Should We Trust Hypothetical Referenda?
Test and Identification Problems

Fredrik Carlsson[a]
Olof Johansson-Stenman[b]

Working Papers in Economics no. 189
January 2006
Department of Economics
School of Business, Economics and Law
Göteborg University

Abstract
In a paper published in the Journal of Political Economy, Cummings et al. experimentally compare hypothetical and real-money referenda. They reject the incentive compatibility hypothesis of hypothetical referenda. However, in a comment, Haab et al. claim that the hypothesis cannot be rejected if one corrects for heteroskedasticity. In this note we show that the methodology used by Haab et al. is flawed, and their conclusions unwarranted. Our results rather support the original conclusion that hypothetical referenda appears not to resemble real referenda (unless one has reasons to believe that the true variance is much larger in the hypothetical case). This paper outlines design and identification difficulties arising when statistically comparing real and hypothetical referenda.

Keywords: Hypothetical referenda, incentive compatibility, non-market valuation, identification

JEL: C12, C35, Q51

Acknowledgments: We thank Wictor Adamowicz and Joffre Swait for comments and discussions. Financial support from Sida to the Environmental Economics Unit at Göteborg University is acknowledged.
[a] Department of Economics, School of Business, Economics and Law, Göteborg University, Box 640, 405 30 Göteborg, Sweden; Ph +46 31 773 41 74; E-mail fredrik.carlsson@economics.gu.se Department of Economics, Göteborg University, Box 640, SE 405 30 Göteborg, Sweden.
[b] Department of Economics, School of Business, Economics and Law, Göteborg University, Box 640, 405 30 Göteborg, Sweden; Ph +46 31 773 25 38; E-mail olof.johansson@economics.gu.se

1. Introduction

Whether hypothetical referenda, as applied in contingent valuation (CV) studies, are valid in the sense of mimicking real referenda has been intensively debated over the years, and there is still no sign of consensus, or even much indication of convergence. One of the reasons is that it is difficult to test this hypothesis, since the possibility to conduct real referenda is quite limited. One of the exceptions is the study by Cummings, Elliot, Harrison and Murphy (1997) (henceforth CEHM) published in the Journal of Political Economy. They conducted an experiment testing the incentive compatibility of hypothetical referenda by comparing the responses from a hypothetical referendum with the responses from a real money referendum (with exactly the same design) directed toward people who lived close to a contaminated area. The respondents were told that if everybody taking part in the study paid 10 USD, the amount of money aggregated across all individuals would be sufficient to cover the costs to produce and distribute a 'citizens' guide' that provides valuable information regarding safe groundwater. In the hypothetical referendum 45% voted yes and 55% voted no, whereas in the real referendum, 27% voted yes and 73% voted no. This rather sizable difference was also found to be statistically significant, implying that they could reject the hypothesis that the hypothetical referendum is incentive compatible.

However, in a comment, also published in this journal, Haab, Huang and Whitehead (1999) (henceforth HHW) claim that the results of CEHM do not reject the hypothesis of incentive compatibility if one allows for a difference in the scale parameter between the real and hypothetical referendum; in other words, if one takes heteroskedasticity into

account.[1] In this note we outline the identification issues involved in testing the equivalence of hypothetical and real referenda.

We conclude that the methodology used by HHW was inappropriate. Our results instead support the original conclusion by CEHM that there appears to be a difference between the hypothetical and the real referenda, with the caveat that the data used does not allow for an appropriate test of heteroskedasticity.

2. Estimating the relative scale parameter with discrete choice data

With discrete choice data we often do not directly observe the variable of interest. For example in an environmental valuation study we seek the willingness to pay (WTP) for an improvement in environmental quality, but, we only observe if the respondent answers yes or no to a certain bid. However, it is possible to estimate the WTP from the discrete choice data given certain assumptions, including assumptions about the functional form of the WTP (or utility) function and the corresponding error term. Since discrete choice data provide limited information, identification problems can arise. In particular this concerns the identification of the variance of the latent variable, in our case WTP. This becomes a problem when one wants to compare and/or pool different data sets. The problem of heteroskedasticity is also more important in limited dependent variable models, since failure to correct for true underlying heteroskedasticity implies inconsistent parameter estimates, contrary to conventional continuous regression models where such an omission still implies consistent parameter estimates (see for example Yatchew and Griliches, 1985; Kiefer and Skoog, 1984).

[1] This comment has been influential and cited in a number of papers including Cameron et al. (2002), List et al. (2004) and Lusk (2003).

In this respect, the comment by HHW is perfectly valid: it is important to correct for heteroskedasticity and investigate whether the results are robust in this respect. The problem is *how* to test for heteroskedasticity in models of referenda. Largely following the notation of HHW, the willingness to pay for the real and hypothetical referendum are assumed to be:

$$WTP^R = \beta^R + \beta x^R + \varepsilon^R, \tag{1}$$

$$WTP^H = \beta^H + \beta x^H + \varepsilon^H,$$

where β^R and β^H are indicators of the effect of real and hypothetical preference revelation processes, respectively; x is a vector of socio-economic characteristics; β is the corresponding parameter vector; and ε reflects the error terms, which are assumed to be normally distributed with mean zero and standard deviations σ^R and σ^H, respectively. The original test of incentive compatibility by CEHM, obtained by simply pooling the data sets and introducing a dummy variable for the real treatment, can be seen as a test of the following hypothesis:

$$H_{0A} : \frac{\beta^R}{\sigma^R} - \frac{\beta^H}{\sigma^H} = 0, \tag{3}$$

whereas what we want to test is

$$H_{0B} : \beta^R - \beta^H = 0. \tag{4}$$

In order to perform this test we need to account for the possible difference in variance, or scale ($\mu = 1/\sigma$) of WTP between the two data sets. An additional complexity arises in the case of the real referendum process data used by CEHM. It is not possible to estimate the scale parameter for both data sets since the bid is not varied in the real data set (Cameron and James, 1987). However, the relative scale factor, $\sigma = \sigma^R / \sigma^H$, can be

estimated either simultaneously (Louviere, Hensher and Swait, 2000) or with a simple grid search procedure (Swait and Louviere, 1993).

HHW claim that they use the grid search procedure suggested by Swait and Louviere, but they do not, in fact. When estimating the relative scale parameter HHW assume that there is no difference in the willingness to pay between the two data sets, i.e. they estimate the model under the restriction that $\beta^R - \beta^H = 0$. They found that the standard deviation is about 25 times larger in the hypothetical case. They then impose the estimated scale parameter from this grid search on a model where it is tested whether $\beta^R - \beta^H$ is significantly different from zero or not. This is clearly not the process outlined by Swait and Louviere; the scale parameter is estimated conditional on a particular model specification. One cannot estimate a scale parameter for a particular model specification and impose it on another specification. Furthermore, a well-known problem is that it is difficult to distinguish between a test for heteroskedasticity and a test for misspecification (Greene, 2002; Davidson and MacKinnon, 1984). The approach used to estimate the scale parameter is particularly problematic because the variable not included in the grid search is exactly the variable we normalize the variance for one of the data sets on. This implies that this variable is very closely correlated with the scale parameter to be estimated, and omission of a variable that is correlated with one of the included variables will lead to inconsistent estimates (Kiefer and Skoog, 1984).

It is perhaps not very surprising that large biases can occur due to mis-specifications if the dataset is small and the explanatory variables have no or low explanatory power, as in this case. A stronger test is whether the large biases prevail in a situation with a large dataset of high quality. To examine this issue we simulate a well-behaved and

large data set to see if the procedure used by HHW would yield biased results, or not. Suppose that we have two groups of individuals, with 10,000 individuals in each group. The WTP in each group is normally distributed with the following functional forms:

$$WTP^R = \beta^R + \beta_1 x_1 + \beta_2 x_2 + \varepsilon^R \qquad (1)$$

$$WTP^H = \beta_1 x_1 + \beta_2 x_2 + \varepsilon^H$$

with scale parameters σ^R and σ^H, where the variables x_1 and x_2 are independent; x_1 has a normal distribution with mean 1 and standard deviation 1, whereas x_2 is a discrete variable that is either 0 or 1 with probability 0.5. Furthermore, we assume the true parameter values to be $\beta^R = -1$, $\beta_1 = -0.5$ and $\beta_2 = 0.5$. To correspond with the discrete choice nature of CV data, an indicator function is defined such that the individual votes "Yes" if WTP>0 and "No" otherwise. To illustrate the strategy of HHW we estimate the relative scale parameter under the incorrect assumption that $\beta^R = 0$.[2] The results of these estimations, with and without a correct assumption about the true underlying WTP function are reported in Table 1. In the first case we assume that there is no true difference in scale, and that $\sigma^R = \sigma^H = \sigma = 1$, while in the second case we assume that $\sigma^R = 1, \sigma^H = 2$, and consequently the relative scale factor is $\sigma = 0.5$.

<<Table 1 around here>>

As expected, when the WTP function is mis-specified (by not including the β^R parameter), the estimated scale parameter becomes highly biased. Thus, by not correctly

[2] Since a simultaneous estimation of the scale parameter and the other WTP parameters gives (by definition) the same estimate as a grid search, but provides additional statistical information that can be used to deduce the significance of the relative scale parameter, we use this method.

specifying the model one can obtain a severely biased parameter estimate even for an otherwise ideal dataset such as the one generated for Monte Carlo analysis here.[3]

Let us therefore look at the original results of CEHM. In the second column of Table 2, the original results of their pooled model are reported. What immediately becomes clear is that their estimated WTP function does not explain much of the variation in WTP. Actually the only significant parameter is the one associated with the dummy variable for the real referendum.[4] In principle, this difference between the two data sets need then not reflect differences in WTP, but could instead reflect differences in scale parameters. However, if there are no other significant parameters, as in the CEHM case, the informational basis for identifying a relative scale difference is very weak. Indeed, when we estimate the scale parameter and at the same time allow for a difference in WTP by including the dummy variable for the real referendum, the likelihood function with this data is actually monotonically increasing in the value of the relative scale parameter – a clear sign of a poorly identified model arising from the lack of variation in the data and the confound between scale and preference parameters. In the last column in Table 2 we report the results of the estimations when the scale parameter is arbitrarily set to 10,000. These results, albeit still suffering from the identification problem, suggest a conclusion quite the opposite of the results of HHM: the dummy variable for the real treatment is highly significant. More importantly, the scale parameter is still not statistically different from 1, even at the 10 % level.

[3] Note that the opposite does not hold in our case. If $\beta^R = 0$ is the correct specification the other parameters will not be biased if we include that variable in the estimations.

[4] In a likelihood ratio test we cannot reject the hypothesis that all parameters expect the intercept is zero (p-value = 0.24).

<<Table 2 around here>>

Thus, we have shown that the parameter reflecting real referendum is statistically significant both at the CEHM assumption of a relative scale equal to unity (i.e. the homoskedastic case), and at large relative scale parameter values. Moreover, this grid search reveals that the parameter reflecting real referendum is significant at the 10% level for all values of the scale parameter above 0.57.

Hence, if we knew that the variance was much larger in the hypothetical case, then there would be no statistical difference. Intuitively, if there is a minority of the respondents with a true WTP above the bid, then if the voters make more errors there will be an increased number of yes responses, keeping the underlying preferences with respect to WTP fixed.[5] Thus, whether one believes the original CEHM conclusions or not would then depend on the perception one may have regarding the relative variance between the samples.

3. Conclusions

The issue raised by HHM is a very important and relevant issue in the measurement of value in referendum models. When comparing models that employ different design mechanisms (e.g. hypothetical versus real payments) the comparison will necessarily involve a comparison of preferences and scale. It is in principle possible to

[5] This is perhaps most obvious in the extreme case where there are no respondents with a true WTP above the bid. Then if we let people make mistakes through an additive error term, there will be some respondents who will vote yes, but no respondents with a true WTP below the bid who will now vote no (since there are no respondents with such preferences). This asymmetry is what results in heteroskedasticity being a much more severe problem in this type of bounded discrete choice model, compared to a linear regression model.

simultaneously estimate relative scale parameters and preference parameters. However, when there is no variation in the bid-variable, and no other significant explanatory variables as in CEHM, to simultaneous identify relative scale parameters and shift variables becomes essentially impossible.

In this note we have made five fundamental conclusions: 1. HHW estimated the scale parameter in an inappropriate way, since they did not estimate it based on the same model as the one that was used to estimate the real treatment parameter. 2. Using simulations we demonstrate why their methodology generates inaccurate and misleading results. We create a well-behaved data set with different underlying "true" probability levels in the different groups, and where there is no true underlying heteroskedasticity, implying that the true relative scale parameters are equal to unity. Using the methodology by HHW on this data produces the same type of large biases as in the HHW paper. 3. While the identification problem remains, the real referendum parameter is statistically significant for a large range of values of the scale parameter in the heteroskedastic model when the parameters are estimated appropriately. This tends to support the original conclusion by CEHM that the probability of accepting the bid is smaller with real money compared to the hypothetical case (unless one has reason to believe that the true variance is much larger in the hypothetical case). 4. When designing experiments such as the one examined here, the potential for heteroskedasticity or scale effects should be recognized. Designing experiments with limited variability in price will preclude the possibility of simultaneously estimating preference and scale parameters. In hypothetical experiments, the experimenter usually has control over the number of bids. The problem may occur with the real experiments. In our opinion it is important to allow for a variation in price (bid) also in the real

experiment if one wants to control for heteroskedasticity when comparing real and hypothetical responses. A corollary is that one should generally be careful when estimating relative scale parameters when the underlying data have limited variation. The grid search approach has clear computational advantages, but should be avoided for data of the type used here where there is no price variation in the bid, and when the other explanatory variables have poor explanatory power.

Several recent papers (e.g. Cameron et al, 2002; Vossler and Kerkvleit, 2003) have examined the difference between hypothetical and real response formats, accounting for scale differences between the response formats. However, as shown here the identification problem precludes easy assessment of preference and scale difference between data formats unless the data are well conditioned. Research is required to address this complex problem.

References

Cameron, T. and M. James (1987) Efficient Estimation Methods for 'Closed-Ended' Contingent Valuation Survey, Review of Economics and Statistics 69, 269-276.

Cameron, T., G. Poe, R. Ethier and W. Schulze (2002) Alternative Non-market Value-Elicitation Methods: Are the Underlying Preferences the Same?, Journal of Environmental Economics and Management 44, 391-425.

Cummings, R., S. Elliot, G. Harrison and J Murphy (1997) Are Hypothetical Referenda Incentive Compatible, Journal of Political Economy 105, 609-621.

Davidson, R. and J. MacKinnon (1984) Convenient Specification Tests for Logit and Probit Models, Journal of Econometrics 25, 241-262.

Greene, W. (2000) Econometric Analysis. New Jersey: Prentice-Hall.

Haab, T., J.-C. Huang and J. Whitehead (1999) Are Hypothetical Referenda Incentive Compatible? A Comment, Journal of Political Economy 107, 186-196.

Islam, T. and J. Louviere (2004) Modeling the Effects of Including/Excluding Attributes in Choice Experiments on Systematic and Random Components, Unpublished Working Paper.

Kiefer, N. and G. Skoog (1984) Local Asymptotic Specification Error Analysis, Econometrica 52, 873-886.

List, J., R. Berrens, A. Bohara and J. Kerkvilet (2004) Examining the Role of Social Isolation on Stated Preferences, American Economic Review 94, 741-752.

Louviere, J., D. Hensher and J. Swait (2000) Stated Choice Methods. Cambridge: Cambridge University Press.

Lusk, J. (2003) Effect of Cheap Talk on Consumer Willingness-to-pay for Golden Rice, American Journal of Agricultural Economics 85, 840-856.

Swait, J. and J. Louviere (1993) The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models, Journal of Marketing Research 30, 305-314.

Vossler, C.A. and J. Kerkvleit (2003) A criterion validity test of the contingent valuation method: comparing hypothetical and actual voting behaviour for a public referendum. Journal of Environmental Economics and Management 45, 631-649.

Yatchew, A. and Z. Griliches (1985) Specification Error in Probit Models, Review of Economics and Statistics 67, 134-139.

Table 1. Results of estimations based on simulated data; p-values in parentheses.

	Case 1 ($\sigma = 1$)		Case 2 ($\sigma = 0.5$)	
	$\beta^R = 0$ (incorrect spec.)	$\beta^R \neq 0$ (correct spec.)	$\beta^R = 0$ (incorrect spec.)	$\beta^R \neq 0$ (correct spec.)
Common intercept	-1.063 (0.000)	-0.018 (0.513)	-1.069 (0.000)	-0.006 (0.894)
β^R		-1.042 (0.000)		-1.056 (0.000)
β_1	0.591 (0.000)	0.524 (0.000)	0.543 (0.000)	0.524 (0.000)
β_2	-0.483 (0.000)	-0.445 (0.000)	-0.446 (0.000)	-0.442 (0.000)
σ	0.140 (0.000)	0.962 (0.000)	0.057 (0.003)	0.502 (0.000)
P($\sigma \neq 1$)	0.000	0.488		
P($\sigma \neq 0.5$)			0.000	0.968

Table 2. Specifications of referendum preference and relative scale parameters; p-values in parentheses.

	Naïve pooling (CEHM)	HHM specification		Simultaneous estimation of relative scale and preference parameters
		Step 1	Step 2	
Constant	-0.6845 (0.772)	-3.5477 (0.497)	-4.7655 (0.431)	-0.000005 (0.985)
Real referendum	-0.4925 (0.014)		1.2073 (0.695)	-0.5791 (0.000)
RH	0.1556 (0.533)	-2.7818 (0.861)	-1.5849 (0.736)	0.00003 (0.376)
Age	0.0129 (0.156)	0.0114 (0.475)	0.0114 (0.476)	0.000002 (0.218)
Sex	-0.0906 (0.594)	-0.2908 (0.399)	-0.2891 (0.401)	0.000001 (0.947)
Race	-0.0209 (0.912)	0.3958 (0.414)	0.3951 (0.414)	-0.00002 (0.476)
Income	-0.0026 (0.596)	0.0062 (0.537)	0.0061 (0.540)	-0.0000005 (0.376)
Married	0.1251 (0.539)	-0.1506 (0.718)	-0.1531 (0.713)	0.00002 (0.351)
Earn	0.0053 (0.978)	0.1853 (0.656)	0.1854 (0.656)	-0.000005 (0.806)
Number	-0.0012 (0.949)	-0.0115 (0.820)	-0.0103 (0.838)	0.0000004 (0.844)
Student	0.1630 (0.586)	0.2892 (0.561)	0.2847 (0.567)	0.000003 (0.945)
Scale		0.0405 (0.814)		10000
Log-L	-178.276	-178.986	-178.909	- 177.287