A Replication of Willingness-to-pay Estimates in ‘An Adding Up Test on Contingent Valuations of River and Lake Quality’ (Land Economics, 2015)

John C. Whitehead

Abstract

Desvousges, Mathews and Train (2015) find that their contingent valuation method (CVM) survey data does not pass the adding-up test using a conservative, nonparametric estimate of mean willingness-to-pay. First I show theoretically that the elicited willingness-to-pay is not appropriate for conduct of an adding-up test. Proceeding as if the theory is not a concern I describe how their data suffers from non-monotonicity, flat bid curve and fat tails problems, each of which can cause willingness-to-pay estimates to be sensitive to the approach chosen to measure the central tendency. Using additional parametric approaches that are standard in the literature, I find that willingness-to-pay for the whole is not statistically different from the sum of the parts in two of three additional estimates. In additional robustness checks, all six of the additional tests find that the WTP estimates do not reject the adding-up hypothesis. The negative result in Desvousges, Mathews and Train (2015) is not robust to these alternative approaches to willingness-to-pay estimation.

JEL Q51

Key Words Contingent valuation, Adding-up test, Willingness-to-pay

Author

John C. Whitehead, Department of Economics, Appalachian State University, Boone, North Carolina, 28608, USA

1 Introduction

The contingent valuation method (CVM) is a stated preference survey approach to the valuation of public goods (Mitchell and Carson 1989, Haab and Whitehead 2015). The scope test is an internal validity test where willingness-to-pay (WTP) estimates are expected to increase with the scope of the public good (i.e., "more is better"). Desvousges, Mathews and Train (2012) catalog CVM studies into those that pass the scope test, those that fail the scope test and those that have mixed results. They find that a significant number of studies fail to pass the test and question the validity of the method. Desvousges, Mathews and Train (2012) go further and critique the Chapman et al. (2009) unpublished natural resource damage assessment technical report, arguing that it does not pass the scope test "adequately."

Desvousges, Mathews and Train's (2012) requirement for adequacy is the so-called "adding-up test." The adding-up test was proposed by Diamond (1996) who provides this description in footnote 14 on page 343:

> *As examples of possible adding-up tests, consider variations on two recent surveys. Schulze et al. used two surveys to ask for WTP for partial and complete cleanups of the Upper Clark Fork River Basin in Montana. For an adding-up test, a third survey would describe a partial cleanup and describe the government as already committed to it, with the costs to be borne as described in the existing survey. The survey would then describe a complete cleanup and ask for WTP to enhance the cleanup from partial to complete. The mean WTP response from this question plus the mean WTP for partial cleanup should be almost exactly the same as the mean WTP for complete cleanup. One could test for the statistical significance of any difference that was found.*

Desvousges, Mathews and Train (2012) reinterpret the two scenario scope test in Chapman et al. (2009) as a three scenario adding-up test. They then assert that the implicit third willingness-to-pay estimate is not of adequate size. Whitehead (2016) critiques the notion of the adding-up test as an adequacy requirement and proposes an alternative measure of the economic significance of the scope test: scope elasticity. In a comment, Chapman et al. (2016) argue that Desvousges, Mathews and Train (2012) misinterpret their scope test and suggest that the Chapman et al. survey design should not be interpreted as an adding-up test. Desvousges, Mathews and Train (2016) reply that they did not misinterpret the Chapman et al. survey design.

In this context, Desvousges, Mathews and Train (DMT, 2015) field the Chapman et al. (2009) survey with new sample data and three additional scenarios. DMT conduct an adding-up test and argue that willingness-to-pay (WTP) for the whole should be equal to willingness-to-pay for the sum of four parts (the first, second, third and fourth increment scenarios). DMT find that "The sum of the four increments … is about three times as large as the value of the whole" (p. 566). In this replication I examine DMT's conclusion using alternative parametric approaches for estimating the central tendency of WTP.[1]

[1] In Appendix A I describe the CVM scenarios in Chapman et al. (2009) and DMT (2015).

This replication is necessary because dichotomous choice contingent valuation questions propose a cost to respondents who then indicate whether or not they are willing to pay the cost. One theoretical validity test is for whether the percentage of respondents who are willing to pay the cost declines as the cost increases. DMT's data suffers from non-monotonicity (i.e., the percentage of affirmative responses does not always decrease as the bid increases), flat portions of the bid curve and fat tails. As such, the WTP estimate may be very sensitive to the assumptions of the estimation approach used.

Following Chapman et al. (2009), DMT choose the ABERS nonparametric estimator for willingness-to-pay (Ayer et al. 1955). Chapman et al. (2009) describe the ABERS estimator as producing a lower bound WTP estimate. The ABERS estimator is a special case of the more familiar Turnbull nonparametric lower bound WTP estimator (Haab and McConnell 1997, Carson and Hanemann 2005, Boyle 2017). When data is non-monotonic, the Turnbull approach smooths non-monotonic bid curves by pooling the percentages of those willing to pay across cost amounts and ignores validity problems associated with non-monotonically decreasing portions of the bid curve. The Turnbull estimates truncate the WTP distribution at the highest bid, ignoring the potential fat tail of the WTP distribution.

In the remainder of this paper I first argue that DMT (2015) fail to elicit willingness-to-pay appropriate for a true adding-up test. Next, ignoring the theory, I replicate the DMT willingness-to-pay estimates with the Turnbull (Haab and McConnell 1997) and reproduce DMT's negative result on the adding-up test. In section four I present two parametric models of WTP that lead to three additional WTP estimates for each scenario. One of these estimates supports DMT's negative adding up test result but two fail to support the negative result. In an appendix I present six additional robustness checks, all of which find that the WTP estimates do not reject the adding up hypothesis. In the conclusions I offer recommendations for future CVM studies on conducting sensitivity analysis for WTP estimation approaches.

2 The Adding-up Test

Consider two public goods, q_1 and q_2. Independent definitions of willingness-to-pay for improvements, $q_1^* > q_1$, $q_2^* > q_2$ and $q_1 + q_2$ are:

$$v(q_1, q_2, Y) = v(q_1^*, q_2, Y - WTP_1)$$

$$v(q_1, q_2, Y) = v(q_1, q_2^*, Y - WTP_2)$$

$$v(q_1, q_2, Y) = v(q_1^*, q_2^*, Y - WTP_{1+2})$$

where $v(\cdot)$ is the indirect utility function and Y is income.

According to standard scope test theory, willingness-to-pay for goods 1 and 2 is expected to be greater than or equal to willingness-to-pay for good 1, $WTP_{1+2} \geq WTP_1$, and good 2, $WTP_{1+2} \geq WTP_2$. Due to substitution effects, valuation of goods 1 and 2 is context dependent. If good 1 (2) is valued first in a sequence then its willingness-to-pay will be higher than if it is valued second (Carson, Flores and Hanemann 1998) due to substitution effects. Therefore, the sum of independently valued goods 1 and 2 will be greater than the willingness-to-pay for independent valuations of goods 1 and 2, $WTP_{1+2} < WTP_1 + WTP_2$.

Suppose WTP_1 is elicited independently as describe above. In an adding up test as described by Diamond (1996), willingness-to-pay for good 2 would be elicited with the following definition:

$$v(q_1, q_2^*, Y - A) = v(q_1^*, q_2^*, Y - A - WTP_2)$$

Willingness-to-pay for the change in good 2 in an adding up test is $WTP_2[\Delta q_2|\Delta q_1, Y - A]$ indicating that the valuation of good 2 proceeds after the provision of good 1 has been made and A is the amount of money taken from the respondent to pay for provision of good 1. The effect of provision of q_1 on willingness-to-pay for q_2 is negative, $\frac{\partial WTP_2}{\partial q_1} < 0$, if q_1 and q_2 are substitutes. The effect of payment for the provision of q_1 on willingness-to-pay for q_2 is also negative, $\frac{\partial WTP_2}{\partial A} < 0$, as the budget constraint tightens.

An explicit description of the conditions under which a valuation is made is necessary to account for income and substitution effects. For n goods, the adding-up test requires $n + 1$ different scenarios. In the case of 2 public goods, there are four valuation steps:

1. Sample 1: Elicit the willingness-to-pay for good 1 in scenario 1
2. Sample 2: Describe that good 1 has been provided at a cost of A to the respondent
3. Sample 2: Elicit the willingness-to-pay for good 2 in scenario 2
4. Sample 3: Elicit the willingness-to-pay for goods 1 and 2 in scenario 3

Following the adding-up test theory, in order to accurately elicit $WTP_2[\Delta q_2|\Delta q_1, Y - A]$ one would need to describe the provision of good 1, describe the extraction of A from the survey respondent and how its provision would reduce the income of the survey respondent before elicitation of WTP_2.[2] The adding-up test is $WTP_{1+2} = WTP_1 + WTP_2[\Delta q_2|\Delta q_1, Y - A]$, where $WTP_2[\Delta q_2, Y] > WTP_2[\Delta q_2|\Delta q_1, Y - A]$.

DMT do not explicitly describe the counterfactual situation required by the adding-up test in step 2 above. Nevertheless, DMT (2015) conduct a two-tailed adding-up test for equality between willingness-to-pay for the whole and willingness-to-pay for the sum of the four parts:

HO: $WTP_{whole} = \sum_{i=1}^{4} WTP_i$
HA: $WTP_{whole} \neq \sum_{i=1}^{4} WTP_i$

Instead of additional survey text, DMT elicit the four parts just as you would elicit willingness-to-pay for each of the four parts independently. Economic theory suggests the appropriate statistical test considering the survey design in DMT (2015) is a one-tailed test[3]:

[2] Inclusion of these two counterfactual conditions in a CVM survey would likely impose additional cognitive burden on the survey respondent.

[3] Note that the Turnbull results in the next section support the null hypothesis.

H0: $WTP_{whole} < \sum_{i=1}^{4} WTP_i$
HA: $WTP_{whole} \geq \sum_{i=1}^{4} WTP_i$

DMT (2015) acknowledges that theory suggests this test, a different null hypothesis than what they test in their paper, with their income effects simulation. Their implicit claim is that income effects are typically so small in CVM studies that an appropriate survey design is not important. However, DMT (2015) do not address the potential for substitution effects.

This section argues that DMT did not elicit the correct willingness-to-pay estimates to conduct an adding-up test. Nevertheless, I will proceed through the remainder of this paper and re-analyze their data as if they did.

3 WTP Replication

The data from DMT is presented in Table 1. The randomly assigned cost amounts presented to respondents for each scenario is presented in the first column. The number of "yes" responses (Yes), the subsample size (N) and the percentage of "yes" responses (%Yes) is presented for the Whole, First, Second, Third and Fourth scenarios. A description of the survey text used by Chapman et al. (2009) and DMT (2015) is presented in Appendix A.

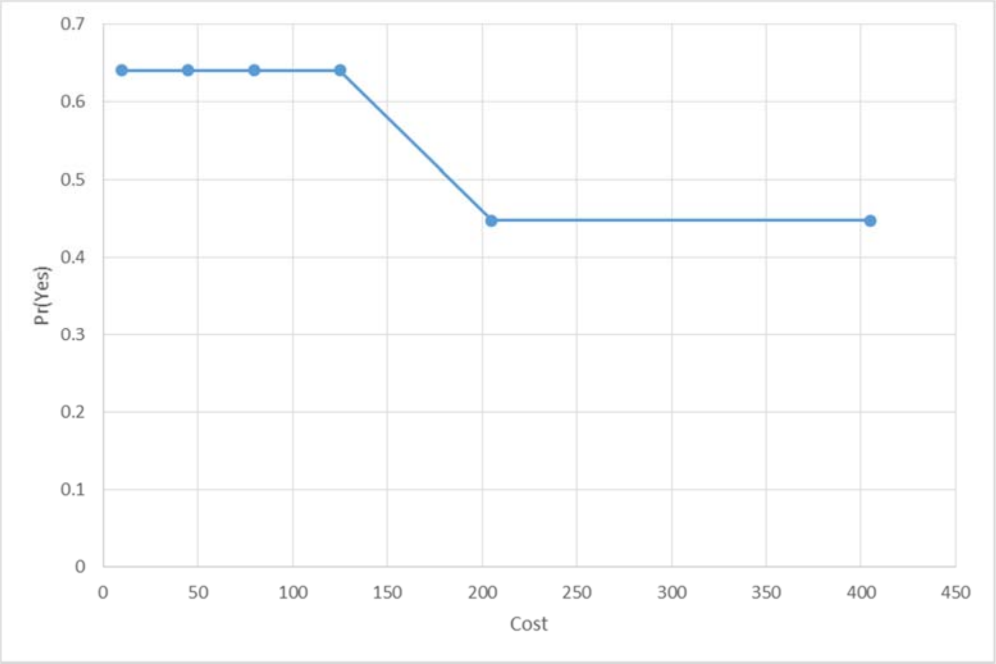
Each of the scenarios exhibits non-monotonicity in at least one of the five cost increases. In the whole scenario the percentage yes is 61 at $45 and 69 at $80 (in bold). The first scenario exhibits non-monotonicity as the cost increases from $45 to $80 and $205 to $405. The second scenario exhibits non-monotonicity as the cost increases from $80 to $125. The third scenario exhibits non-monotonicity as the cost increases from $125 to $205 and then to $405. The fourth scenario exhibits non-monotonicity as the cost increases from $45 to $80 and $125 and $205 and $405.

Table 1. Dichotomous Choice CVM Data (DMT 2015)

	Whole			First			Second			Third			Fourth		
Cost	Yes	N	%Yes	Yes	N	%Yes	Yes	N	%Yes	Yes	N	%Yes	Yes	N	%Yes
10	17	25	68	38	51	75	12	24	50	24	29	81	24	33	73
45	20	33	61	28	48	58	12	32	38	13	27	48	11	25	44
80	18	26	**69**	31	48	**65**	7	24	29	10	31	32	24	37	**65**
125	14	28	50	27	47	57	12	28	**43**	6	26	23	20	32	**63**
205	13	29	45	21	54	39	6	25	24	11	27	**41**	10	28	36
405	14	31	45	18	45	**40**	4	26	15	12	34	**35**	11	27	**41**
Total	96	172	56	163	293	56	53	159	33	76	174	44	100	182	55

Even when the yes responses are monotonically decreasing in the cost amount in Table 1, the slope is not statistically different from zero in large portions of the bid curves. For example, the whole and second scenarios are characterized by two flat portions of the bid curve. A stylized example is illustrated in Figure 1 where the percentage of yes responses is constant over the lower range of cost amounts ($10 to $125), is downward sloping from $125 to $205 and flat from \$205 to \$405.

Figure 1. Bid curve with two flat portions



For the whole scenario, the slope of the bid curve over the entire range of cost amounts (\$10 to \$405) is downward sloping with $b = -.00058$ ($t = -2.09$, n = 172) estimated with a linear probability model ($\Pr(Yes) = a + b \times Cost$). But, the slopes over the lower (\$10 to \$80) and upper (\$125 to \$405) ranges of cost amounts are flat with $b = 0.0019$ ($t = 0.10$, n = 84) and $b = -0.00013$ ($t = -0.29$, n = 88), respectively. Similarly, in the second scenario the slope of the bid curve over the entire range of cost amounts (\$10 to \$405) is downward sloping with $b = -.00074$ ($t = -2.63$, n = 159). But, the slopes over the lower (\$10 to \$125) and upper (\$205 to \$405) ranges of cost amounts are flat with $b = -0.00056$ ($t = -0.49$, n = 109) and $b = -0.00043$ ($t = -0.76$, n = 51), respectively. Flat slopes in the upper range of the bid distribution leads to the fat tails problem.

Estimation of the ABERS and Turnbull requires a valid cumulative distribution function that is non-decreasing in the cost amount. An invalid CDF is non-monotonic. Non-monotonicity can be caused by either a lack of theoretical validity of the data, a lack of attention being paid to cost amounts by survey respondents or due to sampling variability when small sample sizes are employed (as in Table 1). With non-monotonic data, nonparametric WTP estimators require pooling of yes responses across cost amounts until weak monotonicity is achieved (Haab and McConnell 2002). Weak monotonicity occurs in the data when the percentage of yes responses is equal across bid amounts. When the probabilities for two pooled costs are higher than the next lowest cost the pooling continues until the bid curve is non-monotonically non-increasing in the cost amount. The pooled dichotomous choice data are presented in Table 2.

Table 2. Monotonically Non-increasing Probability of a Yes Response

	%Yes				
Cost	Whole	First	Second	Third	Fourth
10	68	75	50	83	73
45	64	61	38	48	59
80	64	61	37	33	59
125	50	57	37	33	59
205	45	39	24	33	38
405	45	39	15	33	38

The lower bound Turnbull WTP estimate is the step function formed by the data in Table 2 (Haab and McConnell 1997, 2002). The Turnbull WTP estimates are presented in Table 3 with standard errors (SE) computed as in Haab and McConnell (1997, 2002), a common approach found in the CVM literature (see e.g., Egan, Corrigan and Dwyer 2015). The Turnbull WTP estimates are equal to the WTP estimates presented by DMT when rounded. The Turnbull standard errors are larger than the DMT standard errors who "used bootstrapping techniques" without sufficient detail to perform the replication.

Table 3. Nonparametric Willingness-to-pay Estimates

	DMT (2015)		Replication	
	WTP	SE	WTP	SE
Whole	200	17.71	200.38	19.65
First	187	12.31	186.63	15.03
Second	97	13.73	97.33	18.16
Third	144	15.34	144.11	22.69
Fourth	181	18.69	181.47	23.66

The null hypothesis consistent with the adding up test as discussed by DMT is $HO\colon WTP_{whole} = \sum_{i=1}^{4} WTP_i$. With the Turnbull estimates $\sum_{i=1}^{4} WTP_i = 610$ which is \$409 greater than WTP_{whole}. The larger Haab and McConnell standard errors will favor the null hypothesis of the adding-up test. Nevertheless, with the standard error for the sum of the four parts constructed as the square root of the sum of the variances of the four parts (SE = 45) (Haab and McConnell 2002)[4], the WTP estimates fail the adding up test, replicating the result in DMT (2015).

4 Parametric Estimates of WTP

In order to investigate the robustness of DMT's results, I combine the data from the sub-samples and estimate linear and log linear parametric dichotomous choice models as described by Boyle (2017): $\ln(\Pr(Yes)/(1-\Pr(Yes)) = a + b \times Cost$ and $\ln(\Pr(Yes)/(1-\Pr(Yes)) = a + b \times lnCost$ These models are specified so that each scenario (whole, first,

[4] DMT "applied the bootstrap method to simulate the sampling distribution of the difference between the mean WTP for the whole and the sum of the mean WTP from the four increments."

second, third and fourth) has its own constant and its own cost coefficient. The models are estimated using LIMDEP version 10 (http://www.limdep.com).

In each of the models the slope coefficients (b) are statistically different from zero (Table 4). In the linear logit model the constants for the whole, first, and fourth scenarios are statistically different from zero. In the log linear logit all constants except in the second scenario are statistically different from zero. The log linear model provides a better statistical fit than the linear logit.

Table 4. Dichotomous Choice Probability Models

	Linear Logit			Log Linear Logit		
Constant (a)	Coefficient	SE	t-stat	Coefficient	SE	t-stat
Whole	0.594	0.235	2.53	1.58	0.653	2.42
First	0.726	0.182	4.00	2.11	0.503	4.19
Second	-0.190	0.249	-0.76	0.96	0.664	1.45
Third	0.145	0.229	0.64	2.19	0.644	3.39
Fourth	0.610	0.225	2.70	1.73	0.617	2.81
Slope (b)						
Whole	-0.0023	0.0012	-2.05	-0.298	0.141	-2.14
First	-0.0035	0.0010	-3.65	-0.422	0.108	-3.90
Second	-0.0039	0.0015	-2.51	-0378	0.154	-2.54
Third	-0.0027	0.0012	-2.29	-0.549	0.146	-3.91
Fourth	-0.0030	0.0011	-2.45	-0.347	0.136	-2.60
χ^2	66.08			80.67		
McFadden R^2	0.05			0.06		
Sample size	980			980		

[a]t-statistics are for the null hypothesis that coefficient estimates are equal to zero.

The parametric willingness-to-pay estimates are presented in Table 5. Mean (and median) WTP from the linear logit, which allows negative WTP, is the negative ratio of the constant and the slope: $WTP = -a/b$ (Hanemann 1984). Estimating WTP only over the positive portion of the distribution from the linear logit uses the formula: $WTP = \left(\frac{-1}{b}\right)\ln(1 + \exp(a))$ (Hanemann 1989). Median WTP from the log linear logit is the exponential of the negative ratio of the constant and slope: $WTP = \exp\left(-\frac{a}{b}\right)$. Mean WTP from the log linear model is undefined when $-\frac{1}{b} > 1$ (Haab and McConnell 2002) as in this model. Standard errors for individual WTP estimates and the sum of the WTP parts are estimated with the Delta Method and the Wald test in LIMDEP (Cameron 1991, Greene 2017).

Table 5. Willingness-to-pay Estimates

	Linear Logit						Log Linear Logit		
	Mean WTP			Mean WTP > 0			Median WTP > 0		
	WTP	SE	t-stat[a]	WTP	SE	t-stat	WTP	SE	t-stat
Whole	250	81	3.09	434	171	2.56	201	126	1.59
First	208	39	5.34	321	66	4.87	149	46	3.21
Second	-49	80	-0.62	156	46	3.42	13	11	1.20
Third	54	69	0.78	285	96	2.96	54	17	3.17
Fourth	205	58	3.53	352	112	3.13	147	71	2.07
Sum of Parts	418	127	3.29	1114	168	6.63	359	92	3.90

[a]t-statistics are for the null hypothesis that WTP estimates are equal to zero.

The parametric WTP estimates are economically different than the nonparametric estimates. Considering the whole scenario, the WTP estimates are 25%, 117% and 0.5% larger than the Turnbull estimates in the three estimates from the two models. The similarity between the mean Turnbull and the median WTP from the log-linear model is only coincidence since the two estimates are based on different measures of central tendency. Considering the sum of the parts, the WTP estimates are -31% smaller, 83% larger and -41% smaller than the Turnbull estimates.

The null hypothesis of equality between WTP for the whole scenario and WTP for the sum of the parts cannot be rejected in two of the three adding up tests. The linear logit that allows for negative mean WTP estimates yields a difference of $168 that is not statistically different from zero (t=1.12). These WTP estimates pass the adding up test. In the linear logit with the mean WTP constrained to be positive the difference between the whole and the sum of the parts is $680 which is statistically different from zero (t=2.85). These WTP estimates fail to pass the adding up test. The log linear logit produces a difference of $187 in median WTP that is not statistically different from zero (t=1.05). The median WTP estimates pass the adding up test.

There are benefits and costs of each of the alternative WTP estimators presented here. Haab and McConnell (2002, page 106) suggest that when “there are concerns about the distribution of response data,” as here, researchers should estimate the Turnbull mean and the log-linear median to present conservative willingness-to-pay estimates. Haab and McConnell do not expand on the appropriate measure of central tendency when conducting hypothesis tests although it is clear they prefer the log-linear. Their concern is that the linear functional form is mis-specified when negative willingness-to-pay is allowed and truncation at zero is arbitrary. Similarly, the Turnbull mean is problematic due to arbitrary truncation at the highest bid and the pooling of data. The most reliable adding-up test is conducted with the medians estimated from the log-linear data.

In Appendix B, I conduct additional tests using DMT’s post-stratification weights and a subsample of complete case data. I estimate the linear model with WTP constrained to be positive. This is the measure of WTP that leads to rejection of the adding-up hypothesis above, supporting DMT’s results. All six of these additional adding-up tests fail to support the negative result in DMT (2015).

5 Conclusions

While I argue that DMT do not elicit WTP appropriate for the conduct of an adding-up test, this replication with the DMT data shows that the adding-up null hypothesis cannot be rejected under two of three alternatives with commonly used parametric econometric specifications. The failure to replicate DMT's results with the parametric models is due to data quality problems: non-monotonicity, flat portions over wide ranges of the bid function and fat tails. Each of these problems leads to high variability in mean WTP across estimation approach and larger standard errors than those associated with nonparametric estimators which rely on smoothed data.

The data quality problems are particularly apparent in the whole and second scenarios which are the versions of the survey developed by Chapman et al. (2009). Chapman et al. (2009) use in-person interviews with a large probability sample, as recommended by Arrow et al. (1993), and provides evidence of consequentiality which may be due to differences in survey mode and sample. DMT (2015) fail to replicate the Chapman et al. (2009) study. DMT (2015) use a relatively inexpensive, small non-probability opt-in sample (Bill Desvousges, personal communication, February 19, 2015) that may provide little incentive for respondent attention (Sandorf et al. 2016) and an online survey that may suggest a lack of consequentiality (Carson and Groves 2007, Carson, Groves and List 2014). The differences in data quality may be a result of these survey differences.

Future studies should attempt to address these data quality problems with experimental designs that lead to statistical tests with sample sizes large enough to provide sufficient power. This can be achieved in three ways. The most costly, of course, is simply increasing the overall sample size. Holding cost constant, researchers could reduce the number of cost amounts used in the experimental design or reduce the number of experimental scenarios presented. For example, DMT could have implemented their adding up test with three (instead of five) separate scenarios as they describe in Desvousges, Mathews and Train (2012). The interested reader is reminded that Appendix C of Mitchell and Carson (1989), the seminal CVM text, has an extensive discussion of sample size and statistical power.

To be clear, I am not asserting that I have shown that the CVM will pass the adding-up test if data are properly analyzed. The only claim that I can confidently make is that the DMT (2015) data is not strong enough to provide credible evidence that the CVM does not pass the test. An adequate adding-up test would require more resources devoted to the study than is apparent in DMT (2015). A survey instrument would need to be developed with extensive focus groups and pretesting to construct believable scenarios with income and substitution effects. Even if researchers devote the necessary resources to survey design a credible adding-up scenario would still impose an amount of cognitive burden on survey respondents that might make the conduct of adding-up tests difficult. Indeed, laboratory experiment studies have found it difficult to impose the adding-up condition for market goods (Bateman et al. 1997, Elbakidze and Nayga 2017). Considering this, self-guided online opt-in surveys are likely not conducive to satisfying the adding-up test. Researchers should consider devoting resources to the in-person survey mode for the conduct of adding-up tests.

Acknowledgements

The author thanks Bill Desvousges for providing the data, David Chapman for providing sample sizes necessary for reconstruction of the Chapman et al. data and three reviewers and the editor for a number of comments which have greatly improved this paper.

References

Arrow, Kenneth, Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, Howard Schuman, Report of the NOAA Panel on Contingent Valuation, January 11, 1993. http://www.darp.noaa.gov/pdf/cvblue.pdf.

Ayer, Miriam, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. 1955. "An Empirical Distribution Function for Sampling with Incomplete Information." Annals of Mathematical Statistics 26 (4): 641–47. http://projecteuclid.org/euclid.aoms/1177728423.

Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer, and Robert Sugden. "Does part–whole bias exist? An experimental investigation." The Economic Journal 107, no. 441 (1997): 322-332. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0013-0133.1997.160.x

Boyle, Kevin J. "Contingent valuation in practice." In A primer on nonmarket valuation, pp. 83-131. Springer Netherlands, 2017. http://www.springer.com/us/book/9789400771031

Cameron, Trudy Ann. "Interval estimates of non-market resource values from referendum contingent valuation surveys." Land Economics 67, no. 4 (1991): 413-421. http://www.jstor.org/stable/3146548?seq=1#page_scan_tab_contents

Carson, Richard, Nicholas E. Flores, and W. Michael Hanemann. "Sequencing and valuing public goods." Journal of Environmental Economics and Management 36, no. 3 (1998): 314-323. https://www.sciencedirect.com/science/article/pii/S0095069698910506

Carson, Richard T., and W. Michael Hanemann. "Contingent valuation." Handbook of environmental economics 2 (2005): 821-936. http://www.sciencedirect.com/science/article/pii/S1574009905020176

Carson, R.T. and Groves, T., 2007. Incentive and informational properties of preference questions. Environmental and resource economics, 37(1), pp.181-210. https://link.springer.com/article/10.1007/s10640-007-9124-5

Carson, R.T., Groves, T. and List, J.A., 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice. Journal of the Association of Environmental and Resource Economists, 1(1/2), pp.171-207. http://www.journals.uchicago.edu/doi/abs/10.1086/676450

Chapman, David, Richard Bishop, Michael Hanemann, Barbara Kanninen, Jon Krosnick, Edward Morey and Roger Tourangeau. 2009. Natural Resource Damages Associated with Aesthetic and Ecosystem Injuries to Oklahoma’s Illinois River System and Tenkiller Lake.

Chapman, David J., Richard C. Bishop, W. Michael Hanemann, Barbara J. Kanninen, Jon A. Krosnick, Edward R. Morey, and Roger Tourangeau. "On the adequacy of scope test results: Comments on Desvousges, Mathews, and Train." Ecological Economics 130 (2016): 356-360. http://www.sciencedirect.com/science/article/pii/S0921800916306139.

Desvousges, William, Kristy Mathews, and Kenneth Train. "Adequate responsiveness to scope in contingent valuation." Ecological Economics 84 (2012): 121-128. http://www.sciencedirect.com/science/article/pii/S0921800912003813.

Desvousges, William, Kristy Mathews, and Kenneth Train. “An Adding Up Test on Contingent Valuations of River and Lake Quality.” Land Economics 91(2015): 556-571. http://le.uwpress.org/content/91/3/556.refs.

Desvousges, William, Kristy Mathews, and Kenneth Train. “Reply to ‘On the adequacy of scope test results: Comments on Desvousges, Mathews, and Train’,” Ecological Economics 130 (2016): 361–362. http://www.sciencedirect.com/science/article/pii/S0921800916306139.

Elbakidze, Levan, and Rodolfo M. Nayga. "The Adding-Up Test in an Incentivized Value Elicitation Mechanism: The Role of the Income Effect." Environmental and Resource Economics (2017): 1-20. https://link.springer.com/article/10.1007/s10640-017-0177-9

Egan, Kevin J., Jay R. Corrigan, and Daryl F. Dwyer. "Three reasons to use annual payments in contingent valuation surveys: Convergent validity, discount rates, and mental accounting." Journal of Environmental Economics and Management 72 (2015): 123-136. http://www.sciencedirect.com/science/article/pii/S0095069615000443

Greene, William H. Econometric Analysis, 8th Edition, Pearson, 2017. https://www.pearson.com/us/higher-education/program/Greene-Econometric-Analysis-8th-Edition/PGM334862.html

Haab, Timothy C., and Kenneth E. McConnell. "Referendum models and negative willingness-to-pay: alternative solutions." Journal of Environmental Economics and Management 32, no. 2 (1997): 251-270. http://www.sciencedirect.com/science/article/pii/S0095069696909687.

Haab, Timothy C., and Kenneth E. McConnell. Valuing Environmental and Natural Resources: The Econometrics of Non-market Valuation. Edward Elgar Publishing, 2002. https://www.elgaronline.com/view/9781840647044.xml.

Hanemann, W. Michael. "Welfare evaluations in contingent valuation experiments with discrete responses." American Journal of Agricultural Economics 66, no. 3 (1984): 332-341. http://ajae.oxfordjournals.org/content/66/3/332.short.

Hanemann, W. Michael. "Welfare evaluations in contingent valuation experiments with discrete response data: reply." American Journal of Agricultural Economics 71, no. 4 (1989): 1057-1061. https://academic.oup.com/ajae/article-lookup/doi/10.2307/1242685

Parsons, George R., and Kelley Myers. "Fat tails and truncated bids in contingent valuation: An application to an endangered shorebird species." Ecological Economics 129 (2016): 210-219. http://www.sciencedirect.com/science/article/pii/S0921800915301567.

Sandorf, Erlend Dancke, Margrethe Aanesen, and Ståle Navrud. "Valuing unfamiliar and complex environmental goods: A comparison of valuation workshops and internet panel surveys with videos." Ecological Economics 129 (2016): 50-61. https://www.sciencedirect.com/science/article/pii/S092180091530389X

Whitehead, John C. "Plausible responsiveness to scope in contingent valuation." Ecological Economics 128 (2016): 17-22. http://www.sciencedirect.com/science/article/pii/S0921800916302890.

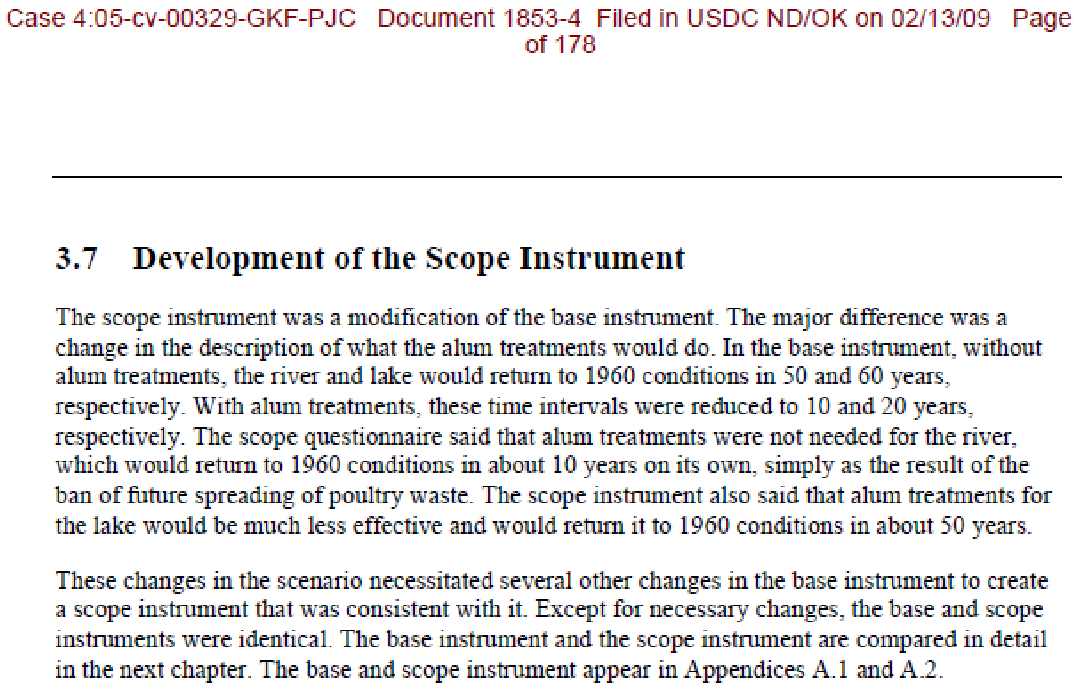
Whitehead, John C. “A Comment on ‘An Adding Up Test on Contingent Valuations of River and Lake Quality.’” No 17-01, Working Papers from Department of Economics, Appalachian State University. http://econpapers.repec.org/paper/aplwpaper/17-01.htm.

Whitehead, John C., and Haab Timothy C. (2013) Contingent Valuation Method. In: Shogren, J.F., (ed.) Encyclopedia of Energy, Natural Resource, and Environmental Economics, Vol. 3, pp. 334-341 Amsterdam: Elsevier. http://www.sciencedirect.com/science/article/pii/B9780123750679000048.

Appendix A. The CVM scenarios

DMT (2015) state that they use the Chapman et al. (2009) survey with minor modifications for the first, third and fourth increment scenarios. Reproduced below in Figures A-1 and A-2 are the initial description (page 3-10) of the Chapman et al. (2009) Base and Scope scenarios. These are called the whole and second increment scenarios in DMT (2015), respectively. Figure 1 from DMT, illustrating their five scenarios, is reproduced in Figure A-3.

Figure A-1. Initial description of Chapman et al.'s (2009) base and scope scenarios (page 3-10)



Case 4:05-cv-00329-GKF-PJC Document 1853-4 Filed in USDC ND/OK on 02/13/09 Page 46 of 178

3.7 Development of the Scope Instrument

The scope instrument was a modification of the base instrument. The major difference was a change in the description of what the alum treatments would do. In the base instrument, without alum treatments, the river and lake would return to 1960 conditions in 50 and 60 years, respectively. With alum treatments, these time intervals were reduced to 10 and 20 years, respectively. The scope questionnaire said that alum treatments were not needed for the river, which would return to 1960 conditions in about 10 years on its own, simply as the result of the ban of future spreading of poultry waste. The scope instrument also said that alum treatments for the lake would be much less effective and would return it to 1960 conditions in about 50 years.

These changes in the scenario necessitated several other changes in the base instrument to create a scope instrument that was consistent with it. Except for necessary changes, the base and scope instruments were identical. The base instrument and the scope instrument are compared in detail in the next chapter. The base and scope instrument appear in Appendices A.1 and A.2.

Figure A-2. Key text from the Chapman et al. (2009) Survey

Base Scenario

> As a result of alum treatments, the river would be back to what it was like in around 1960 about 10 years from now. And the lake would be back to what it was like in around 1960 about 20 years from now. …
>
> Without alum treatments, it will take about 50 years for the river to get back to what is was like in around 1960 instead of about 10 years. That is about 40 years longer. It will take the lake about 60 years to get back to what it was like in around 1960 instead of about 20 years. That is also about 40 years longer. …
>
> If a court bans spreading of poultry litter, the industry will have to safely get rid of all the litter they produce from now on. The industry will have to pay for this,

> and the river and lake will naturally return to what they were like in around 1960. If the people of Oklahoma want this to happen 40 years sooner, there will be an additional cost for the alum treatments. …
>
> *Scope Scenario*
>
> As a result of alum treatments, the lake would be back to what it was like in about 1960 about 50 years from now. …
>
> Without alum treatments, it will take the lake about 60 years to get back to what it was like in around 1960 instead of about 50 years. That is about 10 years longer. …
>
> If a court bans spreading of poultry litter, the industry will have to safely get rid of all the litter they produced from now on. The industry will have to pay for this. The river will naturally return to what it was like in around 1960 in 10 years, and the lake will naturally return to what it was like in around 1960 in 60 years. If the people of Oklahoma want the lake to return to what is was like in around 1960 in 50 years rather than 60 years, there will be an additional cost for the alum treatments. …

DMT's Figure 1 describing their implementation of the survey is below. The whole scenario is an acceleration of river cleanup from 50 to 10 years plus an acceleration of lake cleanup from 60 to 20 years. DMT's whole scenario is the same as Chapman et al.'s "base" scenario. DMT's second increment is the same as Chapman et al.'s "second increment."

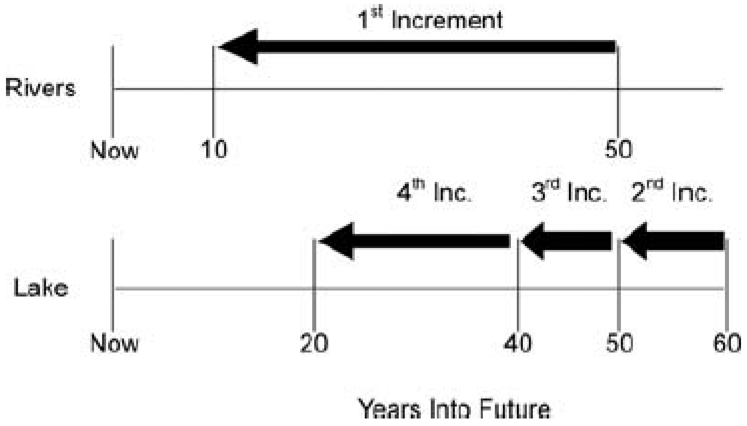


FIGURE 1
Incremental Parts of Accelerated Restoration

Appendix B. Additional tests

DMT report that they conducted sensitivity analysis using post-stratification weights and present regression results with a sample smaller than that used for the mean WTP estimation. In this section I conduct the parametric analysis with these weights and this alternative sample. DMT report that the post-stratification weights do not change the nonparametric results. When I apply the same post-stratification weights, scaled to equal the sample size of n=980, to the models in Table 4 and estimate WTP as in Table 5, none of the three sets of parametric WTP estimates supports rejection of the null hypothesis of equality between WTP for the whole and the sum of the parts.

However, these results are complicated by incorrect signs and statistically insignificant WTP estimates for the most problematic whole and second scenarios. The weighted models produce incorrect signs on the constant and slope in the second scenario (see Figure 1). The incorrect signs lead to a positive weighted WTP estimate of $346 (SE=49) in the second scenario when WTP is estimated over the entire range. The weighted WTP estimate is -$34 (SE=15) when it is estimated only over the positive range. But, both of these WTP estimates are nonsensical given the positive relationship between cost and the probability of a yes response.

Considering the whole scenario, the weighted WTP is $1154 (SE=1289) when estimated only over the non-negative range and the sum of the weighted WTP parts is $811 (SE = 212) (see Figure 1). The statistically insignificant weighted mean WTP for the whole scenario leads to wide confidence intervals for which it is difficult to reject the null hypothesis of equality between WTP for the whole and the sum of the parts.

DMT (2015) conduct their nonparametric WTP estimation with a full sample of n=980. Yet, they conduct regression analysis with a sample of n=950 in order to estimate income effects for their simulation of $Y - A$ (see section 2 of this paper). A close examination of the data reveals that there are only 936 cases that do not suffer from item nonresponse. Forty-three cases have missing income values for which 14 unconditional means of the income variable are imputed for the n=950 regression analysis. There are 30 cases with item nonresponse in the age variable. These 30 cases are dropped for the n=950 regression analysis in DMT (2015). There is one missing age value that occurs with a nonmissing income value so the total number of cases with missing age and/or income values is 44 (see Figure 2).

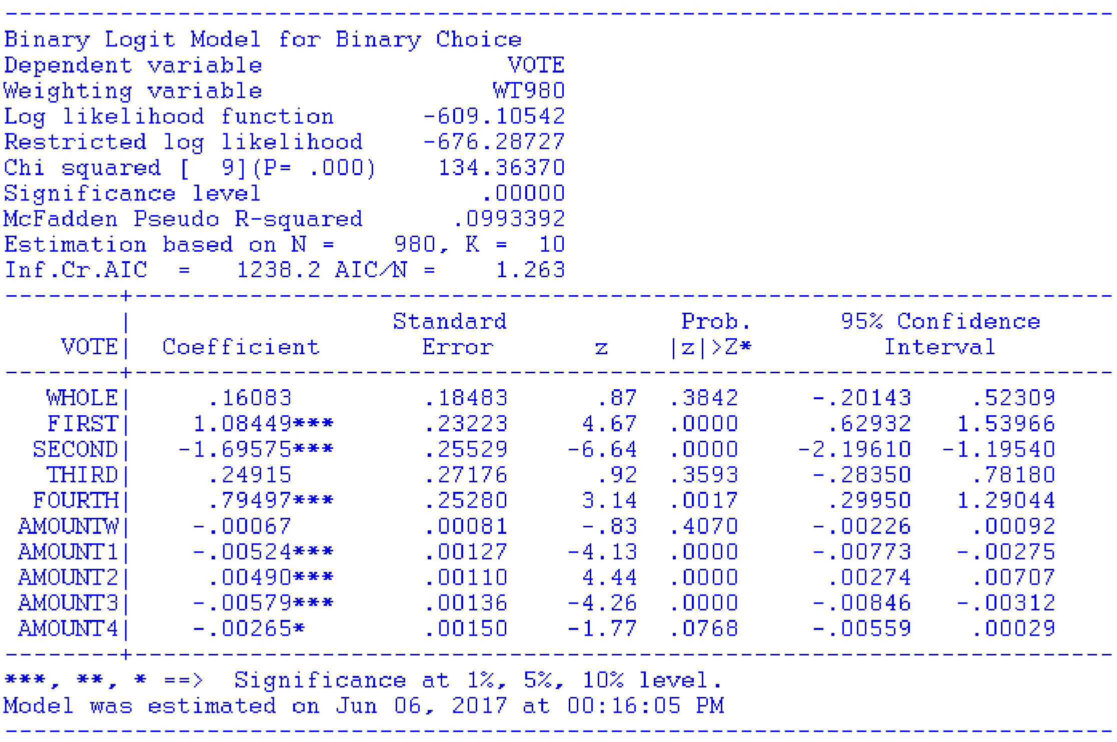
The percentage of yes responses for the 44 respondents who did not answer the age and/or income questions, 66% (n=44), is higher than for the complete case sample, 49% (n=936). Since it appears that this subsample is different than the complete case sample we re-estimate the models in Tables 4 and 5 discarding those who did not answer the age and/or income question. We find that all three of the adding-up tests fail to reject the null hypothesis of equality between WTP for the whole and the sum of the parts for the sample without missing values in age and income. For example, the linear logit model with mean WTP estimated over the positive range is $445 (SE = 193) in the whole scenario and the sum of the WTP parts is $1080 (SE = 174) (see

Figure 3). The 95% confidence intervals for these estimates overlap.

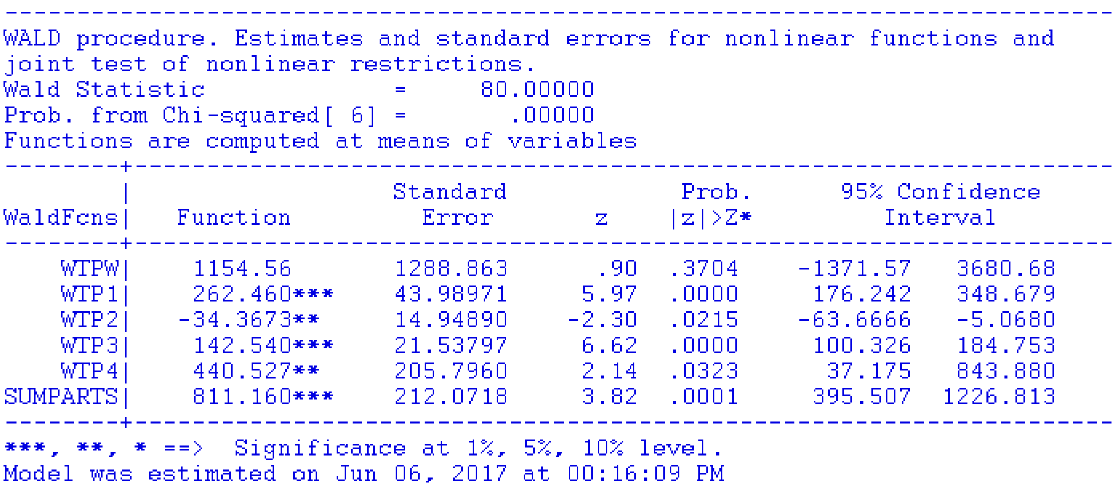
Examination of the income effects estimated by DMT is beyond the scope of this paper. Nevertheless, it is worth mentioning that the unweighted models with the cost coefficient constrained to be equal across scenarios produces statistically insignificant income effects as in DMT (2015). But, applying the post-stratification weights and allowing cost amounts to vary over the scenarios, as is statistically appropriate in this model, leads to statistically significant income effects in the n=980 (with age imputed at the mean), n=950 and n=936 samples. These results suggest that DMT (2015) are using an inappropriate income coefficient for their income effect simulations.

Finally, I test for equivalence of the bid functions (constant and slope) across scenarios with the unweighted data as is used in the paper (Figure 4). The constants are not statistically different from each other for the whole, first and fourth scenarios. The slopes are not statistically different across any of the scenarios. WTP estimates from this model will be equal for the whole, first and fourth scenarios. Naturally, these estimates will not pass the adding-up test. Note that this model does not support the negative results of DMT (2015). These results should be interpreted as a different negative result -- data lacking in divergent validity. For example respondents are willing to pay the same amount for cleanup of the river and lake (whole scenario) and cleanup of just the river (first increment). In this comparison the data do not pass the scope test unless the value of the lake cleanup is equal to zero. However, this is contradicted by an identical WTP for lake cleanup in the fourth increment.

Figure 1. Weighted linear logit model and positive constrained WTP estimates with post-stratification weights (n=980)

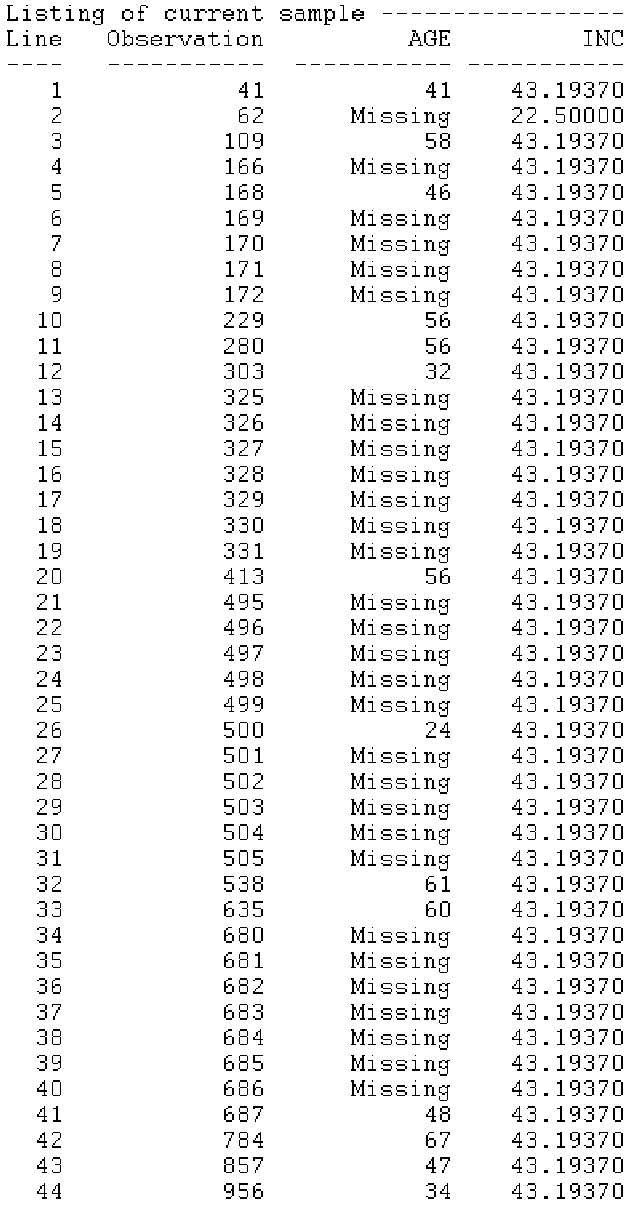


```
-----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                VOTE
Weighting variable               WT980
Log likelihood function      -609.10542
Restricted log likelihood    -676.28727
Chi squared [   9](P= .000)   134.36370
Significance level               .00000
McFadden Pseudo R-squared      .0993392
Estimation based on N =    980, K =  10
Inf.Cr.AIC  =   1238.2 AIC/N =    1.263
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
    VOTE|  Coefficient       Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
   WHOLE|      .16083         .18483      .87  .3842     -.20143     .52309
   FIRST|     1.08449***      .23223     4.67  .0000      .62932    1.53966
  SECOND|    -1.69575***      .25529    -6.64  .0000    -2.19610   -1.19540
   THIRD|      .24915         .27176      .92  .3593     -.28350     .78180
  FOURTH|      .79497***      .25280     3.14  .0017      .29950    1.29044
 AMOUNTW|    -.00067          .00081     -.83  .4070     -.00226     .00092
 AMOUNT1|    -.00524***       .00127    -4.13  .0000     -.00773    -.00275
 AMOUNT2|      .00490***      .00110     4.44  .0000      .00274     .00707
 AMOUNT3|    -.00579***       .00136    -4.26  .0000     -.00846    -.00312
 AMOUNT4|    -.00265*         .00150    -1.77  .0768     -.00559     .00029
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jun 06, 2017 at 00:16:05 PM
-----------------------------------------------------------------------------
```



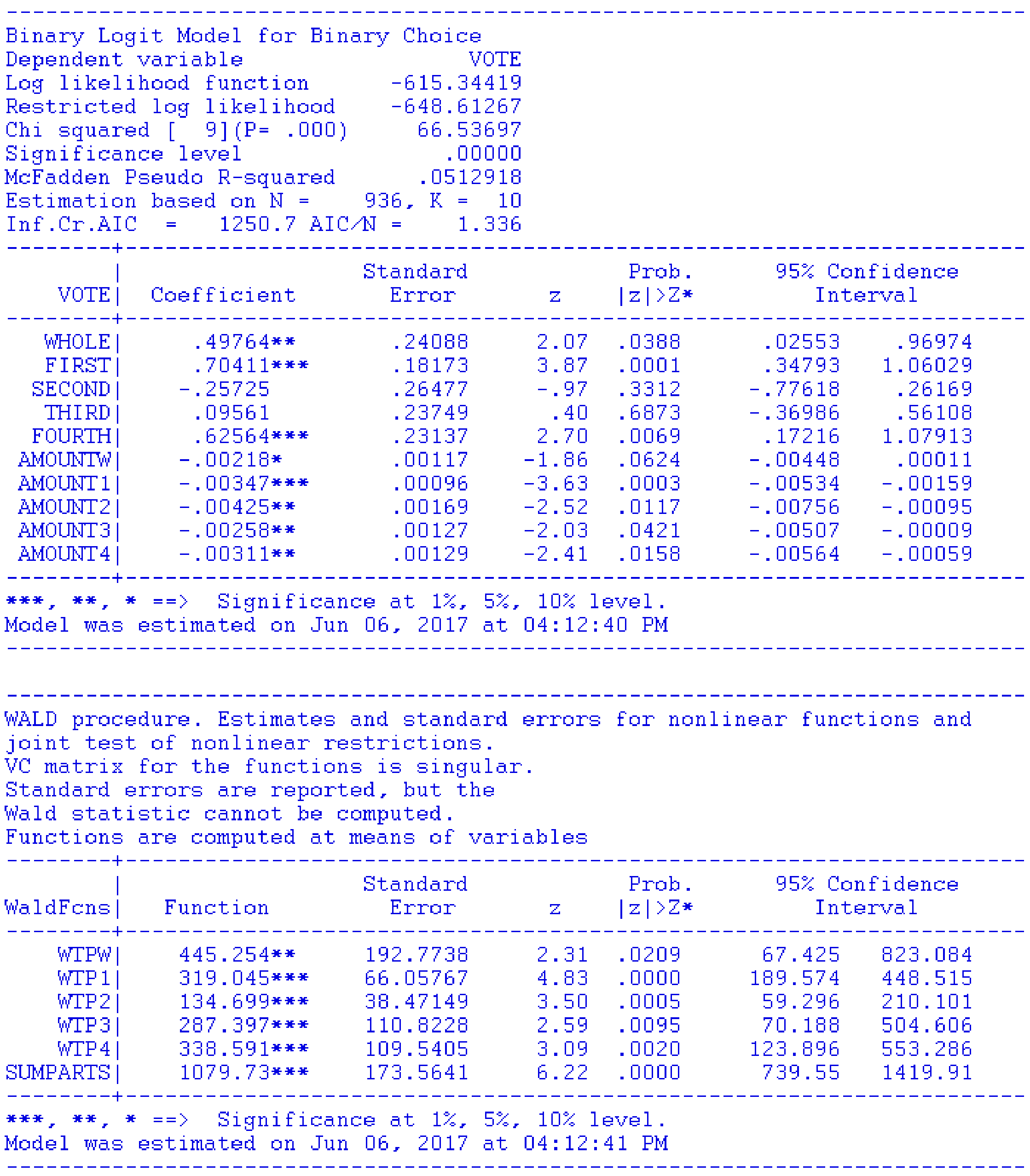
```
-----------------------------------------------------------------------------
WALD procedure. Estimates and standard errors for nonlinear functions and
joint test of nonlinear restrictions.
Wald Statistic                 =     80.00000
Prob. from Chi-squared[ 6] =       .00000
Functions are computed at means of variables
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
WaldFcns|   Function         Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
    WTPW|    1154.56       1288.863      .90  .3704    -1371.57    3680.68
    WTP1|    262.460***    43.98971     5.97  .0000     176.242    348.679
    WTP2|   -34.3673**     14.94890    -2.30  .0215    -63.6666    -5.0680
    WTP3|    142.540***    21.53797     6.62  .0000     100.326    184.753
    WTP4|    440.527**     205.7960     2.14  .0323      37.175    843.880
SUMPARTS|    811.160***    212.0718     3.82  .0001     395.507   1226.813
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jun 06, 2017 at 00:16:09 PM
```

Figure 2. Missing Age and Income values



```
Listing of current sample -----------------
Line   Observation         AGE         INC
----   -----------  ----------- -----------
   1            41           41    43.19370
   2            62      Missing    22.50000
   3           109           58    43.19370
   4           166      Missing    43.19370
   5           168           46    43.19370
   6           169      Missing    43.19370
   7           170      Missing    43.19370
   8           171      Missing    43.19370
   9           172      Missing    43.19370
  10           229           56    43.19370
  11           280           56    43.19370
  12           303           32    43.19370
  13           325      Missing    43.19370
  14           326      Missing    43.19370
  15           327      Missing    43.19370
  16           328      Missing    43.19370
  17           329      Missing    43.19370
  18           330      Missing    43.19370
  19           331      Missing    43.19370
  20           413           56    43.19370
  21           495      Missing    43.19370
  22           496      Missing    43.19370
  23           497      Missing    43.19370
  24           498      Missing    43.19370
  25           499      Missing    43.19370
  26           500           24    43.19370
  27           501      Missing    43.19370
  28           502      Missing    43.19370
  29           503      Missing    43.19370
  30           504      Missing    43.19370
  31           505      Missing    43.19370
  32           538           61    43.19370
  33           635           60    43.19370
  34           680      Missing    43.19370
  35           681      Missing    43.19370
  36           682      Missing    43.19370
  37           683      Missing    43.19370
  38           684      Missing    43.19370
  39           685      Missing    43.19370
  40           686      Missing    43.19370
  41           687           48    43.19370
  42           784           67    43.19370
  43           857           47    43.19370
  44           956           34    43.19370
```

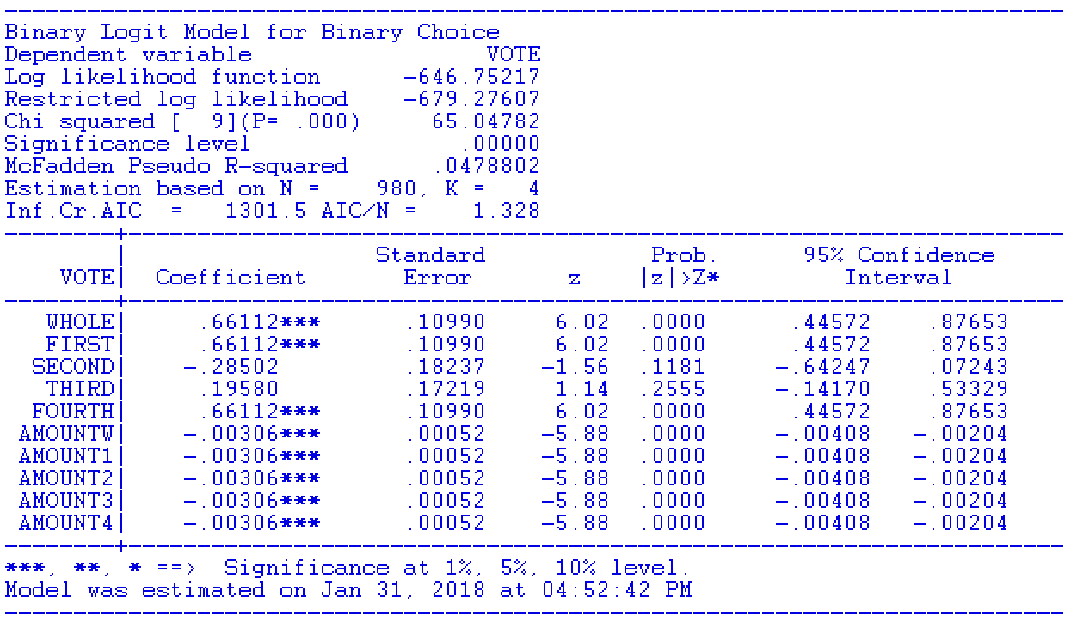
Figure 3. Unweighted linear logit model and positive constrained WTP estimates with the complete case sample (n=936)



```
-----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                 VOTE
Log likelihood function      -615.34419
Restricted log likelihood    -648.61267
Chi squared [   9](P= .000)    66.53697
Significance level               .00000
McFadden Pseudo R-squared      .0512918
Estimation based on N =    936, K =  10
Inf.Cr.AIC  =   1250.7 AIC/N =    1.336
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
    VOTE|  Coefficient       Error      z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
   WHOLE|      .49764**      .24088     2.07  .0388      .02553     .96974
   FIRST|      .70411***     .18173     3.87  .0001      .34793    1.06029
  SECOND|    -.25725         .26477     -.97  .3312    -.77618      .26169
   THIRD|      .09561        .23749      .40  .6873    -.36986      .56108
  FOURTH|      .62564***     .23137     2.70  .0069      .17216    1.07913
 AMOUNTW|    -.00218*        .00117    -1.86  .0624    -.00448      .00011
 AMOUNT1|    -.00347***      .00096    -3.63  .0003    -.00534    -.00159
 AMOUNT2|    -.00425**       .00169    -2.52  .0117    -.00756    -.00095
 AMOUNT3|    -.00258**       .00127    -2.03  .0421    -.00507    -.00009
 AMOUNT4|    -.00311**       .00129    -2.41  .0158    -.00564    -.00059
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jun 06, 2017 at 04:12:40 PM
-----------------------------------------------------------------------------

-----------------------------------------------------------------------------
WALD procedure. Estimates and standard errors for nonlinear functions and
joint test of nonlinear restrictions.
VC matrix for the functions is singular.
Standard errors are reported, but the
Wald statistic cannot be computed.
Functions are computed at means of variables
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
WaldFcns|   Function         Error      z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
    WTPW|    445.254**     192.7738     2.31  .0209      67.425    823.084
    WTP1|    319.045***    66.05767     4.83  .0000     189.574    448.515
    WTP2|    134.699***    38.47149     3.50  .0005      59.296    210.101
    WTP3|    287.397***    110.8228     2.59  .0095      70.188    504.606
    WTP4|    338.591***    109.5405     3.09  .0020     123.896    553.286
SUMPARTS|    1079.73***    173.5641     6.22  .0000      739.55    1419.91
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jun 06, 2017 at 04:12:41 PM
-----------------------------------------------------------------------------
```

Figure 4. Imposition of equality constraints for coefficients in the linear logit with unweighted data and n=980.



```
-----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                VOTE
Log likelihood function      -646.75217
Restricted log likelihood    -679.27607
Chi squared [   9](P= .000)    65.04782
Significance level               .00000
McFadden Pseudo R-squared      .0478802
Estimation based on N =    980, K =   4
Inf.Cr.AIC  =   1301.5 AIC/N =    1.328
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
    VOTE|  Coefficient       Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
   WHOLE|      .66112***      .10990     6.02  .0000      .44572     .87653
   FIRST|      .66112***      .10990     6.02  .0000      .44572     .87653
  SECOND|    -.28502          .18237    -1.56  .1181    -.64247     .07243
   THIRD|     .19580          .17219     1.14  .2555    -.14170     .53329
  FOURTH|      .66112***      .10990     6.02  .0000      .44572     .87653
 AMOUNTW|    -.00306***       .00052    -5.88  .0000    -.00408    -.00204
 AMOUNT1|    -.00306***       .00052    -5.88  .0000    -.00408    -.00204
 AMOUNT2|    -.00306***       .00052    -5.88  .0000    -.00408    -.00204
 AMOUNT3|    -.00306***       .00052    -5.88  .0000    -.00408    -.00204
 AMOUNT4|    -.00306***       .00052    -5.88  .0000    -.00408    -.00204
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
Model was estimated on Jan 31, 2018 at 04:52:42 PM
-----------------------------------------------------------------------------
```