typically find that reasonable agreement between hypothetical and real treatments can be obtained if the set of hypothetical responses used to calculate WTPs is limited to only the survey responses given by those who indicate a degree of certainty that meets or exceeds some cut-off value or, alternatively, opt for qualitative descriptions of their levels of certainty that indicate a relatively high level of certainty. The researcher chooses the appropriate cut-off point for degree of certainty through an ex post determination of which particular value brings the certainty-corrected hypothetical WTP into closest agreement with the WTP derived from a real treatment. The particular cut-off value that is determined in this manner is survey specific and cannot be predicted a priori. For example, of six studies cited in Morrison and Brown (2009) that employed a ten-point certainty scale, two found closest agreement between real and hypothetical values if the analysis of WTP was limited to only responses associated with a degree of certainty of seven or greater, while two other studies found an optimal cut-off at eight, and two found that including only responses equal to ten brought the best fit between hypothetical and real WTP values. Because researchers actively choose, on an ex post basis, the certainty cut-off to be applied to the hypothetical treatment data to mimic the results obtained from an analogous real treatment, calibration factors between certainty-corrected hypothetical WTP results and real treatment WTP values will by design cluster near a value of one. For this reason, our regression models control for observations drawn from "certainty-corrected" or "certainty-calibrated" results. We are unaware of any paper that has analyzed how to set a certainty correction ex ante – that is, there is no procedure available to know what the "right" certainty correction is in advance.

Cheap talk

Cheap talk is set at 1 if the hypothetical treatment used in comparing hypothetical and real responses utilized a "cheap talk" script and set to zero otherwise. In this approach, survey respondents in the hypothetical treatment group are asked to answer any valuation questions only after they have first been presented with a script informing them of the tendency of participants in prior hypothetical surveys to overstate WTPs and asking them to keep this fact in mind whenanswering the survey's questions. Most, though not all, studies on the efficacy of cheap talk scripts have found that mean WTPs derived from treatments utilizing cheap talk scripts tend to be lower than those derived from similar hypothetical treatments lacking a cheap talk script, although the differences in the results obtained between the treatments may or may not be statistically significant. Consistent with the expectation that cheap talk scripts should generally dampen the extent of hypothetical bias, we find that the mean BR for study treatments included in our meta-analysis that include a cheap talk script (1.62) is lower than the mean bias ratio for study treatments lacking a cheap talk script (2.42). This difference is significant at any conventional level of statistical significance.

Same respondents vs different respondents

The indicator variable *Same* is set to one if the survey design has the same person answering the hypothetical and real survey treatments – that is, each participant is first asked to answer hypothetical valuation questions and then is asked to make a real purchase or contribution for the same good. In the alternative "different" sample treatment, participants are divided into two groups, with one group answering only a hypothetical survey and the other group subjected to only the real treatment. Treatments that rely on the same individuals to provide responses for both the hypothetical and then the real valuation exercises are believed to generate smaller bias ratios than treatments that rely on separate and different hypothetical and real treatment groups. When the same participants are asked to respond to a hypothetical treatment and then to a real treatment, their real purchase behavior may be influenced in an upward direction by anchoring or by conscious or unconscious desires to have their real decisions bear a relationship to their answers to the questions asked them in the prior hypothetical treatment. In our sample, the mean bias ratio for within-sample comparisons is just slightly lower than that for between comparisons (2.09 vs 2.43), and the difference between these two values is not statistically significant.

Conjoint/choice experiment

The indicator variable *Conjoint* is set to one if the observation comes from a study using conjoint or choice experiment techniques in which WTPs are derived indirectly from the pattern of choices individual participants express when asked to choose among hypothetical goods that differ in their product attributes. *Conjoint* is otherwise set to zero otherwise, as is the case for studies using some variant of a contingent valuation survey. Some proponents of the choice experiment framework have claimed that it is less susceptible to hypothetical bias than are contingent valuation techniques. In our sample, the mean bias ratio for conjoint/choice experiment elicitation formats is 1.80, versus a mean value of 2.46 for studies using any one of several other elicitation techniques. The difference between these two means is statistically different at any conventional level of statistical significance.

Student

The indicator variable *Student* is set to 1 if a study's participants consist entirely of students and to zero otherwise. It has been hypothesized that survey responses from panels comprised of students are likely to reflect a greater degree of hypothetical bias than are those from panels drawn from predominantly non-student populations. In our sample, the mean bias ratio derived from experiments using only student participants is 2.41, while the mean bias ratio derived from studies that used non-student or mixed survey populations is slightly smaller at 2.29. The difference between these two means is not statistically significant.

Lab experiment

The indicator variable *Lab i*s set to one if the hypothetical and real survey instruments were administered in a laboratory study and to zero otherwise. In our sample, the mean bias ratio derived from experiments conducted in laboratory settings is 1.78, while that derived from studies conducted in other settings is larger, at 2.72. The difference between these two subsamples means is statistically significant at all conventional levels of statistical significance.

Private good/public good

The indicator variable *Private* is set to one if the good that is the subject of the study is a private good and to zero otherwise. Valuations for public goods might be expected to exhibit greater hypothetical bias than those for private goods, given the far greater familiarity survey participants have in engaging in transactions for the purchase of private goods. Contrary to these expectations, in our subsamples the mean value of BR derived

from experiments valuing private goods (2.46) is slightly greater than that derived from the public goods subsample (2.24). The difference between these two means is not statistically significant.

Familiar good

Based upon our own best judgment, we have classified observations into those we believe are for goods that are familiar to the population being surveyed and those that are unfamiliar to them. As might be expected, the mean bias ratio derived from experiments valuing familiar goods is lower than that from experiments valuing unfamiliar goods (2.21 vs 2.42, respectively), although the difference in means is not statistically significant. As far as we are aware, this study is the first of its kind to rely upon this distinction to create an explanatory variable for use in a meta-analysis.

Non-use

This variable, the assignment of which is based upon our best judgment, is set to one for goods that we believe generate all or most of their perceived value from non-use considerations. The mean correction factor in our survey is 2.63 for non-use-value goods and 2.08 for use-value goods. The difference between these two means is statistically significant at the 10% level. As is the case for the creation of the familiar/unfamiliar distinction described above, we believe that this study is the first to use this distinction to create an explanatory variable to be used in a meta-analysis.

REGRESSION ANALYSIS

Base Model

Having assigned the appropriate indicator variable values to each observation retained in our data we then estimated an equation of the form:

$$\ln(BR) = \alpha + \beta 1 * Certaintycorrection + \beta 2 * Cheap\ talk + \beta 3 * Same + \beta 4 * Conjoint + \beta 5 * Student + \beta 6 * Lab + \beta 7 * Private + \beta 8 * Familiar + \beta 9 * Non\text{-}use$$

where ln (BR), the natural logarithm of the bias ratio for each observation, is the dependent variable and the explanatory variables are a constant and the various indicator variables are as defined in the previous section.

Results of this initial OLS regression analysis are displayed in columns 1 and 2 of Table 2. Because some studies contribute multiple observations to the data, we follow the practice of Little and Berrens (2004) and estimate and report clustered standard errors, with each paper represented in the

Table 2 Regression coefficients for reference and linear specifications

Variable	(1) ln(BR) Unweighted Non-use	(2) ln(BR) Weighted Non-use	(3) BR Unweighted Non-use	(4) BR Weighted Non-use
Certaintycorrection	−0.8907***	−0.8845***	−1.8919***	−2.0200***
	(0.10)	(0.15)	(0.37)	(0.46)
Cheaptalk	−0.3627**	−0.4547**	−0.8145	−1.1727*
	(0.17)	(0.22)	(0.56)	(0.63)
Same	−0.1069	0.0175	−0.3865	0.1047
	(0.20)	(0.15)	(0.71)	(0.60)
Conjoint	−0.0331	−0.0748	−0.0507	−0.1277
	(0.11)	(0.19)	(0.42)	(0.64)
Student	0.1070	0.1879	1.4867	1.4773
	(0.23)	(0.21)	(1.08)	(0.94)
Lab	−0.1564	−0.3133*	−1.5703	−1.9390**
	(0.21)	(0.19)	(0.95)	(0.82)
Private	0.5205***	0.6246***	1.1969**	1.7810**
	(0.14)	(0.16)	(0.51)	(0.75)
Familiargood	−0.0871	−0.2716*	−0.2516	−0.6334
	(0.14)	(0.15)	(0.45)	(0.53)
Non-use	0.3716*	0.4503**	0.8287	1.6961**
	(0.19)	(0.18)	(0.63)	(0.77)
Constant	0.3447*	0.4379**	2.1994***	2.0305**
	(0.20)	(0.19)	(0.70)	(0.78)
Observations	432	432	432	432
R-squared	0.2349	0.2329	0.1112	0.1556
RMSE	0.706	0.726	2.990	3.104
Degrees of freedom	422	422	422	422

Note: Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

data forming a separate cluster. In addition, we estimate the equation using both an unweighted and a weighted sample; in the former version, each observation carries equal weight in the estimation, no matter how many other observations may be drawn from the same paper, while in the latter each observation derived from a single study is weighted by the inverse of the number of comparisons in the dataset derived from the same study; that is, for each comparison from a paper contributing n observations to the dataset is weighted by a factor of $1/n$.

In general, most of the coefficients have the signs previously hypothesized for them in the literature. The coefficient on the *Certainty correction*

variables in the unweighted models is relatively large, negative, and statistically significant at the 1% level. The *Cheap talk*, *Same*, and *Conjoint* variables are associated with lower bias ratios, as is the variable meant to indicate whether or not the good being valued is a familiar one. The use of student subjects is associated with higher bias ratios. Potentially offsetting the effects of the *Student* variable, the coefficient on the lab variable is negative, indicating that conducting experiments in a lab setting is associated with lowered bias ratios.[4] The coefficient on the *Private* variable is positive in sign, as is the coefficient on *Non-use*. The coefficients on the *Certainty correction*, *Cheap talk*, *Private* and *Non-use* variables are the only ones to achieve statistical significance at conventional levels ($p < 0.05$) in both the weighted sample and unweighted sample regressions. The coefficient on the *Lab* variable is of marginal statistical significance ($p < 0.10$) in only the weighted sample regression. All of the other indicator variable coefficients fail to achieve statistical significance at conventional levels.

Functional Form

We explore whether the results displayed in columns 1 and 2 of Table 2 are robust with respect to choice of functional form by estimating regressions in which BR replaces ln (BR) as the dependent variable. The original specification is appropriate under an assumption that the effects of the explanatory variables on observed bias ratios are multiplicative in nature, while the choice of BR as dependent variable implicitly assumes that the effects of the same variables in determining observed bias ratios are additive in nature. Results from regression form in which BR serves as the dependent variable can be found in columns 3 and 4 of Table 2. All of the coefficients in the regressions in which BR is the dependent variable exhibit the same signs as those of their counterparts in the ln (BR) specification. With the sole exception of the *Cheap talk* variable in the unweighted BR specification, all of the indicator variables that achieve statistical significance in the ln (BR) specifications also achieve some level of statistical significance in the corresponding BR specifications, with only two exceptions (the *Cheap talk* variable in the unweighted models and the *Familiar* variable in the weighted models). The same relationship holds in the reverse comparison, as all of the variables that achieve statistical significance in the BR specifications also are of statistical significance in the ln (BR) regressions. The

[4] This negative coefficient could reflect differences between the relative frequencies with which lab-administered surveys and field experiments are designed to elicit valuation responses having a basis in "induced values" supplied by the experimenter rather than in survey participants' "homegrown" preferences.

consistency in the patterns of coefficient signs and significance across the two sets of specifications provides reassurance that the results produced with ln (BR) as the dependent variable are not driven by this particular choice of functional form.

Time Trend

We also explore whether the results produced by the reference specification are robust to the inclusion of a *Time trend* variable. Including a time variable in the regression specification controls for the possibility that, after controlling for the effects of the other explanatory variables, at least some of the variation in bias ratios might reflect ongoing refinements in methodology and the gradual adoption of "best practices" in conducting valuation studies. The *Time trend* variable is based on year of publication[5] and is set at 1 for the year 1972, increasing by one unit with each subsequent year.

Table 3 displays a comparison of the results derived from estimating the reference model against those obtained when the *Time trend* variable is included as an additional explanatory variable. The coefficients on the *Time trend* variable are relatively small and are not statistically significant in either the unweighted sample or weighted sample regression and are both positive. The coefficients on the other explanatory variables are little changed by the inclusion of a *Time trend* variable and the pattern of which variables are statistically significant does not change at all, with the exception of the lab variable in the weighted models. Any changes in either the R-squared or root mean-squared error (RMSE) statistics produced by the addition of the *Time trend* variable to the regression equation are sufficiently small as to leave the rounded values reported in Table 3 unchanged. In short, the addition of a *Time trend* variable adds nothing to improve either the explanatory or predictive powers of the reference regression specification.

The results of this latest meta-analysis are broadly consistent with the findings of previous meta-analyses with respect to the pattern of coefficient signs. The regression equations we estimate explain, in the best of circumstances, only about 23% of the overall variance we observe in BRs. ($R^2 = 0.2329$ for the weighted regression and $R^2 = 0.2349$ for the unweighted regression.) Notable, too, is the low predictive power of these regressions, as evidenced by the relatively large RMSEs in both the

[5] We choose to use year of publication to assign time trend values to each study rather than the year in which the underlying research took place. The choice is driven by the availability of year of publication data for each of the studies included in our sample. Commonly, these studies also describe when the underlying survey or experiment was conducted, but this information is not provided in all cases.

Table 3 Time trend regression coefficients

Variables	(1) ln (BR) Unweighted Non-use	(2) ln (BR) Weighted Non-use	(3) ln (BR) Unweighted Non-use Time	(4) ln (BR) Weighted Non-use Time
Certaintycorrection	−0.8907***	−0.8845***	−0.8989***	−0.9045***
	(0.10)	(0.15)	(0.10)	(0.15)
Cheaptalk	−0.3627**	−0.4547**	−0.3795**	−0.4919**
	(0.17)	(0.22)	(0.17)	(0.22)
Same	−0.1069	0.0175	−0.1049	0.0226
	(0.20)	(0.15)	(0.20)	(0.14)
Conjoint	−0.0331	−0.0748	−0.0605	−0.1359
	(0.11)	(0.19)	(0.14)	(0.20)
Student	0.1070	0.1879	0.1028	0.1745
	(0.23)	(0.21)	(0.24)	(0.21)
Lab	−0.1564	−0.3133*	−0.1440	−0.2925
	(0.21)	(0.19)	(0.22)	(0.19)
Private	0.5205***	0.6246***	0.5261***	0.6418***
	(0.14)	(0.16)	(0.13)	(0.16)
Familiargood	−0.0871	−0.2716*	−0.0835	−0.2563
	(0.14)	(0.15)	(0.14)	(0.16)
Non-use	0.3716*	0.4503**	0.3835**	0.4708**
	(0.19)	(0.18)	(0.19)	(0.18)
Time_trend			0.0033	0.0066
			(0.01)	(0.01)
Constant	0.3447*	0.4379**	0.2407	0.2349
	(0.20)	(0.19)	(0.36)	(0.38)
Observations	432	432	432	432
R-squared	0.2349	0.2329	0.2355	0.2353
RMSE	0.706	0.726	0.706	0.726
Degrees of freedom	422	422	421	421

Note: Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

weighted and unweighted sample regressions, 0.726 and 0.706, respectively. This represents the standard error of prediction, a measure of the precision with which the actual value of the dependent variable, ln (BR), can be predicted by the regression line. Even the smaller of these figures indicates that 95% of the observations of ln (BR) should fall within an interval of plus or minus 1.38 logarithmic units from their values as predicted by the regression line. Evaluated at their sample means (ln (BR) = 0.452 for the unweighted sample and ln (BR) = 0.616 for the weighted sample) and

after having been converted from log form into levels, these relatively large RMSEs establish a 95% confidence interval of prediction for BR ranging from 0.394 to 6.27 for the unweighted sample regression and between 0.446 and 7.68 for the weighted sample equation.

These wide ranges clearly indicate that the reference model cannot be used to provide a precise prediction of the bias ratio associated with any particular set of study characteristics. The reference model is thus unsuitable as a tool to offset the hypothetical bias that is inherent in all valuation exercises that attempt to value natural resources on the basis of survey respondents' answers to hypothetical questions.

Fixed Effects Regression

As a final robustness check, and to further explore the usefulness of the regression model in making predictions of the bias ratio associated with any particular survey, we re-estimated the reference model in a version that assigned study-specific fixed effect variables to 76 of the 77 studies from which we obtained our data. Table 4 provides results derived from the fixed effects specification alongside results from the reference model.

A fixed effects regression relies solely on variation within group effects in determining the coefficients to be placed on the other explanatory variables. This characteristic has several implications for our model. First, it means that studies that supply only one observation to the dataset are effectively ignored in estimating the other coefficients of our models. Second, we cannot estimate coefficients for variables that are held constant within each and every paper in which they appear. As a result, we cannot simultaneously estimate fixed effects coefficients and also estimate coefficients for the private, familiar good, and non-use variables.

The regression coefficients associated with five of the six indicator variables that are common to both the fixed effects and reference regressions are generally larger in magnitude and more likely to achieve statistical significance in the fixed effects specifications than is the case for their reference specification counterparts. This general observation is particularly evident when evaluating the coefficients on the potentially offsetting student and lab variables. The exception is the certainty correction variable, which drops in magnitude between the reference and fixed effects regressions, while remaining highly statistically significant.

Comparing the key regression statistics generated by the fixed effects specifications to those produced by the reference model, it is apparent that the fixed effects specifications do a better job than do their reference model counterparts in explaining the data. The fixed effects regressions generate

Table 4 Comparison between reference model and fixed effects regression coefficients

Variables	(1) ln (BR) Unweighted Non-use	(2) ln (BR) Weighted Non-use	(3) ln (BR) Unweighted Non-use F.E.	(4) ln (BR) Weighted Non-use F.E.
Certaintycorrection	−0.8907***	−0.8845***	−0.6332***	−0.6767***
	(0.10)	(0.15)	(0.10)	(0.10)
Cheaptalk	−0.3627**	−0.4547**	−0.5189***	−0.5160***
	(0.17)	(0.22)	(0.11)	(0.14)
Same	−0.1069	0.0175	0.1907*	0.2457
	(0.20)	(0.15)	(0.10)	(0.18)
Conjoint	−0.0331	−0.0748	0.0702	0.1387
	(0.11)	(0.19)	(0.04)	(0.10)
Student	0.1070	0.1879	0.2836*	0.3098*
	(0.23)	(0.21)	(0.17)	(0.17)
Lab	−0.1564	−0.3133*	−0.4132**	−0.4694**
	(0.21)	(0.19)	(0.19)	(0.22)
Private	0.5205***	0.6246***	*NA*	*NA*
	(0.14)	(0.16)		
Familiargood	−0.0871	−0.2716*	*NA*	*NA*
	(0.14)	(0.15)		
Non-use	0.3716*	0.4503**	*NA*	*NA*
	(0.19)	(0.18)		
Constant	0.3447*	0.4379**	0.6085***	0.7027***
	(0.20)	(0.19)	(0.02)	(0.02)
Observations	432	432	432	432
R-squared	0.2349	0.2329	0.6538	0.7532
RMSE	0.706	0.726	0.473	0.411
Degrees of freedom	422	422	347	347
Number of CV			78	78

Note: Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

considerably higher R-squared statistics and lower RMSE statistics than those derived from the estimation of their corresponding reference model equations. This is to be expected, as incorporating a separate fixed effect for every study should improve the goodness of fit of a regression, but the extent of the improvements further illustrates how little of the underlying variation in the data can be attributed to readily observable study characteristics. Our results offer little hope for any efforts to develop "correction

factors" or other tools that would be needed to offset hypothetical bias in any particular instance.

CONCLUSIONS

This study considers whether economists have yet developed any practical and reliable ways to correct for or overcome the well-known phenomenon of hypothetical bias found in survey-based attempts to value environmental or other goods. It does so by updating and extending work done in prior meta-analyses of stated preference methods that has confirmed the widespread presence of hypothetical bias in stated preference studies and that has associated certain factors in survey design with higher or lesser degrees of observable bias. Our meta-analysis, like prior meta-analyses on the same topic, offers no definitive insights that can be used to eliminate or reduce hypothetical bias. While we find some, but generally weak, associations between the presence of a limited number of survey design characteristics and the degree of hypothetical bias likely to be exhibited in particular types of survey treatments, any insights provided by our analysis cannot reliably be used to control for or eliminate the degree of bias likely to be found in any particular survey as the regression coefficients produced by our analyses are typically associated with relatively wide standard errors and the equations can explain only a small fraction of the variance exhibited in the degree of hypothetical bias observed across various studies.

BIBLIOGRAPHY

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**(3), 275–92.*

Barrage, L. and M.S. Lee (2010), "A penny for your thoughts: Inducing truth-telling in stated preference elicitation," *Economics Letters*, **106**(2), 140–42.*

Bishop, R.C. and T.A. Heberlein (1979), "Measuring values of extramarket goods: Are indirect measures biased?," *American Journal of Agricultural Economics*, **61**(5), 926–30.*

Blumenschein, K., M. Johannesson, K.K. Yokoyama, 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**(3), 441–57.*

Blumenschein, K., G.C. Blomquist, M. Johannesson, N. Horn, and P. Freeman (2008), "Eliciting willingness to pay without bias: Evidence from a field experiment," *The Economic Journal*, **118**(525), 114–37.*

Blumenschein, K., M. Johannesson, G.C. Blomquist, B. Liljas, and R.M. O'Conor

(1997), "Hypothetical versus real payments in Vickrey auctions," *Economics Letters*, **56**(2), 177–80.*
Bohm, P. (1972), "Estimating demand for public goods: An experiment," *European Economic Review*, **3**(2), 111–30.*
Botelho, A. and L.C. Costa Pinto (2002), "Hypothetical, real, and predicted real willingness to pay in open-ended surveys: Experimental results," *Applied Economics Letters*, **9**(15), 993–6.*
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 Meeting of the Western Committee on the Benefits and Costs of Natural Resource Planning and the Western Regional Science Association, San Diego, CA.*
Broadbent, C.D. (2014), "Evaluating mitigation and calibration techniques for hypothetical bias in choice experiments," *Journal of Environmental Planning and Management*, **57**(12), 1831–48.*
Brookshire, D.S. and D.L. Coursey (1987), "Measuring the value of a public good: An empirical comparison of elicitation procedures," *American Journal of Agricultural Economics*, **77**(4), 544–66.*
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 & Organization*, **43**(1), 127–39.*
Brown, T.C., I. Ajzen, and D. Hrubes (2003), "Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation," *Journal of Environmental Economics and Management*, **46(**2), 353–61.*
Brown, T.C., P.A. Champ, R.C. Bishop, and D.W. McCollum (1996), "Which response format reveals the truth about donations to a public good?," *Land Economics*, **72**(2), 152–66.*
Brynes, B., C. Jones, and S. Goodman (1999), "Contingent valuation and real economic commitments: Evidence from electric utility green pricing programmes," *Journal of Environmental Planning and Management*, **42**(2), 149–66.*
Camacho-Cuena, E., A. García-Gallego, H. Georgantzís, and G. Sabater-Grande (2003), "An experimental test of response consistency in contingent valuation," *Ecological Economics*, **47**(2–3), 167–82.*
Cameron, T.A., G.L. Poe, R.G. Ethier, and W.D. Schulze (2002), "Alternative nonmarket value-elicitation methods: Are the underlying preferences the same?," *Journal of Environmental Economics and Management*, **44**(3), 391–425.*
Carlson, J.L. (2000), "Hypothetical surveys versus real commitments: Further evidence," *Applied Economics Letters*, **7**(7), 447–50.*
Carlsson, F. and P. Martinsson (2001), "Willingness to pay for reduction in air pollution: A multilevel analysis," *Environmental Economics and Policy Studies*, **4**(1), 17–27.*
Carson, R.T. (2012), *Contingent Valuation: A Comprehensive Bibliography and History*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
Champ, P.A. and T.C. Brown (1997), *A Comparison of Contingent and Actual Voting Behavior*, Rocky Mountain Research Station – USDA Forest Service.*
Champ, P.A. and R.C. Bishop (2001), "Donation payment mechanisms and contingent valuation: An empirical study of hypothetical bias," *Environmental and Resource Economics*, **19**(4), 383–402.*
Champ, P.A., R. Moore, and R.C. Bishop (2009), "A comparison of approaches to

mitigate hypothetical bias," *Agricultural and Resource Economics Review*, **38**(2), 166–80.*

Champ, P.A., R.C. Bishop, T.C. Brown, and D.W. McCollum (1997), "Using donation mechanisms to value non-use benefits from public goods," *Journal of Environmental Economics and Management*, **33**(2), 151–62.*

Chang, J.B., J.L. Lusk, and F.B. Norwood (2009), "How closely do hypothetical surveys and laboratory experiments predict field behavior?," *American Journal of Agricultural Economics*, **91**(2), 518–34.*

Christie, M. (2007), "An examination of the disparity between hypothetical and actual willingness to pay using the contingent valuation methods: The case of red kite conservation in the United Kingdom," *Canadian Journal of Agricultural Economics*, **55**(2), 159–69.*

Cummings, R.G., G.W. Harrison, and E.E. Rutström (1995), "Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive-compatible?," *The American Economic Review*, **85**(1), 260–66.*

Cummings, R.G., S. Elliot, G.W. Harrison, and J. Murphy (1997), "Are hypothetical referenda incentive compatible?" *Journal of Political Economy*, **105**(3), 609–21.*

De Magistris, T., A. Gracia, and R.M. Nayga, Jr. (2011), "On the use of honesty priming task to mitigate hypothetical bias in choice experiments," *Centro de Investigación Y Tecnología Agroalimentaria De Aragón (CITA) Documento de Trabajo* No. 12/01.*

Duffield, J.W. and D.A. Patterson (1991), "Field testing existence values: An instream flow trust fund for Montana rivers," presented at the Association of Environmental and Resource Economics during the Valuing Environmental Goods with Contingent Valuation Session.*

Ehmke, M.D., J.L. Lusk, and J.A. List (2008), "Is hypothetical bias a universal phenomenon? A multinational investigation," *Land Economics*, **84**(3), 489–500.*

Ethier, R.G., G.L. Poe, W.D. Schulze, and J. Clark (2000), "A comparison of hypothetical phone and mail contingent valuation responses for green-pricing electricity programs," *Land Economics*, **76**(1), 54–67.*

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**(2), 123–38.*

Fox, J.A., J.F. Shogren, D.J. Hayes, and J.B. Kliebenstein (1998), "CVM-X: Calibrating contingent values with experimental auction market," *American Journal of Agricultural Economics*, **80**(3), 455–65.*

Frykblom, P. (1997), "Hypothetical question modes and real willingness to pay," *Journal of Environmental Economics and Management*, **34**(3), 275–87.*

Frykblom, P. (2000), "Willingness to pay and the choice of question format: Experimental results," *Applied Economics Letters*, **7**(10), 665–67.*

Getzner, M. (2000), "Hypothetical and real economic commitments, and social status, in valuing a species protection programme," *Journal of Environmental Planning and Management*, **43**(4), 541–59.*

Gregory, R. (1986), "Interpreting measures of economic loss: Evidence from contingent valuation and experimental studies," *Journal of Environmental Economics and Management*, **13**(4), 325–37.*

Griffin, C.C., J. Briscoe, B. Singh, R. Ramasubban, and R. Bhatia (1995), "Contingent valuation and actual behavior: Predicting connections to new

water systems in the State of Kerala, India," *World Bank Economic Review*, **9**(3), 373–95.*

Harrison, G.W. and E.E. Rutström (2008), "Experimental evidence on the existence of hypothetical bias in value elicitation methods," in C. Plott (ed.), *Handbook of Experimental Economics Results Volume 1*, Amsterdam: North-Holland, pp. 752–67.

Heberlein, T.A. and R.C. Bishop (1986), "Assessing the validity of contingent valuation," *The Science of the Total Environment*, **56**(1), 99–107.*

Hofler, R. and J.A. List (2004), "Valuation on the frontier: Calibrating actual and hypothetical statements of value," *American Journal of Agricultural Economics*, **86**(1), 213–21.*

Johannesson, M. (1997), "Some further experimental results on hypothetical versus real willingness to pay," *Applied Economic Letters*, **4**(3), 535–6.*

Johannesson, M., B. Liljas, and P. Johansson (1998), "An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions," *Applied Economics*, **30**(5), 643–7.*

Johansson-Stenman, O. and H. Svedsäter (2008), "Measuring hypothetical bias in choice experiments: The importance of cognitive consistency," *The B.E. Journal of Economic Analysis and Policy*, **8**(1), Article 41, 1–8.*

Johnston, J. (2006), "Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum," *Journal of Environmental Economics and Management*, **52**(1), 469–81.*

Kealy, M.J., J.F. Dovido, and M.L. Rockel (1988), "Accuracy in valuation is a matter of degree," *Land Economics*, **64**(2), 158–71.*

Landry, C.E. and J.A. List (2007), "Using ex ante approached to obtain credible signals for value in contingent markets: Evidence from the field," *American Journal of Agricultural Economics*, **89**(2), 420–29.*

List, J.A. (2003), "Using random nth price auctions to value non-market goods and services," *Journal of Regulatory Economics*, **23**(2), 193–205.*

List, J.A. and C.A. Gallet (2001), "What experimental protocol influence disparities between actual and hypothetical stated values?," *Environmental and Resource Economics*, **20**(3), 241–54.*

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**(2), 193–205.*

List, J., R.P. Berrens, A.K. Bohara, and J. Kerkvliet (2004), "Examining the role of social isolation on stated preferences," *American Economic Review*, **94**(3), 741–52.*

Little, J. and R. Berrens (2004), "Explaining disparities between actual and hypothetical stated values: Further investigation using meta-analysis," *Economics Bulletin*, **3**(6), 1–13.

Little, J., C.D. Broadbent, and R.P. Berrens (2012), "Meta-analysis of the probability of disparity between actual and hypothetical valuation responses: Extension and preliminary new results," *Western Economics Forum*, **11**(1).

Loomis, J., P. Bell, H. Cooney, and C. Asmus (2009), "A comparison of actual and hypothetical willingness to pay of parents and non-parents for protecting infant health: The case of nitrates in drinking water," *Journal of Agricultural and Applied Economics*, **41**(3), 697–712.*

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**(4), 450–61.*

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**(2), 109–23.*

Macmillan, D.C., T.S. Smart, and A.P. Thorburn (1999), "A field experiment involving cash and hypothetical charitable donations," *Environmental and Resource Economics*, **14**(3), 399–412.*

Miller, K.M., R. Hofstetter, H. Krohmer, and Z.H. Zhang (2011), "How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches," *Journal of Marketing Research*, **48**(1), 172–84.*

Morrison, M. and T.C. Brown (2009), "Testing the effectiveness of certainty scales, cheap talk, and dissonance minimization in reducing hypothetical bias in contingent valuation studies," *Environmental Resource Economics*, **44**(3), 307–26.*

Moser, R., R. Raffaelli, and S. Notaro (2014), "Testing hypothetical bias with real choice experiment using respondents' own money," *European Review of Agricultural Economics*, **41**(1), 25–46.*

Murphy, J.J., T. Stevens, and D. Weatherhead (2002), "An empirical study of hypothetical bias in voluntary contribution contingent valuation: Does cheap talk matter?," Paper prepared for the World Congress of Environmental and Resource Economists, No. 789.000.*

Murphy, J.J., P.G. Allen, T.H. Stevens, and D. Weatherhead (2005), "A meta-analysis of hypothetical bias in stated preference valuation," *Environmental and Resource Economics*, **30**(3), 313–25.

Navrud, S. (1992), "Willingness to pay for preservation of species – an experiment with actual payments," in S. Navrud (ed.), *Pricing the European Environment*, Oslo: Scandinavian University Press, pp. 231–46.*

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**(2), 145–54.*

Norwood, B.F. (2005), "Can calibration reconcile stated and observed preferences?," *Journal of Agricultural and Applied Economics*, **37**(1), 237–48.*

Paradiso, M. and A. Trisorio (2001), "The effect of knowledge on the disparity between hypothetical and real willingness to pay," *Applied Economics*, **33**(11), 1359–64.*

Park, J.H. and D.L. MacLachlan (2008), "Estimating willingness to pay with exaggeration bias – corrected contingent valuation method," *Marketing Science*, **27**(4), 691–8.*

Poe, G., 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**, 105–31.*

Schläpfer, F. and B. Fischhoff (2010), "When are preferences consistent? The effects of task familiarity and contextual cues on revealed and stated preferences?," *Working Paper No. 1007*, Socioeconomic Institute, University of Zurich.

Seip, K. and J. Strand (1992), "Willingness to pay for environmental goods in Norway: A contingent valuation study with real payment," *Environmental and Resource Economics*, **2**(1), 91–106.*

Shechter, M., B. Reiser, and N. Zaitsev (1998), "Measuring passive use value:

Pledges, donations, and CV responses in connection with an important natural resource," *Environmental and Resource Economics*, **12**(4), 457–78.*
Silva, A., R.M. Nayaga, B.L. Campbell, and J.L. Park (2007), "On the use of valuation mechanisms to measure consumers' willingness to pay for novel products: A comparison of hypothetical and non-hypothetical values," *International Food and Agribusiness Management Review*, **10**(2), 165–80.*
Silva, A., R.M. Nayga, B.L. Campbell, and J. Park (2012), "Can perceived task complexity influence cheap talk's effectiveness in reducing hypothetical bias in stated choice studies?," *Applied Economics Letters*, **19**(17), 1711–44.*
Sinden, J.A. (1988), "Empirical tests of hypothetical bias in consumer surplus surveys," *Australian Journal of Agricultural Economics*, **32**(2–3), 98–112.*
Spencer, M.A., S.K. Swallow, and C.L. Miller (1998), "Valuing water quality monitoring: A contingent valuation experiment involving hypothetical and real payments," *Agricultural and Resource Economics Review*, **27**(1), 28–42.*
Taylor, L. (1998), "Incentive compatible referenda and the valuation of environmental goods," *Agriculture and Resource Economics Review*, **27**(2), 132–9.*
Veisten, K. and S. Navrud (2006), "Contingent valuation and actual payment for voluntarily provided passive-use values: Assessing the effect of an induced truth-telling mechanism and elicitation formats," *Applied Economics*, **38**(7), 735–56.*
Volinsky, D., W. Adamowicz, and M. Veeman (2011), "Predicting versus testing: A conditional cross-forecasting accuracy measure for hypothetical bias," *The Australian Journal of Agricultural and Resource Economics*, **55**(3), 429–50.*
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*, **45**(3), 631–49.*
Vossler, C.A., R.G. Ethier, G.L. Poe, and M.P. Welsh (2003), "Payment certainty in discrete choice contingent valuation responses: Results from a field validity test," *Southern Economic Journal*, **69**(4), 886–902.*
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 & Competition*, **51**(2), 261–77.*
Willis, K.G. and N.A. Powe (1998), "Contingent valuation and real economic commitments: A private good experiment," *Journal of Environmental Planning and Management*, **41**(5), 611–19.*
Wu, P. and C. Huang (2001), "Actual advertising expenditure versus stated willingness to pay," *Applied Economics*, **33**(4), 277–83.