Journal of Environmental Economics and Management **43,** 219–233 (2002)
doi:10.1006/jeem.2000.1182, available online at http://www.idealibrary.com on IDEAL®

Calibration of Willingness-to-Accept[1]

John A. List[2]

Department of Agricultural and Resource Economics, University of Maryland, College Park, Maryland 20742-5535

and

Jason F. Shogren

Department of Economics and Finance, University of Wyoming, Laramie, Wyoming 82071

Received November 17, 2000; revised July 24, 2000; published online August 21, 2001

This paper calibrates real and hypothetical willingness-to-accept estimates elicited for consumer goods in a multi-unit, random *n*th-price auction. Using a within-subject experimental design, we find that people understated their real willingness to accept in the hypothetical regimes, framed both as demand and non-demand revealing. After controlling for person-specific effects, however, hypothetical and real statements are equivalent on the margin.

1. INTRODUCTION

Some evidence suggests that people tend to overstate their real willingness-to-pay (WTP) in hypothetical markets.[3] This observation triggered a search for a calibration function to correct systematic bias between intentions and actions in valuation exercises.[4] But much less attention has been paid to the relationship between real and hypothetical willingness-to-accept (WTA) compensation, and little is known about the in-sample calibration of WTA offers.[5] This is surprising since the recent hypothetical valuation literature has been driven by compensatory natural resource

[1] We thank Kerry Smith for his excellent comments and diligence in working with us to improve the manuscript. Carol Mansfield, John Horowitz, an Associate Editor, and an anonymous referee also provided very helpful comments. Seminar participants at the *World Congress of Environmental and Resource Economists* in Venice, Italy also made useful suggestions. The University of Central Florida's STAR program, NOW, the Dutch Science Foundation, and the National Science Foundation provided financial support. Apinya Thumaphipol and Priti Manek ably assisted during the experimental sessions.

[2] Address correspondence to: John A. List, Economics Building, Department of Agricultural and Resource Economics, University of Arizona, Tucson, AZ 85721-0023; E-mail: JList@ag.arizona.edu.

[3] See for example, Bohm [8], Bishop and Heberlein [3], Dickie *et al*. [18], Seip and Strand [50], Neill *et al*. [44], Fox *et al*. [22], List and Shogren [37].

[4] See Blackburn *et al*. [7], Fox *et al*. [22], List and Shogren [37], Mansfield [41], and the National Oceanic and Atmospheric Administration [45, 46]. Also see Randall's [47] critique of calibration.

[5] Between-sample comparisons of hypothetical and real WTA offers include, amongst others, Bishop and Heberlein [3], Bishop *et al*. [6], Brookshire and Coursey [10], and Smith and Mansfield [52]. We are unaware of any studies that use within-sample data to calibrate hypothetical and real WTA offers. Although there is no evidence that indicates within-sample procedures are superior to between-sample, in a multi-unit context within-sample tests have a unique advantage in that they can control for individual-specific effects.

0095-0696/01 $35.00

damage assessment, which is closely tied to WTA measures of value (e.g., Exxon Valdez and the Prince William Sound).

This paper uses panel data from a lab valuation experiment to calibrate in-sample real and hypothetical WTA compensation to surrender holiday gifts. Our specific goal is to observe patterns in WTA calibration for market goods with intangible qualities to further efforts aimed at finding generalizations about behavior that can eventually convert experimentation into theory. In this regard, we address two questions: (1) do hypothetical WTA offers differ across demand revealing (random nth price auction) and non-demand revealing (open-ended question) elicitation schemes? No—hypothetical offers were unaffected by the framing of choice. This finding provides further support that differences found in hypothetical valuation exercises cannot be explained away with the argument that people find it difficult to answer open-ended questions; and (2) do hypothetical and real WTA offers differ? Yes—estimates suggest that subjects *understated* their real WTA in both hypothetical scenarios; the ratio of mean real to mean hypothetical is about 1.5. After controlling for subject-specific effects, however, we find that the marginal calibration factor decreases to approximately 1.05, which is not significantly different from unity. Hence, on the margin, hypothetical and real statements are equivalent: a one-dollar increase in a hypothetical valuation statement is associated with a one-dollar increase in actual value.

The remainder of the paper proceeds as follows. Section 2 presents a broad overview of the existing experimental work that includes hypothetical and actual valuation statements. Section 3 includes a description of the data, hypotheses, and empirical methods. Section 4 discusses the empirical results. Section 5 concludes.

2. PREVIOUS WORK

We first set the stage with a brief poll of the key studies in the calibration literature. Table I summarizes the observed relationship between real and hypothetical values.[6] The top panel of Table I presents a chronological ordering of WTP studies that report both hypothetical and actual statements. This line of research began with Bohm's [8] seminal experimental lab study which compared bids in hypothetical and actual experimental markets that elicited subjects' stated value to preview a Swedish television show. His results suggest people moderately overstate their actual values when asked a hypothetical question. Subsequent lab research has generally supported Bohm's findings (e.g., Bishop and Heberlein [3], Seip and Strand [50], Neill *et al*. [44], Frykblom [24], Fox *et al*. [22], List and Shogren [37, 38], and Balistreri *et al*. [2].

Exceptions to this set of results are not difficult to find (e.g., Dickie *et al*. [18], Sinden [51]); but taken as a whole the evidence suggests that the average person seems to exaggerate his or her actual WTP across a broad spectrum of goods with vastly different experimental parameters. For instance, the ratio of hypothetical-to-actual overbidding, which ranged from 2.2 to 3.5 for baseball cards, falls between the calibration factors observed for irradiated/non-irradiated pork and water color paintings and maps (see List and Shogren [37], Fox *et al*. [22], and Neill *et al*. [44] in Table I). These results are within the range of calibration factors, 1.0 to 10.0, observed in earlier work (see Diamond and Hausman's overview [17]), and rein-

[6] See Foster *et al*. [21] for a non-experimental comparison of real and hypothetical WTP statements. We do not include "cheap-talk" approaches in Table I (e.g., Cummings and Taylor [16], List [34]).

force the argument that people tend to overstate their actual WTP when confronted with hypothetical questions. Given that the received literature derives value estimates using quite heterogeneous experimental methods, understanding the relationship between real and hypothetical values is difficult. Nevertheless, data from the top panel of Table I suggest that attempts to bridge the gap statistically might be impossible since calibration factors appear to be good-specific.[7] This finding is supported in related work on the relationship between WTA and WTP (Horowitz and McConnell [29]), and provides an indication that calibration functions are good-specific.

Researchers have spent considerably less energy on understanding the relationship between WTA measures of value. The bottom panel of Table I indicates that experimental evidence from this relatively small lot of studies is mixed. Bishop and associates found that Wisconsin goose hunters' overstated their actual WTA to sell goose-licenses; deer hunters' in a sealed-bid auction understated their actual WTA to sell deer-permits, while hunters in a dichotomous choice institution overstated their real WTA. Coursey *et al*. [13] found that people overstated their actual WTA to experience a drop of the bitter-tasting sucrose octa-acetate; Smith and Mansfield's [52] field results suggest that real and hypothetical WTA statements for the opportunity to spend time in a second set of interviews on an undisclosed topic are statistically indistinguishable. Much like the WTP studies in the top panel of Table I, these results suggest that the real-hypothetical WTA gap might be good-specific.

3. DATA, HYPOTHESES, AND EMPIRICAL METHODS

To provide initial insights into the real/hypothetical WTA relationship from a within-sample design, we use data from List and Shogren's [38] three-stage, multi-good auction administered at the University of Central Florida. Stage 1: *hypothetical open-ended survey*—subjects answered Waldfogel's [55] survey, in which they stated "the amount of cash such that you are indifferent between the gift and the cash" for each of their 1996 Christmas gifts, e.g., a person who received ten gifts submitted ten hypothetical WTA offers. Stage 2: *hypothetical auction*—a monitor asked subjects to again state their hypothetical selling price for each gift in the context of the random nth-price auction.[8] Stage 3: *actual auction*—the monitor

[7] But, see the results in List *et al*. [35], which suggest that similar commodities may have calibration functions that are not statistically different.

[8] After the instructions were read, the monitor ran a candy bar pre-auction to give the subjects some experience with the random nth price auction (see, e.g., Melton *et al*. [42]). The auction works as follows: (i) for each gift received, a subject states his or her selling price; (ii) all gifts from each subject are pooled to create the set of total available gifts; (iii) all gifts are rank-ordered from lowest to highest selling price; (iv) the monitor selects a random number uniformly distributed between 2 and 21 (the most gifts received by a subject); and (v) the monitor purchases the $(n - 1)$ lowest priced gifts and pays the nth lowest price for each gift. We should note that a multi-good, uniform price auction does not necessarily inherit the same demand revealing properties as a second-price or nth price auction in which subjects can only sell one gift (Vickrey [54]). For example, truth-telling can be a non-unique Nash equilibrium in a multiple-unit, uniform price auction, but in general this may not be the case (e.g., Forsythe and Isaac, [20])—there is an incentive for "demand reduction" on units after the first (see Miller and Plott [43], Franciosi *et al*. [23]). Although List and Lucking-Reilly [36] found significant demand reduction in two-unit, two-person uniform-price auctions for high valued goods ($40–$70), demand reduction essentially disappeared when the number of bidders increased to 5 (Engelbrecht-Wiggans *et al*. [19]). We therefore do not concern ourselves with demand reduction since we have 36 bidders.

TABLE I

Summary of Studies Comparing Real and Hypothetical Statements

Study	Year	Type of experiment	Type of good	Type of comparison	Type of elicitation	Calibration factor (hypothetical/actual)
WILLINGNESS-TO-PAY						
Bohm	1972	Laboratory	Private	Within group	Open-ended	Part V with Part II: 10 Part V with Part I: 1.16 Part VIA with Part II: 1.16 Part VIA with Part I: 1.34
Bishop and Heberlein	1979	Field	Private	Between group	Dichotomous choice	0.3–1.6
Bishop and Heberlein	1986	Field	Private	Between group	Open-ended/sealed bid auction dichotomous choice	1.3–2.3 0.8
Samples *et al.*	1986	Non-experimental	Public	Between group	Open-ended	Donations: 1.6 + tax check-offs: 32–300
Brookshire end Coursey	1987	Field and lab	Public	Between	Smith auction/hypothetical in field Actual in laboratory	25 trees: 2 50 trees: 1.85
Dickie *et al.*	1987	Field	Private	Between group	Dichotomous choice	na/but approximately = 1
Coursey *et al.*	1987	Laboratory	Private	Between	Open-ended hypothetical/ Vickrey auction actual	1
Sinden	1988	Laboratory	Public	Within group	Open-ended	0.8–1.5
Kealy *et al.*	1988	Laboratory	Private Public	Between group Within group	Dichotomous choice Dichotomous choice	1.4 1.4
Kealy *et al.*	1990	Laboratory	Private Public	Between group Within group	Dichotomous choice Dichotomous choice	1.3 1.0–2.0
Seip & Strand	1990	Field	Private/ public	Within group	Dichotomous choice	10.3
Brookshire *et al.*	1990	Laboratory	Private	Between	Smith auction	2.7
Navrud	1992	Field	Private Public	Within group	Dichotomous choice	3.2 adjusted 1.6–2.1
Irwin *et al.*	1992	Laboratory	Private	Between	Vickrey auction	1st round: 1.0 all rounds: 2.5

WILLINGNESS-TO-PAY						
Boyce *et al.*	1992	Laboratory and field	Public	Between	Becker, DeGroot, Marschak	No-kill: 1.5 Kill: 2.1 Field-kill: 0.9
McClelland *et al.*	1993	Laboratory	Private	Between	Vickrey auction	1% chance: 2.2 90% chance: 0.8
Neill *et al.*	1994	Laboratory	Private	Between group	Open-ended	3.1–25.1
Cummings *et al.*	1995	Laboratory	Private	Between group	Dichotomous choice	2.6–5.3
				Within group	Dichotomous choice	2.6–10.5
Loomis *et al.*	1996	Laboratory	Private	Between group	Open-ended	1.8–2.9
				Within group	Open-ended	2.0–3.6
Brown *et al.*	1996	Field	Public	Between group	Open-ended	4.1
					Dichotomous choice	6.5
Frykblom	1997		Private			
Spencer *et al.*	1998	Laboratory	Public	Between	Provision point mechanism	Pond A 4.67 Pond B 4.66
List and Shogren	1998	Field	Private	Within	Vickrey auction	1 card: 2.54 10 card: 3.47 Dealers: 2.19
Fox *et al.*	1998	Laboratory	Private	Between/within	Vickrey auction	1.2 1.5
Balistreri *et al.*	1998	Laboratory	Private	Between	Open-ended/dichotomous choice for hypothetical English action for actual	OE: 1.25 DC: 1.54 (no restrictions) .58 (WTP restricted as positive)
COMPENSATION DEMANDED						
Bishop *et al.*	1983	Field	Private	Between group	Dichotomous choice	Goose: 1.6
Coursey *et al.*	1987	Laboratory	Private	Between	Open-ended hypothetical/ Vickrey auction actual	2
Brookshire and Coursey	1987	Field and lab	Public	Between	Smith auction/hypothetical in field Actual in laboratory	25 trees: 28.2 50 trees: 25.79
Bishop and Heberlein	1990	Field	Private	Between group	Sealed bid auction Dichotomous	Deer: 0.70 Deer: 2.74
Smith and Mansfield	1998	Field	Private	Between	Dichotomous choice	na/but approximately = 1

ran the actual random nth-price auction, in which each subject submitted his or her real selling price for each gift. All subjects knew they might potentially have to sell the gift to the monitor if they made one of the nth lowest offers.[9] There were no announcements of experimental Stages 2 or 3 until the actual day they were run; hence subjects were unaware that a real auction would occur after the first two stages, mitigating any strategic effects. After accounting for attrition, our final sample consisted of 244 gifts across 36 recipients. The gifts that were offered were quite heterogeneous, and included automobiles, perfumes, a beer intake facilitator, and various pets including a dog, horse, and an iguana. The most common gift types were clothing, compact disc players, compact discs, and jewelry.

We use these WTA Christmas gift data to examine three questions related narrowly to non-market valuation and broadly to rational choice in the lab. First, does the institutional frame matter for hypothetical WTA offers? Previous lab work examining WTP data has found mixed evidence regarding whether the frame matters to hypothetical bidders (e.g., Neill *et al*. [44]; Frykblom [24], and Balistreri *et al*. [2]). Also, the limited work that compares hypothetical and actual WTA statements has used between-sample designs and has yielded largely mixed results (see, e.g., Bishop and Heberlein [5] and the cites therein). We address this *framing hypothesis* by using our within-sample data to compare mean hypothetical WTA offers: H_0^{HI}: $\mathrm{WTA}^{\mathrm{HOE}} = \mathrm{WTA}^{\mathrm{HR}}$, where HOE and HR are the hypothetical open-ended and hypothetical random nth offers.

Second, do actual and hypothetical WTA offers differ? As noted previously, the few out-of-sample WTA studies provided mixed results. We use two hypotheses to address the potential hypothetical-real WTA gap: (1) the *unconditional calibration hypothesis* considers whether mean real and mean hypothetical WTA offers differ, H_0^{ROE}: $\mathrm{WTA}^{\mathrm{RR}} = \mathrm{WTA}^{\mathrm{HOE}}$ and H_0^{RHR}: $\mathrm{WTA}^{\mathrm{RR}} = \mathrm{WTA}^{\mathrm{HR}}$, where RR denotes real random nth-price auction;[10] and (2) the *no-bias calibration hypothesis* examines whether hypothetical offers are consistent with real economic commitments. Since we have panel data—multiple offers across numerous subjects—our regression approach to test the *no-bias calibration hypothesis* controls for unobserved individual attributes and data dependencies that cannot be accounted for in tests of sample means or pooled ordinary least-squares regression models. Given that subjects received an unequal number of Christmas gifts, we use an unbalanced panel data model to estimate the WTA calibration functions:

$$A_{ig} = \alpha_i + \beta H_{ig} + \varepsilon_{ig}, \qquad i = 1, 2, \ldots, N;\ g = 1, 2, \ldots, G, \tag{1}$$

where A_{ig} represents stated WTA from the real random nth-price auction for the ith subject's gth Christmas gift, H_{ig} is the hypothetical open-ended or hypothetical random nth-price auction WTA statement for the ith subject's gth gift, α_i are estimated fixed or random effects, and ε_{ig} is the well-behaved error term. The *no-bias calibration hypothesis*, using both sets of hypothetical data as the independent variable, is $H_0^{\text{No-bias}}$: $\alpha_i = 0$; $\beta = 1$, where α_i and β are defined above. Rejecting the *no-bias* hypothesis suggests that hypothetical WTA offers are inconsistent with real economic commitments.

[9] The random price was #4, and the monitor made arrangements with the sellers to purchase the three lowest valued gifts each priced at the fourth lowest value.

[10] An unconditional test holds nothing constant across subjects or gifts.

We estimate Eq. (1) using pooled ordinary least-squares (OLS), fixed effects panel data, and random effects panel data models. Both fixed and random effects models control for unmeasured heterogeneity that pooled OLS ignores. Random effects estimates of Eq. (1) yield coefficients that are not conditioned on unmeasured person effects, whereas fixed effects estimates yield coefficients conditioned on the unmeasured characteristics. Fixed effects estimates are inefficient since they only consider within-person variation. Yet, if the person effects are correlated with hypothetical WTA responses, random effects estimates are inconsistent, while the within estimator remains unbiased and consistent. We test for consistency using a Hausman [26] test when comparing estimates from fixed and random effects models.

Our third hypothesis considers the nature of calibration functions across commodities. Previous work suggests statistical bias depends on the characteristics of the good (Fox *et al*. [22], List and Shogren [37]). Extending Eq. (1), our *good-neutral hypothesis*, H_0: $\beta_i = \beta_j$ $(\forall j)$, tests whether goods have heterogeneous calibration functions. We test this hypothesis by clustering goods into three categories based on estimated market value. Two related points justify this approach: (i) Table I suggests that intangible qualities of a good matter—the more distant the good is from an ordinary market good, the larger the calibration factor; and (ii) the good's value seems to matter as well, as suggested by a recent comparison of WTP and WTA for goods of different values (Horowitz and McConnell [29]).

We construct a pragmatic, ad hoc decision rule to split the data into three groups —*low*- (stated market value < \$50), *medium*- (\$50 ≤ value ≤ \$150), and *high*- (value > \$150) valued gifts. Our logic in constructing the three groups proceeded as follows: we wish to split the sample by the intangible nature of each good as suggested by past work; and since each subject also provided a sentimental component of each gift, we focus on sentimental value since goods with more intangible qualities tend to have larger sentimental value. In addition, since stated sentimental value is highly correlated with stated market value, each gift was grouped by the subjects' stated market value. This procedure allows us to run an *ex post* validity test in which we compare these categories with posted prices from a *JC Penney* 1997 catalogue. For those goods that are listed in the catalogue, the validity test indicates that *ex ante* and *ex post* clusters yield similar groupings.

We test the good-neutral hypothesis by estimating

$$A_{ig} = \alpha_i + \beta_1 H_{ig} + \beta_2 D_L H_{ig} + \beta_3 D_M H_{ig} + \varepsilon_{ig}, \quad i = 1, 2, \ldots, N;\ g = 1, 2, \ldots, G, \quad (2)$$

where A_{ig}, H_{ig}, α_i, and ε_{ig} are defined above, and D_L (D_M) is a dichotomous variable that equals 1 if the good is in the low- (high-) valued category, 0 otherwise. Equation (2) uses the high-valued goods as the baseline group and therefore the heterogeneous calibration function hypothesis becomes: H_0: $\beta_2 = \beta_3 = 0$, which can be tested via an F test.[11]

[11] To preserve degrees of freedom and focus on slope changes, we restrict the subject-specific effects, α_i, to be equal across the three categories of goods.

4. RESULTS

Table II presents summary statistics, Wilcoxon signed-rank tests (for matched pairs) of the equivalency of the WTA distributions, and unconditional estimates of the calibration function for each of the three valuation methods. Mean offers in the overall sample range from \$95.77 to \$136.87 per gift in the three designs. Mean

TABLE II
Summary Statistics and Non-Parametric Tests of Equivalency

	Hyp. survey	Hyp. random nth-price auction	Actual random nth-price auction	Hyp. auction vs. hyp. survey	Actual auction vs. hyp. auction	Actual auction vs. hyp. survey
	Mean	Mean	Mean	Wilcoxon test	Wilcoxon test	Wilcoxon test
Mean bid						
Overall	\$95.77	\$96.34	\$136.87	4752	9570[a]	7376[a]
$N = 244$	(\$12.28)	(\$12.48)	(\$15.36)	$z = -0.66$	$z = -5.76$	$z = -5.62$
Low-valued goods	\$28.44	\$30.10	\$21.46	569	1429[a]	1722
$N = 109$	(\$2.47)	(\$2.90)	(\$1.13)	$z = -2.39$	$z = -2.77$	$z = -2.08$
Medium-valued goods	\$67.70	\$67.15	\$77.25	588	719[a]	696[a]
$N = 72$	(\$7.55)	(\$7.43)	(\$2.72)	$z = -0.25$	$z = -3.99$	$z = -3.74$
High-valued goods	\$244.33	\$244.28	\$404.67	252	11[a]	19[a]
$N = 63$	(\$41.20)	(\$42.08)	(\$44.04)	$z = -0.78$	$z = -6.02$	$z = -5.93$
Mean calibration factors						
Overall	—	—	—	1.01	1.42	1.43
Low-valued goods	—	—	—	1.06	0.71	0.75
Medium-valued goods	—	—	—	0.99	1.15	1.14
High-valued goods	—	—	—	1.00	1.66	1.66

Notes: (1) Standard errors in parentheses;

(2) Wilcoxon test is a signed-rank test for matched pairs across gifts. Since the number of paired observations is larger than 30, the large-sample z-test is used. The large sample z-test's null and alternative hypotheses are given by:

H_0: Two sampled populations have identical probability distributions.

H_a: The probability distribution for population A is shifted to the right or to the left of that for population B.

z values are computed as follows: $z = (W - (n(n+1)/4))/(n(n+1)(2n+1)/24)^{1/2}$. Where n is the number of non-tied differences between the two samples;

(3) Mean calibration factor ratios are calculated as the top value in the column header divided by the lower value in the column header. For example, mean bid ratio under the fifth column (labeled *actual auction versus hypothetical auction*), is computed as (mean actual auction/mean hypothetical auction).

[a] Significantly different values at the 1% level.

responses in the three value categories range from \$21.46–\$30.10, \$67.15–\$77.25, and \$244.28–\$404.67 per gift. We now analyze our hypotheses in turn.

4.1. *Framing Does Not Affect Hypothetical Values*

First, we cannot reject the *framing hypothesis*, H_0^{HI}: $\mathrm{WTA}^{\mathrm{HOE}} = \mathrm{WTA}^{\mathrm{HR}}$ (see Table II). Although mean offers differ, we cannot reject the hypothesis that stated values in the hypothetical survey were derived from the same parental population as values from the hypothetical auction (see Table II Wilcoxon test, $W = 4752$ ($z = -0.66$)). This finding is consistent across value categories and suggests the institutional frame does not universally affect hypothetical valuations. If it is meaningful to know which hypothetical mode best approximates real values, this result suggests the gap between intentions and actions cannot be explained away by arguing that people had difficulty answering an open-ended question. Other WTP studies find similar results (e.g., Neill *et al*. [44] and Frykblom [24]). In a much different context, Rutström [48] finds some behavioral differences between alternative incentive-compatible institutions for eliciting home-grown values.

4.2. *Distributions of Hypothetical and Real Offers are Different*

Second, we reject the *unconditional calibration hypothesis*—hypothetical and real WTA offers differ significantly, not controlling for subject effects. The bottom panel of Table II presents the mean WTA ratio across designs. In the overall sample, hypothetical offers **underestimate** real offers—the ratio of mean actual to mean hypothetical is about 1.4. These differences are statistically significant at the $p < 0.01$ level [$W = 9570$ ($z = -5.76$); $W = 7376$ ($z = -5.62$)], implying revealed values in the actual auction are not derived from the same parental population as revealed values in the two hypothetical treatments. In general, these results hold across commodities of different value—one exception is the calibration factors for low-valued goods, which are in the range of 0.75 and imply that hypothetical statements are greater than actual statements.

The calibration factors in Table II suggest that bias increases with the market value of the good. This finding is consonant with previous studies and suggests the further a good is from an ordinary private good, the greater is the tendency for subjects to mis-state their true value. Even so, the range of bias observed in Table II suggests that for most goods people **understated** their true WTA when asked a hypothetical WTA valuation question. Given that many practitioners view WTA measures as the upper bound on value for incremental changes in a good or service (e.g., Cummings *et al*. [14]), this finding may imply that hypothetical offers could actually represent a lower limit on this upper bound.

4.3. *Conditional Calibration Factors Do Not Depend on the Market Value of the Good*

Table III presents panel data regression estimates of Eq. (2) for ordinary least squares, and fixed and random effects models using both hypothetical WTA institutions as the independent variable. When comparing coefficient estimates it is important to note that Hausman [26] tests of the null hypothesis of zero correlation between hypothetical offers and the subject-effects indicate the orthogonality assumption underlying the random effects estimates is violated in both regression

TABLE III
Fixed Effects Estimation Results for Calibration Functions

	Open-Ended Survey			Hypothetical Auction		
Variable	OLS	Fixed effects	Random effects	OLS	Fixed effects	Random effects
Constant	91.3[a]	—	61.8[a]	94.7[a]	—	69.3[a]
	(7.5)		(3.5)	(7.8)		(4.1)
Hyp. offer	0.96[a]	1.08[a]	1.01[a]	0.95[a]	1.05[a]	0.99[a]
	(20.8)	(24.9)	(25.3)	(20.9)	(24.5)	(25.1)
Hyp. offer*D_L	−1.97[a]	−0.22	−1.03[a]	−2.14[a]	−0.46	−1.27[a]
(β_2)	(5.6)	(0.6)	(−3.2)	(5.9)	(1.3)	(4.0)
Hyp. offer*D_M	−1.02	−0.30	−0.68[a]	−1.03[a]	−0.32	−0.71[a]
(β_3)	(5.2)	(1.5)	(−3.9)	(5.2)	(1.7)	(5.2)
R^2	0.68	0.84	0.66	0.67	0.83	0.67
Adj. R^2	0.67	0.80	—	0.67	0.80	—
$F(\alpha_i = 0)$	—	5.7[a]	—	—	5.4[a]	—
(d.f.)		(35, 205)			(35, 205)	
LM($\alpha_i = 0$)	—	—	37.3[a]	—	—	35.4[a]
(d.f.)			(1)			(1)
Hausman	—	—	27.5[a]	—	—	27.0[a]
(d.f.)			(3)			(3)
N	244	244	244	244	244	244

Notes: (1) Dependent variable is actual offer in the random nth-price real auction;
(2) D_L indicates low-valued goods (value < \$50); D_M indicates medium-valued goods (\$50 ≤ value ≤ \$150). High-valued goods (value > \$150) are omitted from the equation and represent the baseline group;
(3) Absolute values of t statistics in parentheses under coefficient estimates.
[a] Significant at the 0.01 level.

models at conventional significance levels ($\chi^2(3) = 27.5$, and 27.0). This result suggests the error components model yields inconsistent and potentially biased coefficient estimates. Accordingly, we focus on estimates from the fixed-effects models.

Although the raw statistics in Table II imply calibration factors critically depend on the good's value, regression estimates in Table III paint a much different picture. Estimates in Table III suggest the slope coefficients on the two value categories (β_2 and β_3) are individually not different from zero at conventional significance levels in either model.[12] Further, we cannot jointly reject H_0: $\beta_2 = \beta_3 = 0$ at the 95% level ($F = 1.2$ open-ended; $F = 1.6$ hypothetical auction), suggesting the slope of the regression calibration function is **not different** across our three

[12] For the high-valued baseline category, we observe that the estimated coefficients on both hypothetical offers are significant at the $p < 0.01$ level. In the hypothetical WTA open-ended survey, the point estimate in column 2 implies that a \$1 increase in the hypothetical offer increases the real offer in the random nth-price auction by \$1.08, which is not significantly different from \$1 at the $p < 0.01$ level. We find a similar result in the hypothetical random nth-price auction fixed effects specification (column 5). Point estimates imply that a \$1 increase in the hypothetical random nth-price offer increases the real offer by \$1.05, which is not significantly different from \$1 at the $p < 0.01$ level.

categories of goods as clustered by market value. Thus, for this particular clustering rule, we cannot reject the *good-neutral hypothesis*.[13]

Although point estimates of the slope terms in each value category are not different from one another at conventional levels, F-tests presented in Table III reject the null hypothesis of homogeneity of unmeasured subject-specific effects at the $p < 0.01$ level ($F = 5.7$; 5.4). These results suggest that individual-specific intercepts are different from one another at conventional significance levels, leading us to reject the *no-bias calibration hypothesis* for both model types. This finding supports the conjecture that there is an individual-specific, systematic component in the error term that may lead to bias in hypothetical responses (see, e.g., Mansfield [41], Andreoni [1], Herriges and Shogren [27]).

4.4. *A Person Understates a Real Offer if He is Older, a Man, or Had Received More Gifts*

Recent studies have suggested that respondent characteristics and attitudes may partly determine whether, and by how much, a person's hypothetical statement differs from his or her real statement (Mansfield [41]). If within-person bias significantly affects the calibration function, the overall relationship between hypothetical and real offers should account for individual-specific effects in the regression model. Our results in Table III suggest that individual-specific factors are important in the relationship between real and hypothetical reported valuations. Given that the fixed effects component of the regression captures any time-invariant subject specific characteristics, we perform an exploratory probe to test for any systematic pattern in the individual effects.

To carry this task out, we use the estimated α_i as the dependent variable in the linear regression model,

$$\alpha_i = \phi + \beta X_i + \varepsilon_i, \tag{3}$$

where α_i are the estimated fixed effects from Eq. (1), ϕ and β are the estimated intercept and slope parameters, X_i are attributes hypothesized to influence the fixed effects, and ε_i is a well-behaved error term. Evidence from the psychology literature suggests that individual response strategies may be a function of personal characteristics (see Krosnick's review [33]). As such, we consider four measurable attributes in X_i—*AGE*, *GENDER* (= 1 male, 0 female), *FAMILY INCOME*, and the number of gifts, *#GIFTS*, that person i could sell in the actual auction.[14]

[13] This result suggests that there are not strong linearities in the relationship. As a robustness test, we also experimented with grouping the goods according to the *JC Penney's* broad categorization of goods: (1) clothing, (2) toys and collectibles, (3) home and leisure, (4) jewelry, and a catch-all category, (5) miscellaneous, for any good that does not fit into one of the first four categories (e.g., beer-intake facilitator). Under this grouping system we find statistical evidence that the nature of the good matters in that we can reject the null hypothesis of homogeneous slope terms at the $p < 0.05$ level. Since we find qualitatively similar results on the margin, we do not present these results, but make them available upon request.

[14] To obtain information on the effects of respondent characteristics we could also estimate a model that combines Eqs. (2) and (3). One approach would be to make (Hypothetical value − Actual value)/Actual value $= f(X_i) + \varepsilon_i$, where X_i includes *AGE*, *GENDER*, *FAMILY INCOME*, and *#GIFTS*. Given that we would not be able to control for other important static factors using this approach because the rank condition would be violated if we included fixed effects in the model, we opt for our two-step regression model.

TABLE IV
Determinants of the Fixed Effects Component of the Calibration Function

Variable	Mean (standard deviation)	Open-ended survey survey	Hypothetical auction
CONSTANT	—	−285.7[a]	−277.5[a]
		(1.8)	(1.8)
AGE	24.22	6.3	6.1
	(7.0)	(1.5)	(1.5)
GENDER	0.58	99.3	105.3
	(0.50)	(1.5)	(1.6)
#GIFTS	6.78	14.9[b]	14.3[a]
	(3.99)	(2.0)	(1.9)
FAMILY	$68,472	−0.12E-03	−0.11E-3
INCOME	($90,385)	(1.1)	(1.0)
R^2	—	0.23	0.22
Adj. R^2	—	0.13	0.12
N		36	36

Notes: (1) Dependent variable is estimated fixed effect α_i from Eq. (2), standard errors are corrected for heteroscedasticity;
(2) Absolute values of ratios are beneath coefficient estimates;
(3) Family income coefficients are in scientific notation;
(4) Gender = 1 if male, 0 if female.
[a] Significant at the 0.10 level.
[b] Significant at the 0.05 level.

Column 1 in Table IV presents the descriptive statistics for the regressors. The sample statistics indicate that the average subject was 24.22 years old, received 6.78 Christmas presents, and had a family income of $68,472. Males comprised 58% of the sample.

Columns 2 and 3 in Table IV contain summary regression estimates of Eq. (3). Given that the dependent variable in Eq. (3) consists of *estimated* person-specific fixed effects, which can introduce heteroscedasticity of an unknown form, the t statistics reported in Table IV use the White heteroscedasticity correction. Regression diagnostics suggest that both models are significant at conventional levels, and across specifications parameter estimates are similar. Focusing on the open-ended survey results in column 2, we find that the fixed effects are at least partially determined by demographic factors, as some coefficients are significantly different from zero at conventional levels. More specifically, the evidence suggests that a person was more likely to understate a real offer if he or she was older, a man, or had received more gifts.[15] For example, the coefficient of *#GIFTS* suggests that for each extra gift received, the fixed effect, α_i, is increased by $14.90—or equivalently, to provide a more accurate predictor of actual behavior, the subject's hypothetical value needs to be adjusted upward by $14.90 for each additional gift he or she receives. This finding suggests that hypothetical auctions do not necessarily provide incentives for people to work through the cognitive processes to evaluate each and every good seriously. Overall, our general observations are consistent with the psychology findings discussed in Krosnick [33]. Although

[15] We should note, however, that using a one-sided alternative the coefficients of *AGE* and *GENDER* are only significant at the $p < 0.07$ level.

intuition suggests people should play it safe by stating large WTA values when their actions or preferences are not well thought out (e.g., Hoehn and Randall [28]), our results suggest otherwise.

5. CONCLUDING COMMENTS

Calibration research may eventually lead to generalizations about behavior that converts experimental results into theory. Such a conversion requires we understand how people behave across both WTP and WTA scenarios. Although WTP statements have been examined thoroughly, much less has been done in the area of compensation demanded. In this study, we fill this gap by using data from List and Shogren's [38] study of gift-giving. Our raw data from a within-subject experimental design suggest that people understated their real WTA in the hypothetical regimes, framed both as demand and non-demand revealing. After controlling for person-specific effects, however, hypothetical and real statements are equivalent on the margin: a one dollar increase in a hypothetical valuation statement is associated with a one dollar increase in actual value.

Calibration studies have thus far been a series of exercises in pattern recognition. One result suggests people understate real WTP; another finds people overstate WTP; and still another observes neither. Why? Is it the exchange institution? The subject pool? The good? The context? All of the above? The answer is not immediately obvious, and it is unanswered questions like these that continue to make the gap between intentions and actions an irascible issue in non-market valuation. This is probably because the perception of a hypothetical stain has never really been systematically removed by an industrial-strength theory in over two decades of debate. We agree with Mansfield's ([41], p. 680) point that "the power of the calibration model could be improved by a better understanding of how individuals answer CV questions, including the traits or attitudes that inspire individuals to give more or less accurate answers." No camp has a completely convincing and axiomatic explanation as to what creates or removes the wedge between intentions and actions. The debate will likely be palliated only when a robust theoretical or behavioral reason emerges as to why the wedge happens and whether it can be controlled systematically. Designing an experiment to understand whether a model of context-dependent preferences can help organize behavior in valuation exercises is the next step in our research program.

REFERENCES

1. J. Andreoni, Cooperation in public goods experiments: kindness and confusion, *Amer. Econ. Rev.* **85**, 891–904 (1995).
2. E. Balistreri, G. McClelland, G. Poe, and W. Schulze, Can Hypothetical Questions Reveal True Values? A Laboratory Comparison of Dichotomous Choice and Open-Ended Contingent Values with Auction Values, Cornell University, Working Paper, 97-15 (1998).
3. R. Bishop and T. Heberlein, Measuring values of extramarket goods: are indirect measures biased? *Amer. J. Agr. Econ.* **61**, 926–930 (1979).
4. R. Bishop and T. Heberlein, Assessing the validity of contingent valuations: three field experiments, *Sci. Total Environ.* **56**, 434–479 (1986).
5. R. Bishop and T. Heberlein, The contingent valuation method, in "Economic Valuation of Natural Resources," (R. Johnson and G. Johnson, Eds.), Westview Press, Boulder, CO (1990), pp. 81–104.

6. R. Bishop, T. Heberlein, and M. Kealy, Contingent valuation of environmental assets: comparisons with a simulated market, *Natural Res. J.* **23**, 619–633 (1983).
7. M. Blackburn, G. Harrison, and E. Rutström, Statistical bias functions and informative hypothetical surveys, *Amer. J. Agr. Econ.* **76**, 1084–1088 (1994).
8. P. Bohm, Estimating demand for public goods: an experiment, *Europ. Econ. Rev.* **3**, 111–130 (1972).
9. R. Boyce, G. McClelland, T. Brown, G. Peterson, and W. Schulze, An experimental examination of intrinsic values as a source of the WTA-WTP disparity, *Amer. Econ. Rev.* **82**, 1366–1373 (1992).
10. D. Brookshire and D. Coursey, Measuring the value of a public good: an empirical comparison of elicitation procedures, *Amer. Econ. Rev.* **77**, 554–566 (1987).
11. D. Brookshire, D. Coursey, and W. Schulze, Experiments in the solicitation of private and public values: an overview, in "Advances in Behavioral Economics" (L. Green and J. Kagel, Eds.), Vol. 2, Ablex, Norwood, NJ (1990).
12. T. Brown, P. Champ, R. Bishop, and D. McCollum, Which response format reveals the truth about donations to a public good? *Land Econ.* **72**, 152–166 (1996).
13. D. Coursey, J. Hovis, and W. Schulze, The disparity between willingness to accept and willingness to pay measures of value, *Quart. J. Econ.* **102**, 679–690 (1987).
14. R. Cummings, D. Brookshire, and W. Schulze (Eds.), "Valuing environmental goods: an assessment of the contingent valuation method," Rowman and Allanheld, Totowa, NJ (1986).
15. R. Cummings, G. Harrison, and E. Rutström, Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive compatible? *Amer. Econ. Rev.* **85**, 260–266 (1995).
16. R. Cummings and L. Taylor, Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method, *Amer. Econ. Rev.* **89**, 649–665 (1999).
17. P. Diamond and J. Hausman, Contingent valuation: is some number better than no number? *J. Econ. Persp.* **8**, 45–64 (1994).
18. M. Dickie, A. Fisher, and S. Gerking, Market transactions and hypothetical demand data: A comparative study, *J. Amer. Statist. Assoc.* **82**, 69–75 (1987).
19. R. Engelbrecht-Wiggans, J. List, and D. Lucking-Reilley, Demand reduction in multi-unit auctions with varying numbers of bidders: theory and field experiments, University of Arizona, Working Paper (2000).
20. R. Forsythe and M. Isaac, Demand-revealing mechanisms for private good auctions, in "Res. in Exper. Econom." (Vernon Smith, Ed.), JAI Press, Inc., Greenwich, CT (1982), Vol. 2, pp. 45–61.
21. V. Foster, I. Bateman, and D. Harley, Real and hypothetical willingness to pay for environmental preservation: a non-experimental comparison, *J. Agr. Econ.* **48**, 123–138 (1997).
22. J. Fox, J. Shogren, D. Hayes, and J. Kliebenstein, CVM-X: calibrating contingent values with experimental auction markets, *Amer. J. Agr. Econ.* **80**, 455–465 (1998).
23. R. Franciosi, M. Isaac, D. Pingry, and S. Reynolds, An experimental investigation of the Hahn-Noll Revenue Neutral Auction for Emissions Licenses, *J. Environ. Econom. Manage.* **24**, 1–24 (1993).
24. P. Frykblom, Hypothetical question modes and real willingness to pay, *J. Environ. Econ. Manage.* **34**, 275–287 (1997).
25. D. Grether and C. Plott, Economic theory of choice and the preference reversal phenomenon, *Amer. Econ. Rev.* **69**, 623–638 (1979).
26. J. Hausman, Specification tests in econometrics, *Econometrica* **46**, 1251–1271 (1978).
27. J. Herriges and J. Shogren, Starting point bias in dichotomous choice valuation with follow-up questioning, *J. Environ. Econ. Manage.* **30**, 112–131 (1996).
28. J. Hoehn and A. Randall, A satisfactory benefit cost indicator from contingent valuation, *J. Environ. Econ. Manage.* **14**, 226–247 (1987).
29. J. Horowitz, and K. McConnell, "A review of WTA/WTP studies," Working Paper 98-05, University of Maryland (1998).
30. J. Irwin, G. McClelland, and W. Schulze, Hypothetical and real consequences in experimental auctions for insurance against low probability risks, *J. Behav. Decision Making* **5**, 107–116 (1992).
31. M. Kealy, J. Dovidio, and M. Rockel, Accuracy in valuation is a matter of degree, *Land Econ.* **64**, 158–171 (1988).
32. M. Kealy, J. Montgomery, and J. Dovidio, Reliability and predictive validity of contingent values: does the nature of the good matter? *J. Environ. Econ. Manage.* **19**, 244–263 (1990).

33. J. Krosnick, Response strategies for coping with the cognitive demands of attitude measures in surveys, *Appl. Cog. Psych.* **5**, 213–236 (1991).
34. J. List, Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards, *Amer. Econ. Rev.* (to be published).
35. J. List, M. Margolis, and J. Shogren, Hypothetical-actual bid calibration of a multi-good auction, *Econ. Letters* **60**, 263–268 (1998).
36. J. List and D. Lucking-Reilly, Demand reduction in a multi-unit auction: evidence from a sportscard field experiment, *Amer. Econ. Rev.* **90**, 961–972 (2000).
37. J. List and J. Shogren, Calibration of the difference between actual and hypothetical valuations in a field experiment, *J. Econ. Behavior Org.* **37**, 193–205 (1998a).
38. J. List and J. Shogren, The deadweight loss of Christmas: comment, *Amer. Econ. Rev.* **88**, 1350–1355 (1998b).
39. J. Loomis, T. Brown, T. Lucero, and G. Peterson, Improving validity experiments of contingent valuation methods: results of efforts to reduce the disparity of hypothetical and actual willingness to pay, *Land Econ.* **72**, 450–461 (1996).
40. G. McClelland, W. Schulze, and D. Coursey, Insurance for low-probability hazards: a bimodal response to unlikely events, *J. Risk Uncertainty* **7**, 95–116 (1993).
41. C. Mansfield, A consistent method for calibrating contingent value survey data, *So. Econ. J.* **64**, 665–681 (1998).
42. B. Melton, W. Huffman, J. Shogren, and J. Fox, Consumer preferences for fresh food items with multiple quality attributes: evidence from an experimental auction of pork chops, *Amer. J. Agr. Econ.* **78**, 916–923 (1996).
43. G. Miller and C. Plott, Revenue generating properties of sealed-bid auctions: an experimental analysis of one-price and discriminative auctions, in "Res. in Experimental Econom." (V. Smith, Ed.), JAI Press, Inc., Greenwich, CT (1985), Vol. 3, pp. 159–182.
44. H. Neill, R. Cummings, P. Ganderton, G. Harrison, and T. McGuckin, Hypothetical surveys and real economic commitments, *Land Econ.* **70**, 145–154 (1994).
45. National Oceanic and Atmospheric Administration, Natural resource damage assessment: proposed rules, Federal Register, 4 May, **59**, 23098–23111 (1994).
46. National Oceanic and Atmospheric Administration, Natural resource damage assessments: final rules, Federal Register, 5 January, **61**, 439 (1996).
47. A. Randall, Calibration of CV responses: discussion, in "The Contingent Valuation of Environmental Resources" (D. Bjornstad and J. Kahn, Eds.), Edgar Elgar, London, (1996), pp. 198–207.
48. E. Rutström, Home-grown values and the design of incentive compatible auctions, *Intl. J. Game Theory* **27**, 427–441 (1998).
49. K. Samples, M. Gowen, and J. Dixon, "The validity of the contingent valuation method for estimating non-use components of preservation value for unique natural resources," Working Paper presented at the American Agricultural Economics Associations meetings in Reno, Nevada (1986).
50. K. Seip and J. Strand, Willingness to pay for environmental goods in Norway: a contingent valuation study with real payment, *Environ. Res. Econ.* **2**, 91–106 (1992).
51. J. Sinden, Empirical tests of hypothetical biases in consumers' surplus surveys, *Australian J. Agr. Econ.* **32**, 98–112 (1988).
52. V. K. Smith and C. Mansfield, Buying time: real and hypothetical offers, *J. Environ. Econ. Manage.* **36**, 209–224 (1998).
53. M. Spencer, S. Swallow, and C. Miller, Valuing water quality monitoring: a contingent valuation experiment involving hypothetical and real payments, *Agr. Res. Econ. Rev.* **27**, 28–42 (1998).
54. W. Vickrey, Counterspeculation, auctions, and competitive sealed tenders, *J. Finance* **16**, 8–37 (1961).
55. J. Waldfogel, The deadweight loss of Christmas, *Amer. Econ. Rev.* **83**, 1328–1336 (1993).