CREI
Centro di **R**icerca
Interdipartimentale
di **E**conomia delle **I**stituzioni

CREI Working Paper no. 3/2014

WILLINGNESS TO PAY CONFIDENCE INTERVAL ESTIMATION METHODS: COMPARISONS AND EXTENSIONS

by

Valerio Gatta
University of Roma Tre

Edoardo Marcucci
University of Roma Tre

Luisa Scaccia
University of Macerata

available online at http://host.uniroma3.it/centri/crei/pubblicazioni.html

ISSN 1971-6907

Willingness to pay confidence interval estimation methods: comparisons and extensions

Valerio Gatta[a], Edoardo Marcucci[b], Luisa Scaccia[c]

[a] *Dipartimento di Scienze Politiche, CREI, Università di Roma Tre, Via G. Chiabrera, 199, 00145 Roma, Italy,*
e-mail: valerio.gatta@uniroma3.it
[b] *Dipartimento di Scienze Politiche, CREI, Università di Roma Tre, Via G. Chiabrera, 199, 00145 Roma, Italy,*
e-mail: emarcucci@uniroma3.it
[c] *Università di Macerata, Dipartimento di Economia e Diritto, via Crescimbeni 20 - 62100 Macerata,*
e-mail: scaccia@unimc.it

Abstract

This paper systematically compares methods to build confidence intervals for willingness to pay measures in a discrete choice context. It contributes to the literature by including methods developed in other research fields. Monte Carlo simulations are used to assess the performance of all the methods considered. The various scenarios evaluated reveal a certain skewness in the estimated willingness to pay distribution. This should be reflected in the confidence intervals. Results show that the commonly used Delta method, producing symmetric intervals around the point estimate, often fails to account for such a skewness. Both the Fieller method and the likelihood ratio test inversion method produce more realistic confidence intervals. Some bootstrap methods also perform reasonably well. Finally, empirical data are used to illustrate an application of the methods considered.

Keywords: Confidence intervals; willingness to pay; discrete choice models; elasticities; standard errors.

---

1. Introduction

Willingness to pay (WTP) is the amount of money an agent would pay to obtain a desired good or service. The derivation of reliable WTP measures is fundamental in transportation economics and in other applied fields. WTP considerations are relevant for: travel time savings

(Hensher, 2010); travel time reliability (Li et al., 2010); transport externalities (Ortúzar et al., 2000); accident risk reduction and value of life (Iraguen and Ortúzar, 2004; Guria et al., 2005); information technologies (Molin and Timmermans, 2006); residential location (Jara-Diaz and Martinez, 1999)[1].

In a choice modeling framework, typically assuming linear-in-attributes utility functions, the WTP for a given attribute is obtained dividing its coefficient by that of cost. Since model estimation yields an estimate of the true coefficients, the computed WTP (i.e. $\widehat{WTP}$) is itself an estimate with a given probability distribution. Thus, it is desirable to calculate confidence intervals (CIs), in addition to point estimates. This is not trivial since the exact distribution of the WTP estimator is not known. When maximum likelihood estimates (MLEs) are used for the coefficients, the distribution of WTP is the ratio between two correlated, asymptotically normal distributions. The distribution of the ratio of two normal variables has been derived by Fieller (1932) and Hinkley (1969), and shown to be approximately normal when the coefficient of variation of the denominator variate is negligible. More recently, Daly et al. (2012a) showed that $\widehat{WTP}$ is itself a MLE and, thus, asymptotically normal. Also Daly et al. (2012b) study WTP distribution and provide conditions for the finiteness of its moments under different cost distributions in random coefficient models.

Notwithstanding the relevant results obtained by Daly et al. (2012a) with respect to the asymptotic properties of $\widehat{WTP}$, its finite sample distribution can be substantially different from the normal distribution. This motivates the development of different methods to calculate CIs for WTP. For example, the Delta method assumes normally distributed $\widehat{WTP}$. Alternatively, Fieller (Fieller, 1940, 1954; Bolduc et al., 2010) and likelihood ratio test inversion methods (Armstrong et al., 2001), only rely on the normality of the coefficients involved in

[1]The list of the subjects reported reflects the seven most cited articles in ISI WEB OF SCIENCE database (accessed on 29th October 2014) resulting from a search using "willingness to pay" as a keyword for *Title* jointly with "transport" for *Topic*.

the ratio. Other methods use bootstrap sampling techniques, thus avoiding any distributional assumption (Efron and Tibshirani, 1993; Davison and Hinkley, 1997).

Only few studies compare methods to construct CIs for WTP. Armstrong et al. (2001) investigate the potentialities of likelihood ratio test inversion method only on real data. Hole (2007) proposes a Monte Carlo study to assess the performance of Delta, Fieller and some bootstrap methods. Bolduc et al. (2010) focus on the advantages of Fieller method when the coefficient in the denominator approaches 0. Hirschberg and Lye (2010) compare the Delta and Fieller methods from a geometrical point of view. The conclusions reached by these studies are not always in accordance and, to the best of our knowledge, a comparison of all the existing methods does not exist.

This paper provides some guidelines for choosing, under different critical conditions, an appropriate method to construct CIs for WTP. It contributes to the literature by comprehensively and systematically comparing all the methods investigated in the discrete choice field, as well as other methods borrowed from different research areas. The comparison is carried out through a Monte Carlo study. Data are simulated under different scenarios mimicking real situations in which the $\widehat{WTP}$ distribution is potentially highly skewed and far from normal. Two real data sets are used to illustrate the practical relevance of the issues raised in the simulation study.

The paper is structured as follows: Section 2 describes WTP estimation within a choice modeling context; Section 3 illustrates the main assumptions, advantages and disadvantages of various methods for CI estimation; Section 4 compares methods through a Monte Carlo study; Section 5 reports the results from real data applications; Section 6 concludes and suggests some general guidelines.

2. Logit models and WTP estimation

Consider a sample of N decision makers, facing J alternatives, in T choice experiments. A choice performed by individual n, for $n = 1, \dots, N$, can be modeled, in a random utility framework, as follows:

$$y_{int} = \begin{cases} 1 & \text{if } U_{int} \geq U_{jnt} \text{ for } j = 1, \ldots, J \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$U_{int} = V_{int} + \epsilon_{int} \tag{2}$$

is the unobservable utility that individual n derives from alternative i (for $i = 1, \ldots, J$), in choice experiment t (for $t = 1, \ldots, T$), V_{int} is the observable utility and ϵ_{int} is an error term. Observable utility is generally assumed linear-in-the-attributes so that

$$V_{int} = X_{int}\beta, \tag{3}$$

where X_{int} is a $(1 \times K)$ vector of attributes and β is a $(K \times 1)$ vector of coefficients. The choice probability associated with the alternative i chosen by individual n in choice experiment t, is defined as:

$$P_{int} = P\left(U_{int} \geq U_{jnt}, \text{ for } j = 1, \ldots, J\right).$$

Different model specifications can be derived from (2), depending on the assumptions made on the error term. For example, assuming that the error vector ϵ_n, obtained by stacking the vectors $\epsilon_{nt} = (\epsilon_{1nt} \cdots \epsilon_{Jnt})$, is independent, identically Gumbel distributed, leads to the well known Multinomial Logit (MNL) model, for which P_{int} can be analytically determined.

From now on, all subscripts, unless strictly necessary, are dropped to lighten notation and utility is simply denoted as $U = V + \epsilon$. When utility is specified as in (3), the total derivative of U with respect to changes in the k-th attribute X_k and the cost attribute X_C is given by $dU = \beta_k dX_k + \beta_C dX_C$. Setting this expression equal to 0 and solving for $dX_C = dX_k$ yields the change in cost that keeps utility unchanged, given a variation in X_k:

$$\frac{dX_C}{dX_k} = WTP_k = -\frac{\beta_k}{\beta_C},$$

representing the WTP_k for an improvement in X_k.

Dropping the subscript for WTP, a point estimate is calculated as follows:

$$\widehat{WTP} = -\frac{\hat{\beta}_k}{\hat{\beta}_C}, \tag{4}$$

where $\hat{\beta}_k$ and $\hat{\beta}_C$ are the MLEs of β_k and β_C, respectively, which are asymptotically normally distributed, as well as $\widehat{WTP}$ (Daly et al., 2012a). $\widehat{WTP}$ distribution is needed for constructing CIs. As stressed, despite its asymptotic normality, finite $\widehat{WTP}$ distribution can be heavily skewed and relevant for practical purposes. The uncertainty existing on finite $\widehat{WTP}$ distribution gives rise to various methods for constructing CIs.

3. Methods to construct WTP confidence intervals

This section illustrates, for each method, the procedure to construct CIs, the assumptions made, the pros and cons. Figure 1 shows all the methods considered classifying them into two sets (approximation *vs.* simulation) and three families (pivotal, percentile and test inversion).

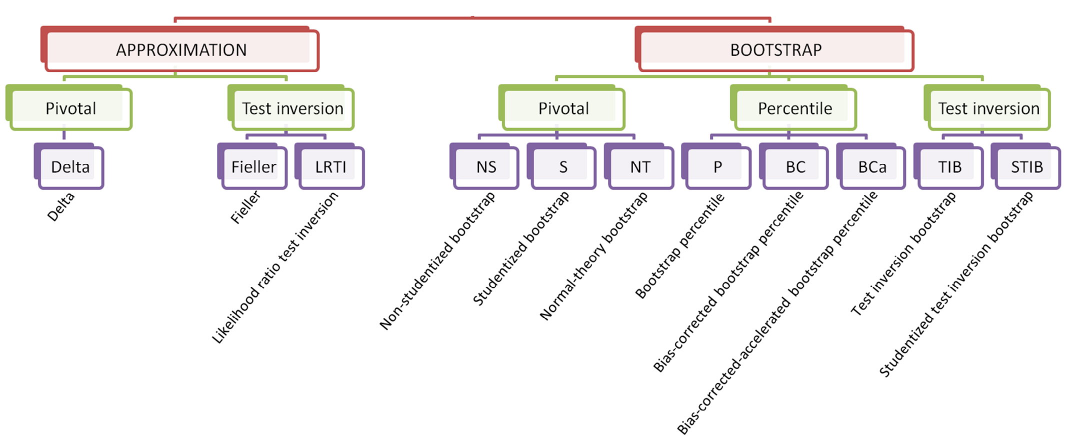

Figure 1: Classification of methods to build WTP confidence intervals

The distinction between approximation and simulation depends on the use of either an analytic or simulated distribution of $\widehat{WTP}$.

The methods belonging to the pivotal family use a pivotal function of $\widehat{WTP}$ and the percentiles of its analytic or simulated distribution to construct CIs. Percentile methods directly consider the simulated distribution of $\widehat{WTP}$ and its percentiles. Finally, the test inversion methods exploit the duality between hypothesis testing and CIs.

Eleven methods are illustrated. Nine have already been used in the choice modeling literature, while the remaining, derived from different research contexts, have not.

3.1. Approximation methods

The three methods hereby described are based on the analytic distribution of $\widehat{WTP}$. The first belongs to the pivotal family and the others to the test inversion one.

3.1.1. Delta method

The first method discussed is the Delta method (Delta) due to its widespread adoption given it is simple and often incorporated in commercial software packages. Delta relies on the normality assumption of MLEs coefficients *and* their ratio. $\widehat{WTP}$ is asymptotically normal and its variance is obtained by taking a first order Taylor expansion around the mean of the variables involved in the ratio and estimating the variance for this expression, i.e.

$$\widehat{WTP} \sim N\left(-\frac{\beta_k}{\beta_C}; \mathrm{var}(\widehat{WTP})\right),$$

where

$$\begin{aligned}
\mathrm{var}(\widehat{WTP}) &= (\widehat{WTP}_{\beta_k})^2\hat{\sigma}^2_{\hat{\beta}_k} + (\widehat{WTP}_{\beta_C})^2\hat{\sigma}^2_{\hat{\beta}_C} + 2\widehat{WTP}_{\beta_k}\widehat{WTP}_{\beta_C}\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C} = \\
&= (-1/\hat{\beta}_C)^2\hat{\sigma}^2_{\hat{\beta}_k} + (\hat{\beta}_k/\hat{\beta}^2_C)^2\hat{\sigma}^2_{\hat{\beta}_C} + 2(-1/\hat{\beta}_C)(\hat{\beta}_k/\hat{\beta}^2_C)\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C},
\end{aligned}$$

where $\widehat{WTP}_{\beta_k}$ and $\widehat{WTP}_{\beta_C}$ are the partial derivatives of $\widehat{WTP}$ with respect to β_k and β_C, evaluated at the MLEs, and with $\hat{\sigma}^2_{\hat{\beta}_k}$, $\hat{\sigma}^2_{\hat{\beta}_C}$ and $\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C}$ representing, respectively, the estimated variances and covariance of $\hat{\beta}_k$ and $\hat{\beta}_C$.

The CI's lower and upper bounds at the $(1-\alpha)$-level are:

$$WTP_L = \widehat{WTP} - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})} \qquad \text{and} \qquad WTP_U = \widehat{WTP} + z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})} \tag{5}$$

where $z_{\alpha/2}$ indicates the $100(1-\alpha/2)th$ percentile of the standard normal density.

Daly et al. (2012a) show that the standard errors obtained using Delta are correct estimates of $\widehat{WTP}$ accuracy characterized by full maximum likelihood properties. Delta generally produces narrow CIs since it delivers errors achieving the Cramér-Rao lower bound. However, this holds only for continuous functions and β_C can never be equal to 0. Bolduc et al. (2010) show that CIs' effective coverage rate, when using Delta, rapidly deteriorates, independently of sample size, as β_C gets closer to 0. Finney (1971) suggests to use Delta whenever the t-statistic of the coefficient at the denominator, t_c, is above 8.75. Marsaglia (2006), on the other hand, considering also the possible correlation among coefficients, suggests a more stringent bound for the ratio variable requiring t_c to be greater than 4 and $(t_b - \rho t_c)/(1-\rho^2)^{0.5}$ be less than 2.26, where t_b is the t-statistic for the numerator. Note that increasing sample size does not guarantee meeting this condition. In addition, this method always produces symmetric CIs around $\widehat{WTP}$ point estimates. This might represent a serious drawback since, has shown in practice, the finite sample $\widehat{WTP}$ distribution is often non-symmetric and far from normal (Armstrong et al., 2001).

3.1.2. Fieller method

The Fieller method[2] (Fieller) exploits the duality between CIs and hypothesis testing. This method makes no assumption on $\widehat{WTP}$ distribution as Delta does, assuming normality only for estimates of attributes' coefficients. This represents a considerable advantage in all those cases where the normality assumption for $\widehat{WTP}$ might not hold. Moreover, Fieller does not present discontinuity points, as for Delta in $\beta_C = 0$, and CIs are defined for all β_C. However, some computational effort is required.

The asymptotic t-test is generally used to check whether a parameter, whose estimator is normally distributed, is significantly different from 0. Ben-Akiva and Lerman (1985) extend

[2]Also known as asymptotic t-test inversion method.

this test to a linear combination of parameters. Recalling (4) and postulating:

$$H_0 : \beta_k + WTP\beta_C = 0. \tag{6}$$

one can derive the following test statistic (Garrido and Ortúzar, 1993):

$$T(WTP) = \frac{\hat{\beta}_k + WTP\hat{\beta}_C}{\sqrt{WTP^2\hat{\sigma}^2_{\hat{\beta}_C} + 2WTP\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C} + \hat{\sigma}^2_{\hat{\beta}_k}}}. \tag{7}$$

Under the null hypothesis, (7) is asymptotically standard normal. The CI for WTP is given by the set of WTP values for which it is not possible to reject H_0 at a predetermined significance level. Thus, the $(1-\alpha)$-level interval corresponds to the WTP_0 values such that $|T(WTP_0)| \leq z_{\alpha/2}$ or equivalently $T^2(WTP_0) \leq z^2_{\alpha/2}$. Garrido and Ortúzar (1993) derive upper and lower bounds of the CI for WTP and Bolduc et al. (2010) extend the result to the simultaneous CI case. Upper and lower bounds are obtained solving the following second-degree-polynomial inequality for WTP_0: $A(WTP_0)^2 + 2B(WTP_0) + C \leq 0$, where

$$A = \hat{\beta}^2_C - z^2_{\alpha/2}\hat{\sigma}^2_{\hat{\beta}_C} \quad , \quad B = \hat{\beta}_k\hat{\beta}_C - z^2_{\alpha/2}\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C} \quad , \quad C = \hat{\beta}^2_k - z^2_{\alpha/2}\hat{\sigma}^2_{\hat{\beta}_k}. \tag{8}$$

One can compute CIs using the following algorithm:

1. fit the model and obtain MLEs of the parameter vector β along with its variance-covariance matrix;
2. compute A, B and C as in (8) and let $\Delta = B^2 - AC$;
3. calculate the interval as:

$$\begin{array}{rl} [WTP_L\ ;\ WTP_U] & \text{if } \Delta > 0 \text{ and } A > 0 \\ (-\infty\ ;\ WTP_L]\bigcup[WTP_U\ ;\ \infty) & \text{if } \Delta > 0 \text{ and } A < 0 \\ (-\infty\ ;\ \infty) & \text{if } \Delta < 0 \text{ (which implies } A < 0) \end{array} \tag{9}$$

where $WTP_L = \dfrac{-B - \sqrt{\Delta}}{A}$ and $WTP_U = \dfrac{-B + \sqrt{\Delta}}{A}$.

Notice that the CI in (9) can be bounded or unbounded (including the entire real line). The unbounded solution occurs if $|\hat{\beta}_C/\hat{\sigma}_{\hat{\beta}_C}| \leq z_{\alpha/2}$, i.e. when β_C is not significantly different from

0 at level α. Fieller coverage rate does not deteriorate as β_C approaches 0. Notice, also, that the bounded CI in (9) is, in general, not symmetric around $\widehat{WTP}$. In fact, the interval's mid-point is usually greater than $\widehat{WTP}$. The CI becomes progressively symmetric when $\hat{\sigma}^2_{\hat{\beta}_C}/\hat{\beta}_C$ and $\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C}/\hat{\beta}^2_C$ tend to 0. In presence of asymmetrically distributed WTP, Fieller is likely to yield more accurate CIs than Delta. Asymptotically, the two methods produce the same interval endpoints (Bolduc et al. (2010)).

3.1.3. Likelihood ratio test inversion method

The likelihood ratio test inversion method (LRTI) is similar to Fieller since it also takes advantage of the duality between CIs and hypothesis testing. They share similar assumptions and have equivalent implications. The likelihood ratio test for the null hypothesis in (6) compares the likelihood of the unrestricted model to that of the restricted, with the restriction being that imposed under the null hypothesis. The test statistic is:

$$LR = -2[l(\hat{\beta}^R) - l(\hat{\beta})], \tag{10}$$

where $l(\hat{\beta}^R)$ and $l(\hat{\beta})$ represent the logarithm of the likelihood at the MLEs for the restricted and unrestricted models, respectively. Under the null hypothesis, the statistic is distributed χ^2 with one degree of freedom, corresponding to the single linear restriction $\beta_k + WTP\beta_C = 0$. Inverting the test statistic (10) to obtain a CI for WTP, requires a search for the maximum and minimum values of WTP for which $-2[l(\hat{\beta}^R) - l(\hat{\beta})] \leq \chi^2_{1,\alpha}$. The following algorithm (Armstrong et al., 2001) can be used to compute WTP_L (similarly for WTP_U). First, fit the model to the unconstrained systematic utility function

$$V = \beta_k X_k + \beta_C X_C + \sum_{h=1}^{K} \beta_h X_h \tag{11}$$

and obtain MLEs $\hat{\beta}$, the corresponding $\widehat{WTP}$ and the unrestricted log-likelihood $l(\hat{\beta})$. Then, initialize the algorithm by letting Inf $= \widehat{WTP} - \lambda$, with λ being a sufficiently large positive value, Sup $= \widehat{WTP}$, Tol $= 1,000$ and ϵ be an arbitrarily small tolerance limit. Perform the following steps until Tol $> \epsilon$:

1. let $WTP_L = \dfrac{\text{Inf} + \text{Sup}}{2}$;
2. fit the constrained model using the constrained utility function

$$V_{\text{con}} = \beta_C(-WTP_L X_k + X_C) + \sum_{h=1}^{K} \beta_h X_h, \tag{12}$$

   obtain restricted MLEs and restricted log-likelihood and then calculate LR as in (10);
3. if $LR < \chi^2_{1,\alpha}$, let Sup $= WTP_L$ and $WTP_L = \dfrac{\text{Inf} + \text{Sup}}{2}$, otherwise if $LR > \chi^2_{1,\alpha}$ let Inf $= WTP_L$ and $WTP_L = \dfrac{\text{Inf} + \text{Sup}}{2}$;
4. set Tol $= |LR - \chi^2_{1,\alpha}|$.

When the algorithm stops, the last WTP_L value is the lower bound of the interval.

In addition to the advantages of Fieller, the usage of LRTI is not restricted to linear utility functions. A drawback is the iterative procedure needed to obtain each interval limit which makes it computationally more demanding than Fieller method while much less intensive than any bootstrap method.

3.2. Bootstrap methods

Bootstrap methods use the simulated distribution of parameter estimates in place of their analytical one (Efron, 1987; DiCiccio and Efron, 1996). Most of these methods are discussed in detail by Hall (1992) and DiCiccio and Efron (1996) while those belonging to the test inversion family are reviewed by Carpenter (1999); Carpenter and Bithell (2000). Efron and Tibshirani (1993) and Davison and Hinkley (1997) provide practical examples of CI construction along with some S-plus software code. All these methods are computationally intensive and affected by Monte Carlo error (Carpenter and Bithell, 2000).

Before describing the eight bootstrap methods, resampling algorithms are discussed in a regression context. Different sampling strategies, either parametric or non-parametric, can be used to produce a bootstrap sample and, thus, a simulated $\widehat{WTP}$ distribution.

Parametric resampling. A parametric model for the data is assumed known up to the unknown parameter vector, generally replaced by its MLE. In a regression context 'assuming the model' implies treating model assumptions as true. In other words, the predictors are known without error (i.e. the natural framework of a stated preference study) and the error term follows a specific distribution (e.g. Gumbel).

Let $\hat{\beta}$ be the MLE of β obtained by fitting the logit model (e.g. MNL) to the original data. The algorithm producing the $\widehat{WTP}$ bootstrap distribution, under the parametric resampling scheme, performs the following steps, for $b = 1, \dots, B$:

1. generate a vector, $e^{\star}_{(b)}$, of residuals parametrically (equal in size to the number of observations in the data set), by drawing each component, $e^{\star}_{int(b)}$, independently from the same specified distribution;
2. compute $\hat{U}^{\star}_{int(b)} = X_{int}\hat{\beta} + e^{\star}_{int(b)}$ and, thus, $y^{\star}_{int}$ according to (1), $\forall i, n, t$, and produce a parametric bootstrap sample, $y^{\star}_{(b)}$;
3. regress the bootstrapped values $y^{\star}_{(b)}$ on the fixed predictors to obtain bootstrap replications of the estimated regression coefficients, $\hat{\beta}^{\star}_{(b)}$, and bootstrap replications of the estimated WTP parameter, $\widehat{WTP}^{\star}_{(b)}$.

Non-parametric resampling. In this case, no assumptions are made concerning the data generating process. Let the original sample of observations be $w_{int} = (y_{int}\ , X_{int})$, for $i = 1, \dots, J$, $n = 1, \dots, N$ and $t = 1, \dots, T$. Then, for $b = 1, \dots, B$:

1. resample the observations w_{int} with replacement to generate a new sample; let this sample be $w^{\star}_{(b)}$ and have the same number of observations as the original one;
2. fit the logit model to the bootstrap sample $w^{\star}_{(b)}$ to obtain $\hat{\beta}^{\star}_{(b)}$ and $\widehat{WTP}^{\star}_{(b)}$.

Notice that, under this sampling scheme, predictors too are treated as random. This potentially implies loosing all the desirable experimental design properties a researcher might have developed in a stated preference study. Nevertheless non-parametric random-x resampling plans are appealing mainly for the following reason. Fixed-x resampling enforces the assumption that the errors are identically distributed by resampling residuals from a common distribution. Consequently, if the model is incorrectly specified (e.g.

unmodelled nonlinearity, non-constant error variance, outliers, etc.) these characteristics will not carry over into the resampled data sets.

Krinsky and Robb resampling. This resampling method consists in drawing many values of the parameters of the model from a multivariate normal distribution with mean $\hat{\beta}$ and variance-covariance matrix $\hat{\Sigma}_{\hat{\beta}}$, the variance-covariance matrix of the estimates. This method can be considered as a parametric sampling scheme, since it samples from a specified distribution. The sampling algorithm, for $b = 1, \ldots, B$, proceeds as follows:

1. draw a vector $\hat{\beta}^{\star}_{(b)}$ from a $N(\hat{\beta}, \hat{\Sigma}_{\hat{\beta}})$;
2. use the vector $\hat{\beta}^{\star}_{(b)}$ to calculate $\widehat{WTP}^{\star}_{(b)}$.

This sampling scheme, originally proposed by Krinsky and Robb (1986, 1990), has both been widely applied in transportation research and also misinterpreted (Daly et al., 2012a). In fact, Krinsky and Robb, assuming random parameters derived from linear models, consider them as exactly normally distributed. In logit models, instead, parameter estimates are only asymptotically normal and the assumption of normality might be inappropriate. This is particularly true for small samples. Furthermore, the elasticity functions considered in those papers did not involve a ratio of parameters as in the WTP case. Since for a ratio of random normal variables the variance does not exist, using Krinsky and Robb resampling can be seriously misleading Daly et al. (2012a,b). The method can be purposely used in the case of WTP percentiles when the required result actually exists.

3.2.1. Non-Studentized bootstrap method

A natural way of constructing a CI for WTP is to seek a function of $\widehat{WTP}$ and WTP whose distribution is known and use its quantiles to construct a CI. When drawing observations from an unknown population distribution, it is not clear which function should be chosen. However, since many estimators are asymptotically normally distributed around their mean, it is reasonable to use

$$W = \widehat{WTP} - WTP. \tag{13}$$

If the distribution of W were known, $\left[\widehat{WTP} - w_{1-\alpha/2}\ ;\ \widehat{WTP} - w_{\alpha/2}\right]$ would represent a $(1-\alpha)$-level CI for WTP, where w_α is the quantile of W such that $P(W < w_\alpha) = \alpha$. When the distribution of W is unknown, the non-studentized bootstrap method[3] (NS) suggests to replace the quantile, w_α, with the appropriate quantile, $w^\star_\alpha$, of $W^\star$, calculated through the following algorithm:

1. set $W^\star_{(b)} = \widehat{WTP}^\star_{(b)} - \widehat{WTP}$, for $b = 1, \ldots, B$;
2. estimate the α-th quantile of $W^\star$ as $\hat{w}^\star_\alpha$, the ordered value of $\{W^\star_{(b)}, b = 1, \ldots, B\}$ which occupies the position $\lceil \alpha B \rceil$, where $\lceil x \rceil$ denotes the integer ceiling of the real positive number x, thus $x \leq \lceil x \rceil < x + 1$.
3. calculate the $(1-\alpha)$-level non-Studentized pivotal interval as:

$$\left[\widehat{WTP} - \hat{w}^\star_{1-\alpha/2}\ ;\ \widehat{WTP} - \hat{w}^\star_{\alpha/2}\right]. \tag{14}$$

Unfortunately, the distributions of W and $W^\star$ might differ markedly, leading to substantial coverage errors. Moreover, if there is a parameter constraint (such as $WTP > 0$) then the interval might include invalid parameter values. On the other hand, this procedure provides simple to calculate CIs. Davison and Hinkley (1997) prove that NS is particularly accurate for some parameters such as the median.

3.2.2. Studentized bootstrap method

The Studentized bootstrap method[4] (S), first suggested by Efron (1979), tries to overcome the shortcomings of NS. However, some poor numerical results reduced its appeal. Hall (1988) showed the bootstrap-t's good second-order properties, thus reviving interest in its use. In

[3] Also known as basic bootstrap method.

[4] Also known as bootstrap-t interval or studentized pivotal method.

line with Student's t-statistic, S replaces (13) with

$$W = \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}}, \tag{15}$$

where $\sqrt{\mathrm{var}(\widehat{WTP})}$ is an estimate of $\widehat{WTP}$ standard deviation. The endpoints of a $(1-\alpha)$-level two-sided CI for WTP are:

$$\left[\widehat{WTP} - w_{1-\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\ ;\ \widehat{WTP} - w_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\right]. \tag{16}$$

In the usual Student's-t case, the percentiles w_α are those of the Student distribution, while S estimates the percentiles of W by bootstrapping, through the following algorithm:

1. set $W^\star_{(b)} = \dfrac{\widehat{WTP}^\star_{(b)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^\star_{(b)})}}$, for $b = 1, \ldots, B$, where $\mathrm{var}(\widehat{WTP}^\star_{(b)})$ must be numerically computed for each bootstrap data set, using, for example, Delta estimates;
2. estimate the α-th quantile of $W^\star$ as $\hat{w}^\star_\alpha$, the ordered value of $\{W^\star_{(b)}, b = 1, \ldots, B\}$ which occupies the position $\lceil \alpha B \rceil$.
3. calculate the CI as in (16), replacing w_α with $\hat{w}^\star_\alpha$.

The quantiles used represent the only difference with respect to the CI in (5). An advantage of this approach compared to Delta, when the distribution of $\widehat{WTP}$ is skewed, is that it produces not necessarily symmetric CIs.

3.2.3. Normal-theory bootstrap method

Assuming that $\widehat{WTP}$ is approximately normal, a bootstrap CI can be obtained as in (5), where now $\mathrm{var}(\widehat{WTP})$ is estimated on the bootstrap sample. The Normal-theory bootstrap method (NT) is based on the following algorithm:

1. estimate $\mathrm{var}(\widehat{WTP}^\star) = \dfrac{1}{B-1}\displaystyle\sum_{b=1}^{B}(\widehat{WTP}^\star_{(b)} - \overline{WTP}^\star)^2$, where $\overline{WTP}^\star = \sum_{b=1}^{B} \widehat{WTP}^\star_{(b)}/B$ is the mean of the B bootstrap replicates of $\widehat{WTP}$;
2. calculate the $(1-\alpha)$-level bootstrap CI as:

$$\left[\widehat{WTP} - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP}^\star)}\ ;\ \widehat{WTP} + z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP}^\star)}\right]. \tag{17}$$

In this case, the $\widehat{WTP}$ standard deviation, rather than the quantiles, is replaced by its bootstrap estimate. It is important to note that this method, like Delta, always delivers symmetric CIs.

3.2.4. Bootstrap percentile method

The bootstrap percentile method (P) uses empirical percentiles of $\widehat{WTP}$ bootstrap distribution to obtain a CI through the following algorithm:

1. let $\widehat{WTP}^\star_{[1]}, \dots, \widehat{WTP}^\star_{[B]}$ be the ordered bootstrap replicates of $\widehat{WTP}$;
2. calculate $L = (B+1)\alpha/2$ and $U = (B+1)(1-\alpha/2)$ and build the CI for WTP as:

$$\left[\widehat{WTP}^\star_{\lceil L \rceil} \quad ; \quad \widehat{WTP}^\star_{\lceil U \rceil}\right]. \tag{18}$$

The rationale, which is then pushed forward to get the methods described in Section 3.2.5 and 3.2.6 is the following. Assuming $g(\cdot)$ to be a monotonically increasing function, let $\phi = g(WTP)$, $\hat{\phi} = g(\widehat{WTP})$ and $\hat{\phi}^\star = g(\widehat{WTP}^\star)$. Choose $g(\cdot)$, such that

$$\hat{\phi} - \phi \sim \hat{\phi}^\star - \hat{\phi} \sim N(0, \sigma^2) \tag{19}$$

so to deliver the following $(1-\alpha)$-level CI for WTP:

$$\left[g^{-1}(\hat{\phi} - \sigma z_{\alpha/2}) \quad ; \quad g^{-1}(\hat{\phi} + \sigma z_{\alpha/2})\right]. \tag{20}$$

However, (19) implies that $\hat{\phi} - \sigma z_{\alpha/2} = F^{-1}_{\hat{\phi}^\star}(\alpha/2)$ and $\hat{\phi} + \sigma z_{\alpha/2} = F^{-1}_{\hat{\phi}^\star}(1-\alpha/2)$, with $F^{-1}_{\hat{\phi}^\star}(\cdot)$ being the inverse of the cumulative distribution of $\hat{\phi}^\star$. Since $g(\cdot)$ is monotonically increasing $F^{-1}_{\hat{\phi}^\star}(\alpha/2) = g(F^{-1}_{\widehat{WTP}^\star}(\alpha/2))$ and analogously for $F^{-1}_{\hat{\phi}^\star}(1-\alpha/2)$, where $F^{-1}_{\widehat{WTP}^\star}$ is the bootstrap inverse cumulative distribution of $\widehat{WTP}^\star$. Interval (20) becomes

$$\left[F^{-1}_{\widehat{WTP}^\star}(\alpha/2) \quad ; \quad F^{-1}_{\widehat{WTP}^\star}(1-\alpha/2)\right], \tag{21}$$

which is exactly the interval in (18).

The simplicity of P is particularly appealing. In fact, neither the estimate of $\mathrm{var}(\widehat{WTP})$ nor the specification of $g(\cdot)$ are required. An important advantage over the methods belonging to the pivotal family is that no invalid parameter values can be included within the interval. Unfortunately, the method rests on the existence of a $g(\cdot)$ such that (19) holds, but in many practical situations such a $g(\cdot)$ does not exist. This determines a substantial coverage error, whenever $\widehat{WTP}$ distribution is not nearly symmetric.

3.2.5. Bias-corrected bootstrap percentile method

The bias-corrected bootstrap percentile method (BC) tries to improve over P, by relaxing the assumption of a symmetric $\widehat{WTP}$ distribution. It considers a monotonically increasing function $g(\cdot)$, such that

$$\hat{\phi} - \phi \sim \hat{\phi}^{\star} - \hat{\phi} \sim N(-c\sigma, \sigma^2), \tag{22}$$

for some constant c. In this case, the interval, slightly more complex than (21), is:

$$\left[F^{-1}_{\widehat{WTP}^{\star}}\left(\Phi(2c - z_{\alpha/2})\right) \quad ; \quad F^{-1}_{\widehat{WTP}^{\star}}\left(\Phi(2c + z_{\alpha/2})\right)\right], \tag{23}$$

with the bias-correction parameter estimated as:

$$c = \Phi^{-1}\left(\frac{\#\{\widehat{WTP}^{\star}_{(b)} \leq \widehat{WTP}\}}{B}\right) \tag{24}$$

where $\dfrac{\#\{\widehat{WTP}^{\star}_{(b)} \leq \widehat{WTP}\}}{B}$ is the proportion of bootstrap replicates at or below the original-sample estimate $\widehat{WTP}$. If $\widehat{WTP}$ is unbiased and its bootstrap distribution symmetric, this proportion will be close to 0.5, and c will be close to 0, making the interval (23) equal to that in (21).

The algorithm to compute CIs is sketched below:

1. estimate c as in (24);
2. calculate $L = (B+1)\Phi(2c - z_{\alpha/2})$ and $U = (B+1)\Phi(2c + z_{\alpha/2})$ and build the CI for WTP as in (18).

BC represents an improvement over P in presence of non-symmetric $\widehat{WTP}$ distributions. Similar considerations on the existence of $g(\cdot)$ apply also in this case.

3.2.6. Bias-corrected-accelerated bootstrap percentile method

The bias-corrected-accelerated bootstrap percentile method (BC$_a$) accounts for both lack of symmetry in $\widehat{WTP}$ distribution and changes in shape (i.e. skewness) as WTP varies. Two key parameters characterize BC$_a$, namely the bias-correction c and the acceleration a. The function $g(\cdot)$ is such that

$$\hat{\phi} - \phi \sim N(-c\sigma(\phi), \sigma^2(\phi)) \quad \text{and} \quad \hat{\phi}^\star - \hat{\phi} \sim N(-c\sigma(\hat{\phi}), \sigma^2(\hat{\phi})), \tag{25}$$

where $\sigma(x) = 1 + ax$ and the CI is:

$$\left[F^{-1}_{\widehat{WTP}^\star}\left(\Phi\left(c + \frac{c - z_{\alpha/2}}{1 - a(c - z_{\alpha/2})}\right)\right) \quad ; \quad F^{-1}_{\widehat{WTP}^\star}\left(\Phi\left(c + \frac{c + z_{\alpha/2}}{1 - a(c + z_{\alpha/2})}\right)\right)\right]. \tag{26}$$

A simple jackknife estimate of a is used (DiCiccio and Efron, 1996). It is obtained as:

$$a = \frac{\sum_{h=1}^{NT}(\widehat{WTP}_{(-h)} - \overline{WTP})^3}{6\left[\sum_{h=1}^{NT}(\widehat{WTP}_{(-h)} - \overline{WTP})^2\right]^{\frac{3}{2}}}, \tag{27}$$

where $\widehat{WTP}_{(-h)}$, for $h = 1, \ldots, NT$, represents the estimate of WTP when the h-th observation is deleted from the original sample and $\overline{WTP}$ represents the $\widehat{WTP}_{(-h)}$ average, that is $\overline{WTP} = \sum_{h=1}^{NT} \widehat{WTP}_{(-h)}/NT$.

The following algorithm can be used to compute the CI:

1. estimate c as in (24) and a as in (27);
2. calculate $L = (B+1)\Phi\left(c + \dfrac{c - z_{\alpha/2}}{1 - a(c - z_{\alpha/2})}\right)$ and $U = (B+1)\Phi\left(c + \dfrac{c + z_{\alpha/2}}{1 - a(c + z_{\alpha/2})}\right)$ and build the CI for WTP as in (18).

When $a = 0$ and $c = 0.5$, BC$_a$ reduces to P. In all other cases, BC$_a$ is characterized by a smaller coverage error with respect to P and BC. However, coverage error increases as α tends to 0 and caution should be used when $\alpha < 0.025$ (Davison and Hinkley, 1997, p. 205, p. 231).

3.2.7. Test inversion bootstrap method

The test inversion bootstrap method (TIB), first proposed by Kabaila (1993) in time series, is here applied within a choice modeling context. The duality between CIs and hypothesis testing implies that, if $[WTP_L \,;\, WTP_U]$ are the correct endpoints of the $(1-\alpha)$-level interval and a bootstrap sample is drawn after setting $WTP = WTP_L$, then under some natural monotonicity conditions,

$$P\left(\widehat{WTP}^{\star} \geq \widehat{WTP} \mid WTP = WTP_L\right) = \alpha/2. \tag{28}$$

Similarly, if a sample is taken under $WTP = WTP_U$, then

$$P\left(\widehat{WTP}^{\star} \leq \widehat{WTP} \mid WTP = WTP_U\right) = \alpha/2. \tag{29}$$

Solving (28) and (29) with respect to WTP_L and WTP_U produces a CI estimate. In this case, one has to simulate from the bootstrap distribution at different WTP values which is possible only within a parametric resampling scheme. Suppose that WTP_L is the current lower bound estimate. A bootstrap sample can be obtained according to the previously described parametric resampling scheme. The utility function is computed as $\hat{U}^{\star} = V_{\text{con}} + e^{\star}$, where V_{con}, expressed in WTP space, is given as in (12), with WTP replaced by WTP_L.

A stochastic root finding algorithm is needed to solve (28) and (29). The Robbins-Monro algorithm is the most efficient for our purpose among those proposed in the literature (Garthwaite and Buckland, 1992; Carpenter, 1999). Let $g = 1$ and $WTP_L^{(g)}$ be an initial estimate of WTP_L. According to the Robbins-Monro algorithm:

1. generate a bootstrap sample with WTP set equal to $WTP_L^{(g)}$ and let $\widehat{WTP}^{(g)}$ be the estimate of WTP from this sample;
2. set
$$\begin{cases} WTP_L^{(g+1)} = WTP_L^{(g)} + \ell \dfrac{\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} < \widehat{WTP} \\ WTP_L^{(g+1)} = WTP_L^{(g)} - \ell \dfrac{1-\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} \geq \widehat{WTP} \end{cases},$$
where ℓ is the step length constant.

Each step is expected to reduce the distance from WTP_L. The algorithm is iterated a predetermined number of times equal to G, so that $WTP_L^{(G)}$ is taken as an estimate of WTP_L. An independent search is needed for WTP_U. Assuming $WTP_U^{(g)}$ is, after g steps, the estimate of WTP_U, then $WTP_U^{(g+1)}$ can be calculated as:

$$\begin{cases} WTP_U^{(g+1)} = WTP_U^{(g)} - \ell \dfrac{\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} > \widehat{WTP} \\ WTP_U^{(g+1)} = WTP_U^{(g)} + \ell \dfrac{1-\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} \leq \widehat{WTP} \end{cases}.$$

Garthwaite and Buckland (1992) provide details about WTP_L and WTP_U starting value estimates, stopping rule and the choice of ℓ.

TIB is characterized by the advantages pertaining to the test inversion family methods (i.e. no assumptions on $\widehat{WTP}$ distribution, no discontinuity points, no invalid parameter values included in the intervals) as well as those of the bootstrap methods (i.e. no assumptions on the distribution of the test statistic). The main disadvantage pertains to its computational burden due to the different searches needed for the lower and upper confidence limits, with a bootstrap sample needed at each search step. In addition, assessing CI limits convergence requires careful monitoring.

3.2.8. Studentized test inversion bootstrap method

The studentized test inversion bootstrap method (STIB), never used in the choice modeling context, aims at reducing TIB coverage error by replacing $\widehat{WTP}$ in (28) and (29) with a studentized statistic. If $[WTP_L \ ; \ WTP_U]$ are the correct endpoints of the $(1-\alpha)$-level interval and a bootstrap sample is drawn after setting $WTP = WTP_L$, then

$$P\left(\frac{\widehat{WTP}^{\star} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}} \geq \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}} \mid WTP = WTP_L\right) = \alpha/2$$

where the variances can be estimated using Delta. Similarly, if a resample is taken under $WTP = WTP_U$, then

$$P\left(\frac{\widehat{WTP}^{\star} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}} \leq \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}} \mid WTP = WTP_U\right) = \alpha/2.$$

The same algorithm employed in TIB can be used for constructing CIs, where the estimates of WTP_L and WTP_U are now updated as:

$$\begin{cases} WTP_L^{(g+1)} = WTP_L^{(g)} + \ell\dfrac{\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} < \dfrac{\widehat{WTP} - \widehat{WTP}_L^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \\ WTP_L^{(g+1)} = WTP_L^{(g)} - \ell\dfrac{1-\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} \geq \dfrac{\widehat{WTP} - \widehat{WTP}_L^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \end{cases}$$

and

$$\begin{cases} WTP_U^{(g+1)} = WTP_U^{(g)} - \ell\dfrac{\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} > \dfrac{\widehat{WTP} - \widehat{WTP}_U^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \\ WTP_U^{(g+1)} = WTP_U^{(g)} + \ell\dfrac{1-\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} \leq \dfrac{\widehat{WTP} - \widehat{WTP}_U^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \end{cases}.$$

STIB has the same advantages and disadvantages of TIB but is expected to have a smaller coverage error.

4. Simulation study

This section compares the performance of the methods described in Section 3 through a Monte Carlo study. The comparison is carried out within a MNL framework. This choice is motivated by the fact that, in MNL models, choice probabilities have a closed form, leading to quick parameter estimates. This is fundamental in Monte Carlo simulations, where estimation is performed thousands of times. Considering, for example, a mixed logit framework would have been prohibitive[5]. In addition, only few methods have been extended so far to a mixed logit context. Hensher and Greene (2003) adapt P, based on Krinsky and Robb sampling, to the mixed logit model, while Bliemer and Rose (2013) extend Delta, providing formulas for many commonly used random parameter distributions. Bliemer and Rose (2013) also provide a comparison based on real data between Delta and the method in Hensher and Greene (2003).

[5]In our simulation study more than 750,000,000 different parameter estimations were performed.

In Section 6 we provide some discussion on the extensibility of our findings to the mixed logit environment.

In line with Hole (2007), data sets mimicking actual choices are constructed. N hypothetical subjects in $T = 16$ different choice exercises, choose among $J = 2$ alternatives each characterized by X_1, X_2 (2-level attributes) and X_C (4-level cost attribute). Dropping subscripts for simplicity, the deterministic difference in utility is:

$$V_1 - V_2 = \beta_0 + \beta_1(X_{11} - X_{12}) + \beta_2(X_{21} - X_{22}) + \beta_C(X_{C1} - X_{C2}),$$

where the values of the parameters are opportunely set.

A single data set can be simulated by drawing from an appropriate distribution, independently for each N and T, a value for the error difference $\epsilon_1 - \epsilon_2$. If this value is less than $V_1 - V_2$, the first alternative is chosen and the choice variable y is set equal to 1 for the first alternative and to 0 for the second. Otherwise, the second alternative is chosen.

Several scenarios, under various sample size conditions, are simulated to assess the performance of the different methods: 1) the effect of β_C approaching 0, determining WTP values close to its discontinuity point; 2) the correlation between numerator and denominator estimates having the same sign of $\widehat{WTP}$; 3) the impact of model mis-specification, due to heteroscedasticity arising from a dishomogeneous population.

A number of $M = 1000$ different data sets is generated, drawing the error differences from logistic distributions. A MNL model is fitted to each data set, and its parameters estimated via MLE. Then, $\widehat{WTP}_1$, $\widehat{WTP}_2$ and their relative CIs are calculated. The M sample values of the CIs are used to calculate: coverage rates, median interval length and median interval shape attained by the various methods. Let $WTP_L^{(m)}$ and $WTP_U^{(m)}$ represent, respectively, the lower and the upper limits of the CI, calculated with a certain method, for the m-th Monte Carlo data set, and define:

$$c^{(m)} = I\left(WTP_L^{(m)} \leq WTP \leq WTP_U^{(m)}\right)$$
$$\ell^{(m)} = WTP_U^{(m)} - WTP_L^{(m)}$$

$$s^{(m)} = \frac{WTP_U^{(m)} - WTP}{WTP - WTP_L^{(m)}},$$

where $I(\cdot)$ is the indicator function. Coverage, median length, and median shape[6] are calculated as follows:

$$\text{Coverage} = \frac{1}{M}\sum_{m=1}^{M} c^{(m)}, \qquad \text{Length} = \ell^{(\lceil 0.5M \rceil)}, \qquad \text{Shape} = s^{(\lceil 0.5M \rceil)},$$

after sorting, in non-decreasing order, the series $\ell^{(m)}$ and $s^{(m)}$ for $m = 1, \ldots, M$.

Left rejection probability (LRP) and right rejection probability (RRP) are also considered in analyzing the effective coverage. The two indexes are calculated as follows:

$$\text{LRP} = \frac{1}{M}\sum_{m=1}^{M} I\left(WTP \leq \widehat{WTP}_L^{(m)}\right) \quad \text{and} \quad \text{RRP} = \frac{1}{M}\sum_{m=1}^{M} I\left(WTP \geq \widehat{WTP}_U^{(m)}\right).$$

Monte Carlo estimates of confidence limits are derived calculating the $100\alpha/2th$ and $100(1-\alpha/2)th$ percentiles of the M $\widehat{WTP}^{(m)}$ estimates. Monte Carlo CI serves as benchmark for evaluating the accuracy of all the methods considered.

4.1. Cost parameter approaching zero

This section describes the effects a cost parameter approaching 0 has on CI estimates. More in detail, the specific β_C considered are: -1, -0.5 and -0.25. The remaining parameters are set as follows: $\beta_0 = 0.5$, $\beta_1 = 1$ and $\beta_2 = 0.5$. Performance indicators of the different 95% level CIs for WTP_1 and WTP_2 are reported in Tables 1, 2 and 3, for various sample sizes (i.e. $N = 10,\ 25, 50$).

When β_C is far from 0 and its coefficient of variation is small[7], most of the methods considered perform well even for small sample sizes (see Tables 1); few methods have inadequate

[6] Notice that median length and median shape are used since the median is more robust to extreme values than the mean and also because the median length can be calculated in the presence of Fieller CIs with infinite limits. For such intervals, the shape index $s^{(m)}$ cannot, instead, be computed and they are excluded from determination of the median shape.

[7] In Table 1, the t-statistic is equal to -4.88, -7,71 and -10,90, respectively for $N = 10$, $N = 25$ and $N = 50$.

	N	Method	$\widehat{WTP}_1$ Length	Shape	LRP	RRP	Coverage	$\widehat{WTP}_2$ Length	Shape	LRP	RRP	Coverage
	10	Monte Carlo	0.9576	1.1303	0.0250	0.0250	0.9500	0.8439	1.0449	0.0250	0.0250	0.9500
	10	Delta	0.8918	1.0000	0.0220	0.0310	0.9470	0.8428	1.0000	0.0200	0.0220	0.9580
	10	LRTI	0.9082	1.1184	0.0310	0.0270	0.9420	0.8594	1.1020	0.0240	0.0250	0.9510
	10	Fieller	0.9439	1.1639	0.0263	0.0172	0.9570	0.8940	1.1235	0.0232	0.0171	0.9600
Parametric sampling	10	TIB	0.9587	1.0420	0.0160	0.0090 **	0.9750 ***	0.9389	1.1487	0.0180	0.0250	0.9570
	10	STIB	0.9258	1.0979	0.0400 **	0.0430 ***	0.9170 ***	0.8892	1.0969	0.0250	0.0360 *	0.9390
	10	NS	1.0258	0.9415	0.0160	0.0240	0.9600	0.8944	0.9736	0.0100 **	0.0390 **	0.9510
	10	S	0.9022	1.1127	0.0460 ***	0.0450 ***	0.9090 ***	0.8538	1.1026	0.0250	0.0280	0.9470
	10	NT	1.0192	1.0000	0.0170	0.0230	0.9600	0.8939	1.0000	0.0120 **	0.0360 *	0.9520
	10	P	1.0258	1.0621	0.0250	0.0350 *	0.9400	0.8944	1.0272	0.0170	0.0330	0.9500
	10	BC	1.0327	1.0764	0.0200	0.0210	0.9590	0.9023	1.1190	0.0180	0.0280	0.9540
	10	BC$_a$	1.0287	1.0411	0.0180	0.0300	0.9520	0.9007	1.0961	0.0180	0.0280	0.9540
Non param. sampling	10	NS	1.0619	0.9584	0.0130 *	0.0230	0.9640 *	0.9092	0.9829	0.0110 **	0.0410 **	0.9480
	10	S	0.8912	1.1127	0.0460 ***	0.0500 ***	0.9040 ***	0.8490	1.1026	0.0240	0.0410 **	0.9350 *
	10	NT	1.0515	1.0000	0.0150 *	0.0220	0.9630	0.9120	1.0000	0.0100 **	0.0360 *	0.9540
	10	P	1.0619	1.0434	0.0260	0.0320	0.9420	0.9092	1.0174	0.0120 **	0.0360 *	0.9520
	10	BC	1.0717	1.0561	0.0210	0.0230	0.9560	0.9222	1.1233	0.0130 *	0.0280	0.9590
	10	BC$_a$	1.0685	1.0190	0.0200	0.0280	0.9520	0.9182	1.0976	0.0120 **	0.0280	0.9600
Krinsky and Robb sampling	10	NS	0.9432	0.8566	0.0170	0.0400 **	0.9430	0.8951	0.8817	0.0090 **	0.0300	0.9610
	10	S	0.8475	1.1505	0.0500 ***	0.0320	0.9180 ***	0.8049	1.1168	0.0360 *	0.0350 *	0.9290 *
	10	NT	0.9387	1.0000	0.0180	0.0280	0.9540	0.8868	1.0000	0.0110 **	0.0270	0.9620
	10	P	0.9432	1.1653	0.0300	0.0220	0.9480	0.8951	1.1318	0.0240	0.0290	0.9470
	10	BC	0.9453	1.1574	0.0300	0.0200	0.9500	0.8966	1.1262	0.0320	0.0200	0.9480
	10	BC$_a$	0.9442	1.1049	0.0340	0.0280	0.9380	0.8957	1.0936	0.0310	0.0200	0.9490
	25	Monte Carlo	0.5750	0.9714	0.0250	0.0250	0.9500	0.5239	1.0282	0.0250	0.0250	0.9500
	25	Delta	0.5667	1.0000	0.0200	0.0320	0.9480	0.5389	1.0000	0.0190	0.0280	0.9530
	25	LRTI	0.5615	1.0702	0.0240	0.0320	0.9440	0.5420	1.0591	0.0220	0.0210	0.9570
	25	Fieller	0.5800	1.0925	0.0210	0.0280	0.9510	0.5510	1.0681	0.0200	0.0230	0.9570
Parametric sampling	25	TIB	0.5793	1.0176	0.0180	0.0350 *	0.9470	0.5511	1.0398	0.0230	0.0190	0.9580
	25	STIB	0.5877	1.0731	0.0270	0.0260	0.9470	0.5597	1.0923	0.0250	0.0190	0.9560
	25	NS	0.5743	0.9594	0.0150 *	0.0380 **	0.9470	0.5417	0.9615	0.0140 *	0.0260	0.9600
	25	S	0.5802	1.0733	0.0230	0.0280	0.9490	0.5461	1.0756	0.0190	0.0230	0.9580
	25	NT	0.5743	1.0000	0.0190	0.0350 *	0.9460	0.5409	1.0000	0.0160	0.0270	0.9570
	25	P	0.5743	1.0423	0.0200	0.0370 *	0.9430	0.5417	1.0400	0.0180	0.0250	0.9570
	25	BC	0.5755	1.0501	0.0170	0.0310	0.9520	0.5428	1.0756	0.0220	0.0240	0.9540
	25	BC$_a$	0.5748	1.0297	0.0170	0.0310	0.9520	0.5427	1.0649	0.0220	0.0240	0.9540
Non param. sampling	25	NS	0.5796	0.9628	0.0160	0.0340	0.9500	0.5422	0.9686	0.0140 *	0.0290	0.9570
	25	S	0.5820	1.0793	0.0230	0.0270	0.9500	0.5430	1.0826	0.0210	0.0210	0.9580
	25	NT	0.5800	1.0000	0.0180	0.0350 *	0.9470	0.5437	1.0000	0.0150 *	0.0270	0.9580
	25	P	0.5796	1.0386	0.0210	0.0340	0.9450	0.5422	1.0324	0.0200	0.0260	0.9540
	25	BC	0.5819	1.0540	0.0200	0.0280	0.9520	0.5438	1.0694	0.0210	0.0230	0.9560
	25	BC$_a$	0.5811	1.0308	0.0180	0.0300	0.9520	0.5433	1.0599	0.0200	0.0230	0.9570
Krinsky and Robb sampling	25	NS	0.5800	0.9133	0.0130 *	0.0380 **	0.9490	0.5494	0.9318	0.0120 **	0.0250	0.9630
	25	S	0.5570	1.0858	0.0270	0.0340	0.9390	0.5283	1.0649	0.0240	0.0260	0.9500
	25	NT	0.5767	1.0000	0.0180	0.0360 *	0.9460	0.5472	1.0000	0.0170	0.0230	0.9600
	25	P	0.5800	1.0949	0.0230	0.0280	0.9490	0.5494	1.0732	0.0200	0.0210	0.9590
	25	BC	0.5810	1.0926	0.0180	0.0260	0.9560	0.5506	1.0665	0.0200	0.0220	0.9580
	25	BC$_a$	0.5799	1.0675	0.0180	0.0280	0.9540	0.5504	1.0565	0.0180	0.0230	0.9590
	50	Monte Carlo	0.3861	1.0631	0.0250	0.0250	0.9500	0.3875	1.0254	0.0250	0.0250	0.9500
	50	Delta	0.4047	1.0000	0.0180	0.0190	0.9630	0.3847	1.0000	0.0250	0.0270	0.9480
	50	LRTI	0.4089	1.0494	0.0210	0.0180	0.9610	0.3784	1.0400	0.0260	0.0260	0.9480
	50	Fieller	0.4092	1.0626	0.0190	0.0180	0.9630	0.3889	1.0461	0.0250	0.0240	0.9510
Parametric sampling	50	TIB	0.4128	1.0186	0.0170	0.0230	0.9600	0.3890	1.0097	0.0200	0.0260	0.9540
	50	STIB	0.4144	1.0513	0.0220	0.0150 *	0.9630	0.3956	1.0484	0.0280	0.0250	0.9470
	50	NS	0.4064	0.9708	0.0170	0.0200	0.9630	0.3849	0.9755	0.0210	0.0320	0.9470
	50	S	0.4093	1.0491	0.0210	0.0150 *	0.9640 *	0.3876	1.0493	0.0250	0.0250	0.9500
	50	NT	0.4067	1.0000	0.0190	0.0180	0.9630	0.3856	1.0000	0.0240	0.0290	0.9470
	50	P	0.4064	1.0300	0.0220	0.0170	0.9610	0.3849	1.0251	0.0250	0.0270	0.9480
	50	BC	0.4073	1.0336	0.0210	0.0160	0.9630	0.3859	1.0461	0.0270	0.0250	0.9480
	50	BC$_a$	0.4071	1.0196	0.0210	0.0170	0.9620	0.3860	1.0400	0.0270	0.0250	0.9480
Non param. sampling	50	NS	0.4085	0.9698	0.0140 *	0.0190	0.9670 *	0.3874	0.9765	0.0230	0.0280	0.9490
	50	S	0.4102	1.0492	0.0200	0.0150 *	0.9650 *	0.3880	1.0512	0.0260	0.0250	0.9490
	50	NT	0.4088	1.0000	0.0180	0.0170	0.9650 *	0.3872	1.0000	0.0250	0.0270	0.9480
	50	P	0.4085	1.0312	0.0220	0.0160	0.9620	0.3874	1.0241	0.0250	0.0260	0.9490
	50	BC	0.4097	1.0330	0.0210	0.0160	0.9630	0.3884	1.0437	0.0270	0.0220	0.9510
	50	BC$_a$	0.4099	1.0181	0.0190	0.0180	0.9630	0.3884	1.0369	0.0270	0.0220	0.9510
Krinsky and Robb sampling	50	NS	0.4086	0.9404	0.0150 *	0.0220	0.9630	0.3887	0.9539	0.0200	0.0290	0.9510
	50	S	0.4020	1.0589	0.0220	0.0180	0.9600	0.3820	1.0428	0.0260	0.0260	0.9480
	50	NT	0.4079	1.0000	0.0170	0.0180	0.9650 *	0.3878	1.0000	0.0240	0.0260	0.9500
	50	P	0.4086	1.0634	0.0220	0.0160	0.9620	0.3887	1.0483	0.0260	0.0250	0.9490
	50	BC	0.4099	1.0628	0.0230	0.0150 *	0.9620	0.3891	1.0426	0.0250	0.0240	0.9510
	50	BC$_a$	0.4095	1.0478	0.0230	0.0150 *	0.9620	0.3890	1.0357	0.0240	0.0240	0.9520

Table 1: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. Model simulated: MNL model with orthogonal experimental design. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$.

	N	Method	$\widehat{WTP}_1$ Length	Shape	LRP	RRP	Coverage	$\widehat{WTP}_2$ Length	Shape	LRP	RRP	Coverage
	10	Monte Carlo	2.0553	1.3507	0.0250	0.0250	0.9500	1.7400	1.2915	0.0250	0.0250	0.9500
	10	Delta	1.9566	1.0000	0.0000 ***	0.0520 ***	0.9480	1.6368	1.0000	0.0020 ***	0.0390 **	0.9590
	10	LRTI	2.2900	1.8844	0.0180	0.0250	0.9570	1.8359	1.4929	0.0170	0.0330	0.9500
	10	Fieller	2.3146	1.8929	0.0140 *	0.0190	0.9670 *	1.8749	1.5069	0.0180	0.0250	0.9570
Parametric sampling	10	TIB	2.0473	1.1043	0.0000 ***	0.0530 ***	0.9470	1.6717	1.0163	0.0020 ***	0.0400 **	0.9580
	10	STIB	2.2612	1.9265	0.0440 ***	0.0220	0.9340 *	1.8296	1.4412	0.0310	0.0260	0.9430
	10	NS	2.2821	0.5431	0.0000 ***	0.0700 ***	0.9300 **	1.8100	0.6775	0.0000 ***	0.0390 **	0.9610
	10	S	1.8834	1.8686	0.0810 ***	0.0350 *	0.8840 ***	1.5708	1.4647	0.0550 ***	0.0360 *	0.9090 ***
	10	NT	2.3424	1.0000	0.0000 ***	0.0430 ***	0.9570	1.8260	1.0000	0.0010 ***	0.0330	0.9660 *
	10	P	2.2821	1.8411	0.0170	0.0220	0.9600	1.8100	1.4760	0.0160	0.0290	0.9550
	10	BC	2.2865	1.8797	0.0170	0.0250	0.9580	1.8249	1.4902	0.0170	0.0300	0.9530
	10	BC_a	2.2981	1.8850	0.0200	0.0270	0.9520	1.8238	1.4767	0.0190	0.0310	0.9500
Non param. sampling	10	NS	2.3200	0.5360	0.0000 ***	0.0680 ***	0.9320 **	1.8491	0.6805	0.0000 ***	0.0370 *	0.9630
	10	S	1.8854	1.8770	0.0830 ***	0.0340	0.8830 ***	1.5721	1.4798	0.0560 ***	0.0350 *	0.9090 ***
	10	NT	2.3859	1.0000	0.0000 ***	0.0420 ***	0.9580	1.8835	1.0000	0.0000 ***	0.0330	0.9670 *
	10	P	2.3200	1.8658	0.0130 *	0.0230	0.9640 *	1.8491	1.4695	0.0140 *	0.0280	0.9580
	10	BC	2.3433	1.9106	0.0140 *	0.0230	0.9630	1.8583	1.4831	0.0140 *	0.0270	0.9590
	10	BC_a	2.3411	1.9287	0.0170	0.0260	0.9570	1.8565	1.4670	0.0150 *	0.0270	0.9580
Krinsky and Robb sampling	10	NS	2.3284	0.5232	0.0000 ***	0.0690 ***	0.9310 **	1.8847	0.6614	0.0000 ***	0.0350 *	0.9650 *
	10	S	1.8475	1.8395	0.0780 ***	0.0430 ***	0.8790 ***	1.5049	1.4592	0.0620 ***	0.0400 **	0.8980 ***
	10	NT	2.4304	1.0000	0.0000 ***	0.0430 ***	0.9570	1.9329	1.0000	0.0010 ***	0.0310	0.9680 **
	10	P	2.3284	1.9112	0.0130 *	0.0190	0.9680 **	1.8847	1.5119	0.0140 *	0.0250	0.9610
	10	BC	2.3326	1.9113	0.0120 **	0.0200	0.9680 **	1.8898	1.5148	0.0150 *	0.0240	0.9610
	10	BC_a	2.3326	1.9181	0.0170	0.0200	0.9630	1.8760	1.4946	0.0150 *	0.0240	0.9610
	25	Monte Carlo	1.3094	1.2689	0.0250	0.0250	0.9500	1.1109	1.2022	0.0250	0.0250	0.9500
	25	Delta	1.2633	1.0000	0.0020 ***	0.0410 **	0.9570	1.0421	1.0000	0.0100 **	0.0350 *	0.9550
	25	LRTI	1.3379	1.4692	0.0230	0.0300	0.9470	1.0986	1.2644	0.0340	0.0250	0.9410
	25	Fieller	1.3453	1.4753	0.0180	0.0230	0.9590	1.0971	1.2822	0.0250	0.0230	0.9520
Parametric sampling	25	TIB	1.2918	1.0389	0.0030 ***	0.0390 **	0.9580	1.0626	1.0043	0.0140 *	0.0350 *	0.9510
	25	STIB	1.3392	1.4854	0.0310	0.0270	0.9420	1.1015	1.2831	0.0290	0.0250	0.9460
	25	NS	1.3384	0.6823	0.0000 ***	0.0520 ***	0.9480	1.0867	0.7875	0.0000 ***	0.0390 **	0.9610
	25	S	1.2407	1.4698	0.0420 ***	0.0290	0.9290 **	1.0256	1.2683	0.0390 **	0.0290	0.9320 **
	25	NT	1.3368	1.0000	0.0000 ***	0.0390 **	0.9610	1.0834	1.0000	0.0040 ***	0.0340	0.9620
	25	P	1.3384	1.4656	0.0200	0.0250	0.9540	1.0867	1.2699	0.0270	0.0270	0.9460
	25	BC	1.3405	1.4770	0.0210	0.0240	0.9550	1.0898	1.2712	0.0260	0.0270	0.9470
	25	BC_a	1.3400	1.4754	0.0210	0.0250	0.9540	1.0889	1.2610	0.0260	0.0270	0.9470
Non param. sampling	25	NS	1.3400	0.6852	0.0000 ***	0.0500 ***	0.9500	1.0889	0.7944	0.0000 ***	0.0400 **	0.9600
	25	S	1.2355	1.4765	0.0400 **	0.0260	0.9340 *	1.0261	1.2761	0.0360 *	0.0300	0.9340 *
	25	NT	1.3402	1.0000	0.0000 ***	0.0390 **	0.9610	1.0810	1.0000	0.0020 ***	0.0320	0.9660 *
	25	P	1.3400	1.4595	0.0200	0.0240	0.9560	1.0889	1.2588	0.0250	0.0250	0.9500
	25	BC	1.3424	1.4864	0.0190	0.0240	0.9570	1.0940	1.2737	0.0260	0.0240	0.9500
	25	BC_a	1.3402	1.4858	0.0200	0.0260	0.9540	1.0920	1.2640	0.0260	0.0250	0.9490
Krinsky and Robb sampling	25	NS	1.3444	0.6718	0.0000 ***	0.0570 ***	0.9430	1.0990	0.7734	0.0000 ***	0.0390 **	0.9610
	25	S	1.2297	1.4663	0.0440 ***	0.0290	0.9270 ***	1.0083	1.2728	0.0400 **	0.0320	0.9280 **
	25	NT	1.3421	1.0000	0.0000 ***	0.0360 *	0.9640 *	1.0881	1.0000	0.0040 ***	0.0290	0.9670 *
	25	P	1.3444	1.4886	0.0180	0.0250	0.9570	1.0990	1.2930	0.0280	0.0230	0.9490
	25	BC	1.3471	1.4812	0.0180	0.0250	0.9570	1.1031	1.2874	0.0250	0.0200	0.9550
	25	BC_a	1.3492	1.4803	0.0180	0.0250	0.9570	1.1028	1.2780	0.0250	0.0210	0.9540
	50	Monte Carlo	0.9554	1.2955	0.0250	0.0250	0.9500	0.7773	1.2883	0.0250	0.0250	0.9500
	50	Delta	0.8962	1.0000	0.0110 **	0.0360 *	0.9530	0.7439	1.0000	0.0180	0.0250	0.9570
	50	LRTI	0.9241	1.3087	0.0300	0.0200	0.9500	0.7666	1.1918	0.0310	0.0180	0.9510
	50	Fieller	0.9236	1.3117	0.0290	0.0220	0.9490	0.7633	1.1888	0.0310	0.0180	0.9510
Parametric sampling	50	TIB	0.9132	1.0175	0.0140 *	0.0360 *	0.9500	0.7569	0.9987	0.0190	0.0270	0.9540
	50	STIB	0.9308	1.3247	0.0290	0.0220	0.9490	0.7719	1.1971	0.0320	0.0210	0.9470
	50	NS	0.9240	0.7642	0.0000 ***	0.0460 ***	0.9540	0.7556	0.8443	0.0060 ***	0.0290	0.9650 *
	50	S	0.8899	1.3084	0.0340	0.0240	0.9420	0.7376	1.1818	0.0360 *	0.0200	0.9440
	50	NT	0.9217	1.0000	0.0050 ***	0.0340	0.9610	0.7564	1.0000	0.0170	0.0230	0.9600
	50	P	0.9240	1.3086	0.0300	0.0240	0.9460	0.7556	1.1844	0.0330	0.0170	0.9500
	50	BC	0.9255	1.3098	0.0300	0.0220	0.9480	0.7585	1.1822	0.0330	0.0180	0.9490
	50	BC_a	0.9256	1.3095	0.0300	0.0230	0.9470	0.7575	1.1765	0.0320	0.0180	0.9500
Non param. sampling	50	NS	0.9297	0.7616	0.0000 ***	0.0430 ***	0.9570	0.7548	0.8461	0.0100 **	0.0300	0.9600
	50	S	0.8930	1.3122	0.0370 *	0.0240	0.9390	0.7335	1.1808	0.0380 **	0.0170	0.9450
	50	NT	0.9255	1.0000	0.0080 ***	0.0330	0.9590	0.7514	1.0000	0.0150 *	0.0230	0.9620
	50	P	0.9297	1.3130	0.0330	0.0190	0.9480	0.7548	1.1819	0.0300	0.0180	0.9520
	50	BC	0.9331	1.3129	0.0320	0.0230	0.9450	0.7560	1.1818	0.0310	0.0190	0.9500
	50	BC_a	0.9331	1.3119	0.0320	0.0230	0.9450	0.7549	1.1746	0.0310	0.0200	0.9490
Krinsky and Robb sampling	50	NS	0.9247	0.7597	0.0000 ***	0.0480 ***	0.9520	0.7618	0.8329	0.0080 ***	0.0310	0.9610
	50	S	0.8898	1.3039	0.0370 *	0.0250	0.9380	0.7336	1.1898	0.0370 *	0.0220	0.9410
	50	NT	0.9234	1.0000	0.0060 ***	0.0340	0.9600	0.7595	1.0000	0.0160	0.0230	0.9610
	50	P	0.9247	1.3163	0.0320	0.0230	0.9450	0.7618	1.2006	0.0320	0.0190	0.9490
	50	BC	0.9223	1.3062	0.0320	0.0240	0.9440	0.7656	1.1896	0.0320	0.0180	0.9500
	50	BC_a	0.9234	1.3058	0.0320	0.0250	0.9430	0.7649	1.1833	0.0320	0.0190	0.9490

Table 2: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. Model simulated: MNL model with orthogonal experimental design. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -0.5$.

	N	Method	$\widehat{WTP}_1$ Length	Shape	LRP	RRP	Coverage	$\widehat{WTP}_2$ Length	Shape	LRP	RRP	Coverage
	10	Monte Carlo	15.2290	3.4325	0.0250	0.0250	0.9500	8.4442	3.0840	0.0250	0.0250	0.9500
	10	Delta	6.8556	1.0000	0.0000 ***	0.1030 ***	0.8970 ***	4.2240	1.0000	0.0000 ***	0.0770 ***	0.9230 ***
	10	LRTI	20.6641	8.7843	0.0100 **	0.0300	0.9600	10.6641	5.4118	0.0190	0.0270	0.9540
	10	Fieller	21.3345	4.5744	0.0000 ***	0.0347	0.9580	11.6398	3.2236	0.0000 ***	0.0229	0.9620
Parametric sampling	10	TIB	29.6260	5.5833	0.0030 ***	0.0820 ***	0.9150 ***	26.3406	1.0481	0.0010 ***	0.0760 ***	0.9230 ***
	10	STIB	8.4122	3.0851	0.0100 **	0.0280	0.9620	5.4977	1.8749	0.0210	0.0280	0.9510
	10	NS	15.1248	0.3136	0.0060 ***	0.1710 ***	0.8230 ***	8.2964	0.4599	0.0050 ***	0.0950 ***	0.9000 ***
	10	S	7.1503	4.8277	0.2390 ***	0.0580 ***	0.7030 ***	3.8728	3.0769	0.1930 ***	0.0680 ***	0.7390 ***
	10	NT	189.1937	1.0000	0.0000 ***	0.0420 ***	0.9580	93.3485	1.0000	0.0000 ***	0.0280	0.9720 **
	10	P	15.1248	3.1887	0.0000 ***	0.0250	0.9750 ***	8.2964	2.1745	0.0000 ***	0.0220	0.9770 ***
	10	BC	16.7205	5.4917	0.0030 ***	0.0280	0.9690 **	8.9018	4.0587	0.0080 ***	0.0250	0.9670 *
	10	BC$_a$	17.9803	6.3633	0.0040 ***	0.0290	0.9670 *	9.0352	4.2639	0.0090 **	0.0250	0.9660 *
Non param. sampling	10	NS	15.4022	0.3310	0.0050 ***	0.1720 ***	0.8230 ***	8.7385	0.4649	0.0050 ***	0.0980 ***	0.8970 ***
	10	S	7.1528	4.8388	0.2380 ***	0.0570 ***	0.7050 ***	3.9462	3.1201	0.1920 ***	0.0680 ***	0.7400 ***
	10	NT	166.6002	1.0000	0.0000 ***	0.0350 *	0.9650 *	76.6297	1.0000	0.0000 ***	0.0240	0.9760 ***
	10	P	15.4022	3.0209	0.0000 ***	0.0220	0.9780 ***	8.7385	2.1509	0.0000 ***	0.0200	0.9800 ***
	10	BC	17.5345	5.9005	0.0020 ***	0.0280	0.9700 **	9.3534	4.2173	0.0070 ***	0.0230	0.9700 **
	10	BC$_a$	18.4542	6.9539	0.0020 ***	0.0280	0.9700 **	9.4445	4.4733	0.0070 ***	0.0230	0.9700 **
Krinsky and Robb sampling	10	NS	15.2320	0.3211	0.0050 ***	0.1730 ***	0.8220 ***	8.4909	0.4692	0.0060 ***	0.0890 ***	0.9050 ***
	10	S	6.9235	4.4932	0.2300 ***	0.0620 ***	0.7080 ***	3.8029	2.9941	0.1910 ***	0.0740 ***	0.7350 ***
	10	NT	150.9794	1.0000	0.0000 ***	0.0370 *	0.9630	73.3779	1.0000	0.0000 ***	0.0200	0.9800 ***
	10	P	15.2320	3.1142	0.0000 ***	0.0220	0.9780 ***	8.4909	2.1313	0.0000 ***	0.0150 *	0.9850 ***
	10	BC	16.2477	5.4088	0.0010 ***	0.0240	0.9750 ***	9.1288	3.9373	0.0080 ***	0.0190	0.9730 ***
	10	BC$_a$	17.3287	6.3764	0.0010 ***	0.0230	0.9760 ***	9.3681	4.2903	0.0100 **	0.0190	0.9710 **
	25	Monte Carlo	5.9592	2.2693	0.0250	0.0250	0.9500	3.3758	1.8337	0.0250	0.0250	0.9500
	25	Delta	4.2692	1.0000	0.0000 ***	0.0700 ***	0.9300 **	2.6410	1.0000	0.0000 ***	0.0610 ***	0.9390
	25	LRTI	5.6396	2.9577	0.0290	0.0250	0.9460	3.3398	2.4000	0.0270	0.0280	0.9450
	25	Fieller	5.7322	2.8140	0.0062 ***	0.0280	0.9490	3.4067	2.3026	0.0145 *	0.0239	0.9530
Parametric sampling	25	TIB	4.5211	1.2532	0.0000 ***	0.0720 ***	0.9280 **	2.6836	1.0399	0.0000 ***	0.0630 ***	0.9370
	25	STIB	4.8668	2.9605	0.0990 ***	0.0310	0.8700 ***	2.9293	2.0964	0.0770 ***	0.0300	0.8930 ***
	25	NS	5.7330	0.3539	0.0000 ***	0.1350 ***	0.8650 ***	3.3658	0.4456	0.0000 ***	0.0940 ***	0.9060 ***
	25	S	4.2578	2.8970	0.1540 ***	0.0370 *	0.8090 ***	2.5172	2.2041	0.1360 ***	0.0410 **	0.8230 ***
	25	NT	7.2512	1.0000	0.0000 ***	0.0490 ***	0.9510	4.0064	1.0000	0.0000 ***	0.0470 ***	0.9530
	25	P	5.7330	2.8257	0.0030 ***	0.0270	0.9700 **	3.3658	2.2441	0.0100 **	0.0240	0.9660 *
	25	BC	5.7668	2.9512	0.0060 ***	0.0280	0.9660 *	3.3647	2.3327	0.0160	0.0240	0.9600
	25	BC$_a$	5.8327	3.0240	0.0060 ***	0.0280	0.9660 *	3.3728	2.3472	0.0160	0.0240	0.9600
Non param. sampling	25	NS	5.8012	0.3520	0.0000 ***	0.1350 ***	0.8650 ***	3.4098	0.4377	0.0000 ***	0.0900 ***	0.9100 ***
	25	S	4.2666	2.9024	0.1520 ***	0.0380 **	0.8100 ***	2.5059	2.2015	0.1310 ***	0.0400 **	0.8290 ***
	25	NT	7.9278	1.0000	0.0000 ***	0.0460 ***	0.9540	4.2000	1.0000	0.0000 ***	0.0450 ***	0.9550
	25	P	5.8012	2.8408	0.0030 ***	0.0240	0.9730 ***	3.4098	2.2849	0.0110 **	0.0220	0.9670 *
	25	BC	5.8818	3.0035	0.0060 ***	0.0230	0.9710 **	3.4286	2.3526	0.0160	0.0230	0.9610
	25	BC$_a$	5.9124	3.0882	0.0070 ***	0.0220	0.9710 **	3.4285	2.3745	0.0160	0.0230	0.9610
Krinsky and Robb sampling	25	NS	5.6870	0.3641	0.0000 ***	0.1320 ***	0.8680 ***	3.4123	0.4403	0.0000 ***	0.0920 ***	0.9080 ***
	25	S	4.2240	2.8264	0.1480 ***	0.0410 **	0.8110 ***	2.4685	2.1847	0.1340 ***	0.0420 ***	0.8240 ***
	25	NT	8.3479	1.0000	0.0000 ***	0.0510 ***	0.9490	4.4484	1.0000	0.0000 ***	0.0400 **	0.9600
	25	P	5.6870	2.7462	0.0010 ***	0.0290	0.9700 **	3.4123	2.2711	0.0100 **	0.0260	0.9640 *
	25	BC	5.7434	2.8614	0.0040 ***	0.0260	0.9700 **	3.4339	2.3344	0.0120 **	0.0260	0.9620
	25	BC$_a$	5.8148	2.9367	0.0040 ***	0.0260	0.9700 **	3.4338	2.3551	0.0120 **	0.0260	0.9620
	50	Monte Carlo	3.6449	1.7233	0.0250	0.0250	0.9500	5.5166	-12.0957	0.0250	0.0250	0.9500
	50	Delta	3.0629	1.0000	0.0000 ***	0.0630 ***	0.9370	1.9070	1.0000	0.0000 ***	0.0560 ***	0.9440
	50	LRTI	3.4180	2.0522	0.0280	0.0350 *	0.9370	2.1875	1.7638	0.0200	0.0250	0.9550
	50	Fieller	3.5108	2.0342	0.0280	0.0280	0.9440	2.1449	1.7774	0.0180	0.0230	0.9590
Parametric sampling	50	TIB	3.1429	1.1046	0.0000 ***	0.0630 ***	0.9370	1.9360	1.0097	0.0000 ***	0.0560 ***	0.9440
	50	STIB	3.2540	2.1617	0.0770 ***	0.0290	0.8940 ***	2.0092	1.7775	0.0510 ***	0.0310	0.9180 ***
	50	NS	3.5301	0.4821	0.0000 ***	0.1020 ***	0.8980 ***	2.1503	0.5612	0.0000 ***	0.0810 ***	0.9190 ***
	50	S	3.0508	2.0766	0.0910 ***	0.0340	0.8750 ***	1.8479	1.7691	0.0680 ***	0.0360 *	0.8960 ***
	50	NT	3.5871	1.0000	0.0000 ***	0.0530 ***	0.9470	2.1816	1.0000	0.0000 ***	0.0490 ***	0.9510
	50	P	3.5301	2.0745	0.0260	0.0280	0.9460	2.1503	1.7818	0.0170	0.0260	0.9570
	50	BC	3.5439	2.1001	0.0280	0.0270	0.9450	2.1695	1.7931	0.0190	0.0270	0.9540
	50	BC$_a$	3.5520	2.1322	0.0290	0.0260	0.9450	2.1701	1.7957	0.0190	0.0270	0.9540
Non param. sampling	50	NS	3.5028	0.4821	0.0000 ***	0.1040 ***	0.8960 ***	2.1332	0.5609	0.0000 ***	0.0770 ***	0.9230 ***
	50	S	2.9910	2.0568	0.0890 ***	0.0340	0.8770 ***	1.8567	1.7823	0.0700 ***	0.0330	0.8970 ***
	50	NT	3.5788	1.0000	0.0000 ***	0.0560 ***	0.9440	2.1648	1.0000	0.0000 ***	0.0450 ***	0.9550
	50	P	3.5028	2.0744	0.0320	0.0280	0.9400	2.1332	1.7828	0.0170	0.0300	0.9530
	50	BC	3.5390	2.0989	0.0300	0.0270	0.9430	2.1466	1.7841	0.0180	0.0290	0.9530
	50	BC$_a$	3.5533	2.1204	0.0310	0.0270	0.9420	2.1452	1.7845	0.0180	0.0290	0.9530
Krinsky and Robb sampling	50	NS	3.5218	0.4880	0.0000 ***	0.1050 ***	0.8950 ***	2.1598	0.5586	0.0000 ***	0.0760 ***	0.9240 ***
	50	S	3.0155	2.0431	0.0910 ***	0.0350 *	0.8740 ***	1.8520	1.7599	0.0690 ***	0.0410 **	0.8900 ***
	50	NT	3.6298	1.0000	0.0000 ***	0.0540 ***	0.9460	2.1870	1.0000	0.0000 ***	0.0470 ***	0.9530
	50	P	3.5218	2.0492	0.0280	0.0290	0.9430	2.1598	1.7903	0.0170	0.0230	0.9600
	50	BC	3.5089	2.0457	0.0280	0.0290	0.9430	2.1632	1.8021	0.0170	0.0240	0.9590
	50	BC$_a$	3.5290	2.0696	0.0290	0.0280	0.9430	2.1616	1.8047	0.0170	0.0240	0.9590

Table 3: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. Model simulated: MNL model with orthogonal experimental design. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -0.25$.

coverage rates for $N = 10$ (e.g. TIB and STIB, probably due to convergence problems, as discussed later). Pivotal bootstrap methods show their limits, no matter the sampling scheme used. Approximation methods and percentile bootstrap methods perform well and have good coverage rates. For this last family, and for $N = 10$, the non parametric sampling scheme produces slightly less satisfactory results with respect to Krinsky and Robb and parametric schemes. A possible explanation is that when such a small sample size is available, respondents may be sampled many times inducing unstable parameter estimates. The larger the sample size the better all methods perform[8].

Tables 2 and 3 show that β_C approaching 0 reduces the performance of most methods. In particular, already for $\beta_C = -0.5$, Delta produces LRP and RRP significantly different from $\alpha/2$ and CIs shifted towards 0, even if the total coverage rate is unaffected[9]. The problem persists even for $N = 25$ and, to some extent, also for $N = 50$. With $\beta_C = -0.25$ this shift is more marked and the total coverage rate deteriorates, falling well below the nominal level[10].

Fieller and LRTI both produce accurate CIs for $\beta_C = -0.5$ but show significantly smaller LRP values than $\alpha/2$ for $\beta_C = -0.25$ and $N = 10$. This problem disappears as N increases. Notice, however, that LRTI seems to perform slightly better than Fieller and its LRP recovers earlier to its nominal value as N raises.

Percentile bootstrap methods also work well for $\beta_C = -0.5$. Some problems are detected for the Krinsky and Robb and for the non parametric sampling scheme when $N = 10$. In this case, BC_a seems superior to P and BC. The performance of these three methods, however, deteriorates when β_C goes to -0.25. Despite BC_a seems to confirm its superiority, the three percentile methods give raise to LRPs significantly smaller than $\alpha/2$ producing overconservative

[8]Notice that, in all of the tables reported, significance levels are not corrected to account for multiple testing problems. Thus, coverage rates or rejection probabilities only significant at the 5% level should not be given too much credit.

[9]In Table 2, the t-statistic is equal to -3.85, -6.09 and -8.61, respectively for $N = 10$, $N = 25$ and $N = 50$.

[10]In Table 3, the t-statistic is equal to -2.22, -3.50 and -4.95, respectively for $N = 10$, $N = 25$ and $N = 50$.

CIs. However, these methods regain their accuracy well before Delta as N increases.

Pivotal bootstrap methods show very poor coverage rates, confirming their inadequacy in delivering reliable CIs. NT gives huge CIs for $N = 10$ and $\beta_C = -0.25$, due to the instability of WTP estimates across bootstrap samples determining an inflated bootstrap estimate of the standard error of $\widehat{WTP}$.

STIB performs reasonably well for $\beta_C = -0.5$ but not for $\beta_C = -0.25$, showing, in this last case, a counterintuitive worsening as N increases[11].

Looking at the shape index for the Monte Carlo CI in Tables 1, 2 and 3 one notices a positive asymmetry, which increases as β_C approaches 0 and decreases as N rises. The median shape obtained with all methods, except Delta and NT (both symmetric by construction), reflects such a positive asymmetry. The median length of the intervals behave as the median shape, increasing for small β_C or N. It is interesting to note that Delta produces, in general, the shortest CIs, which would be desirable where the coverage rate correct; this is not always the case for small β_C values. Fieller and LRTI are characterized by more satisfactory coverage rates and CIs of comparable lengths. Those produced by Fieller are slightly larger than those produced by LRTI, making the latter somehow preferable. Percentile methods produce intervals of length similar to Fieller and LRTI, except for $\beta_C = -0.25$ and $N = 10$, when they are much shorter but heavily shifted and with unsatisfactory coverage rates. The median length of bootstrap CIs based on non-parametric sampling is generally higher than that obtained using alternative resampling schemes.

Figure 2 reports the q-q plots of the sample quantiles of $\widehat{WTP}_1$ and $\widehat{WTP}_2$ when $\beta_C = -0.25$ with increasing N values, showing that such a small cost parameter provokes substantial departures from normality in the distribution of $\widehat{WTP}$. The sample distributions of $\widehat{WTP}_1$

[11] TIB and STIB require careful monitoring to assess convergence to interval limits. They turned out particularly sensitive to both step length and stopping rule. These issues might explain the controversial results obtained in the simulation study, casting doubts on convergence in some cases.

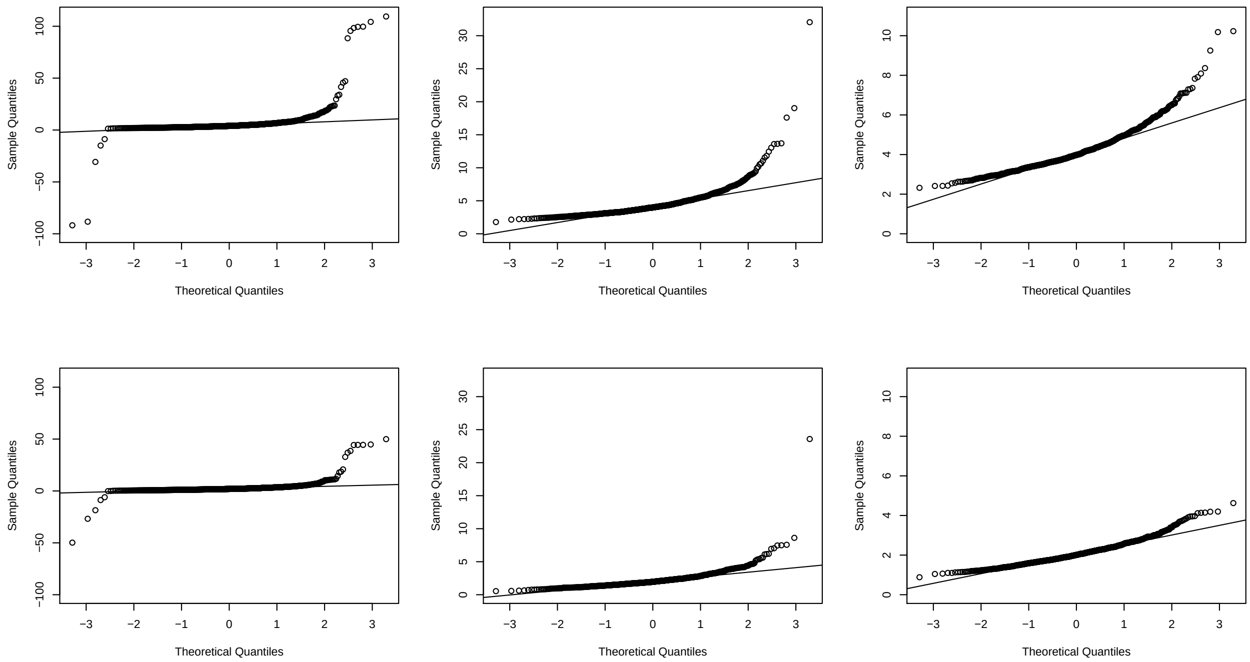

Figure 2: Quantile-quantile plots of the sample quantiles of $\widehat{WTP}_1$ (upper panel) and $\widehat{WTP}_2$ (bottom panel) for increasing values of N ($N = 10$, left panel; $N = 25$, central panel; $N = 50$, right panel). Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -0.25$.

and $\widehat{WTP}_2$ are positively skewed even for $N = 50$, notwithstanding skewness decreases as N rises. This might explain both the good performance of Fieller, LRTI and percentile methods, which do not rely on $\widehat{WTP}$ normality assumption, and the poor performance of Delta, rendering symmetric CIs. Additionally, Figure 2 reveals a larger departure from normality in the distribution of $\widehat{WTP}_1$ (top panel) compared to $\widehat{WTP}_2$ (bottom panel). This suggests that Delta reliability cannot simply rest on the coefficient of variation of WTP denominator. *Ceteris paribus*, the approximation of WTP distribution to normality improves as the coefficient of variation of the numerator increases explaining the overall improvement in performance of all methods for WTP_2 compared to WTP_1.

Figure 3 shows the CIs obtained through Delta, Fieller, LRTI and BC$_a$ on 10 different data sets simulated under $\beta_C = -0.25$ and $N = 50$ (same settings as Table 3). The last

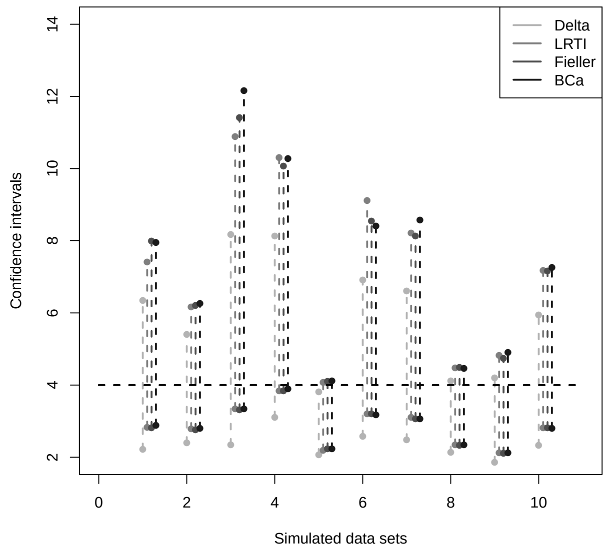

Figure 3: Confidence intervals for WTP_1 obtained through Delta, Fieller, LRTI and BC_a, for 10 different simulated data sets. $N = 50$. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -0.25$. Horizontal dashed line represents the true value of WTP_1.

three methods not only render CIs with the right coverage rate, LRP and RRP (see Table 3), but also produce very similar intervals for the single data set. CIs produced by Delta differ, in some cases, quite substantially. In particular, the shift towards 0 and the much smaller superior limits further underline Delta inability to account for positive skewness in the $\widehat{WTP}$ distribution.

4.2. Correlation between numerator and denominator estimates

Most of the literature investigating the conditions under which Delta is likely to work well have only focused on the coefficient of variation of the denominator. Nevertheless, in some cases, it is acknowledged that the correlation between numerator and denominator plays an important role in determining $\widehat{WTP}$ distribution (Hirschberg and Lye, 2010; Marsaglia, 2006). As shown in Hirschberg and Lye (2010), Delta and Fieller intervals may diverge even when the denominator has a high level of precision, if the sign of the estimated correlation between the numerator and denominator is the same as that of $\widehat{WTP}$. The performance of all the methods under such a situation is illustrated by inducing positive correlation between the numerator and denominator estimates through a non-orthogonal experimental design. In particular, the first level of X_1 is associated with any level of X_C except the first one, and the second level of X_1 is associated with any level of X_C except the fourth one[12]. This setting gives $\text{cor}(X_1, X_C) = -0.589$ and introduces, as a side effect, a negligible negative correlation between X_2 and X_C, i.e. $\text{cor}(X_2, X_C) = -0.119$. Letting $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$, as in Table 1, one obtains $\sigma_{\hat{\beta}_1,\hat{\beta}_C} = 0.555$ and $\sigma_{\hat{\beta}_2,\hat{\beta}_C} = -0.390$.

Table 4 shows the effects of positive correlation between the numerator and denominator of $\widehat{WTP}$. Delta produces unsatisfactory CIs for WTP_1 in terms of LRP, RRP and total coverage rate. The LRP, even for $N = 50$, is significantly different from its nominal level, while Fieller, LRTI, BC and BC$_a$ perform well independently of sample size[13]. Additionally, Delta intervals for WTP_1 include the value 0 in 73.5% of the cases when $N = 10$, while such a percentage ranges from 27.2% to 32.0% when using Fieller, LRTI, BC and BC$_a$. Since the coefficient of variation for $\hat{\beta}_1$ is 0.380 (t-statistic equal to 2.631), this implies Delta often produces WTP_1 intervals including 0 when the numerator of WTP_1 is significantly different from 0. This is counterintuitive and might have serious implications on policy making decisions when WTP measures are used as a benchmark. Delta intervals with $N = 25$ still contain 0 in 5.6% of

[12]This apparently counterintuitive situation might, under some circumstances, happen in real-life service industries production. In fact, one could have a raising cost associated with a decreasing level in a desirable attribute of the available alternatives. In public transportation services, for instance, this might happen with respect to frequency, comfort and price. In fact, imagine a situation where price is high and frequency is high during peak time, while the opposite is true in off-peak. Now suppose that the rationing effect of the increase in price is not sufficient to improve the level of comfort which remains low, while comfort is high during off-peak. In this case, one would witness a negative correlation between comfort and price. Additionally, the same phenomenon might appear also in stated preference data due both to efficient, non-orthogonal experimental designs and missing responses.

[13]The only exception being non parametric sampling for BC and BC$_a$ for $N = 10$.

			$\widehat{WTP}_1$					$\widehat{WTP}_2$				
	N	Method	Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
	10	Monte Carlo	2.4135	1.5135	0.0250	0.0250	0.9500	1.2654	1.0584	0.0250	0.0250	0.9500
	10	Delta	2.2165	1.0000	0.0000 ***	0.0780 ***	0.9220 ***	1.1861	1.0000	0.0090 **	0.0220	0.9690 **
	10	LRTI	2.5000	2.1429	0.0230	0.0350 *	0.9420	1.3281	1.1082	0.0310	0.0290	0.9400
	10	Fieller	2.7841	2.4090	0.0180	0.0220	0.9600	1.3604	1.0894	0.0250	0.0210	0.9540
Parametric sampling	10	TIB	3.3362	2.3762	0.0060 ***	0.0170	0.9770 ***	1.2182	0.9835	0.0120 **	0.0220	0.9660 *
	10	STIB	2.8687	2.5036	0.0560 ***	0.0150 *	0.9290 **	1.4341	1.0890	0.0290	0.0220	0.9490
	10	NS	2.4930	0.5180	0.0000 ***	0.1200 ***	0.8800 ***	1.2928	0.8825	0.0010 ***	0.0130 *	0.9860 ***
	10	S	2.4167	2.2684	0.0740 ***	0.0270	0.8990 ***	1.1544	1.0765	0.0700 ***	0.0280	0.9020 ***
	10	NT	2.5713	1.0000	0.0000 ***	0.0680 ***	0.9320 **	1.2808	1.0000	0.0040 ***	0.0190	0.9770 ***
	10	P	2.4930	1.9305	0.0140 *	0.0430 ***	0.9430	1.2928	1.1331	0.0290	0.0230	0.9480
	10	BC	2.5850	2.0892	0.0170	0.0320	0.9510	1.2980	1.0624	0.0360 *	0.0280	0.9360 *
	10	BC$_a$	2.5347	2.0117	0.0160	0.0330	0.9510	1.2970	1.0659	0.0360 *	0.0250	0.9390
Non param. sampling	10	NS	2.5032	0.5215	0.0000 ***	0.1260 ***	0.8740 ***	1.2893	0.8685	0.0000 ***	0.0110 **	0.9890 ***
	10	S	2.4148	2.3259	0.0810 ***	0.0350 *	0.8840 ***	1.1382	1.0686	0.0710 ***	0.0410 **	0.8880 ***
	10	NT	2.5678	1.0000	0.0000 ***	0.0750 ***	0.9250 ***	1.2919	1.0000	0.0050 ***	0.0130 *	0.9820 ***
	10	P	2.5032	1.9174	0.0130 *	0.0470 ***	0.9400	1.2893	1.1515	0.0320	0.0190	0.9490
	10	BC	2.5675	2.0745	0.0140 *	0.0400 **	0.9460	1.2940	1.0914	0.0350 *	0.0270	0.9380
	10	BC$_a$	2.5141	1.9675	0.0130 *	0.0410 **	0.9460	1.2926	1.1002	0.0350 *	0.0260	0.9390
Krinsky and Robb sampling	10	NS	2.7973	0.4116	0.0000 ***	0.1270 ***	0.8730 ***	1.3643	0.9113	0.0000 ***	0.0030 ***	0.9970 ***
	10	S	2.1271	2.3431	0.1040 ***	0.0430 ***	0.8530 ***	1.0632	1.0819	0.0780 ***	0.0460 ***	0.8760 ***
	10	NT	3.1509	1.0000	0.0000 ***	0.0460 ***	0.9540	1.3777	1.0000	0.0010 ***	0.0080 ***	0.9910 ***
	10	P	2.7973	2.4294	0.0170	0.0230	0.9600	1.3643	1.0973	0.0280	0.0200	0.9520
	10	BC	2.7843	2.4007	0.0180	0.0220	0.9600	1.3691	1.0880	0.0270	0.0210	0.9520
	10	BC$_a$	2.7204	2.2957	0.0170	0.0240	0.9590	1.3722	1.0880	0.0270	0.0190	0.9540
	25	Monte Carlo	1.4374	1.3413	0.0250	0.0250	0.9500	0.7921	1.0891	0.0250	0.0250	0.9500
	25	Delta	1.3953	1.0000	0.0040 ***	0.0600 ***	0.9360 *	0.7469	1.0000	0.0250	0.0270	0.9480
	25	LRTI	1.4453	1.6000	0.0230	0.0340	0.9430	0.7764	1.0537	0.0320	0.0280	0.9400
	25	Fieller	1.5098	1.6747	0.0220	0.0220	0.9560	0.7830	1.0434	0.0290	0.0280	0.9430
Parametric sampling	25	TIB	1.5213	1.4306	0.0090 **	0.0260	0.9650 *	0.7566	0.9725	0.0190	0.0270	0.9540
	25	STIB	1.4959	1.6545	0.0300	0.0230	0.9470	0.8007	1.0347	0.0310	0.0260	0.9430
	25	NS	1.4547	0.6675	0.0000 ***	0.0800 ***	0.9200 ***	0.7687	0.9292	0.0080 ***	0.0200	0.9720 **
	25	S	1.4387	1.6288	0.0320	0.0250	0.9430	0.7345	1.0282	0.0390 **	0.0290	0.9320 **
	25	NT	1.4583	1.0000	0.0000 ***	0.0560 ***	0.9440	0.7647	1.0000	0.0200	0.0250	0.9550
	25	P	1.4547	1.4980	0.0180	0.0310	0.9510	0.7687	1.0762	0.0290	0.0270	0.9440
	25	BC	1.4745	1.5657	0.0190	0.0260	0.9550	0.7708	1.0274	0.0290	0.0300	0.9410
	25	BC$_a$	1.4617	1.5080	0.0160	0.0300	0.9540	0.7704	1.0270	0.0290	0.0290	0.9420
Non param. sampling	25	NS	1.4589	0.6644	0.0000 ***	0.0810 ***	0.9190 ***	0.7690	0.9322	0.0080 ***	0.0230	0.9690 **
	25	S	1.4315	1.6327	0.0330	0.0250	0.9420	0.7303	1.0286	0.0440 ***	0.0360 *	0.9200 ***
	25	NT	1.4608	1.0000	0.0000 ***	0.0550 ***	0.9450	0.7665	1.0000	0.0160	0.0240	0.9600
	25	P	1.4589	1.5051	0.0180	0.0310	0.9510	0.7690	1.0728	0.0300	0.0240	0.9460
	25	BC	1.4769	1.5639	0.0210	0.0280	0.9510	0.7701	1.0310	0.0320	0.0300	0.9380
	25	BC$_a$	1.4625	1.5029	0.0180	0.0270	0.9550	0.7700	1.0308	0.0310	0.0280	0.9410
Krinsky and Robb sampling	25	NS	1.5194	0.5928	0.0000 ***	0.0870 ***	0.9130 ***	0.7842	0.9496	0.0080 ***	0.0150 *	0.9770 ***
	25	S	1.3757	1.6666	0.0430 ***	0.0300	0.9270 ***	0.7178	1.0447	0.0430 ***	0.0380 **	0.9190 ***
	25	NT	1.5275	1.0000	0.0000 ***	0.0480 ***	0.9520	0.7785	1.0000	0.0140 *	0.0220	0.9640 *
	25	P	1.5194	1.6869	0.0230	0.0240	0.9530	0.7842	1.0531	0.0300	0.0280	0.9420
	25	BC	1.5240	1.6802	0.0260	0.0240	0.9500	0.7862	1.0398	0.0290	0.0280	0.9430
	25	BC$_a$	1.5067	1.6109	0.0220	0.0250	0.9530	0.7863	1.0377	0.0300	0.0280	0.9420
	50	Monte Carlo	0.9882	1.3020	0.0250	0.0250	0.9500	0.5455	1.0106	0.0250	0.0250	0.9500
	50	Delta	0.9943	1.0000	0.0040 ***	0.0340	0.9620	0.5283	1.0000	0.0250	0.0300	0.9450
	50	LRTI	0.9961	1.3846	0.0210	0.0230	0.9560	0.5322	1.0297	0.0290	0.0320	0.9390
	50	Fieller	1.0330	1.4315	0.0250	0.0200	0.9550	0.5414	1.0311	0.0270	0.0290	0.9440
Parametric sampling	50	TIB	1.0422	1.2809	0.0130 *	0.0180	0.9690 **	0.5375	0.9907	0.0240	0.0290	0.9470
	50	STIB	1.0353	1.4165	0.0230	0.0200	0.9570	0.5553	1.0159	0.0270	0.0340	0.9390
	50	NS	1.0188	0.7492	0.0000 ***	0.0490 ***	0.9510	0.5355	0.9565	0.0120 **	0.0290	0.9590
	50	S	1.0095	1.4035	0.0250	0.0220	0.9530	0.5244	1.0190	0.0330	0.0360 *	0.9310 **
	50	NT	1.0132	1.0000	0.0040 ***	0.0300	0.9660 *	0.5331	1.0000	0.0220	0.0280	0.9500
	50	P	1.0188	1.3347	0.0210	0.0250	0.9540	0.5355	1.0454	0.0300	0.0280	0.9420
	50	BC	1.0241	1.3660	0.0200	0.0240	0.9560	0.5371	1.0178	0.0260	0.0320	0.9420
	50	BC$_a$	1.0220	1.3289	0.0190	0.0240	0.9570	0.5368	1.0168	0.0270	0.0320	0.9410
Non param. sampling	50	NS	1.0206	0.7501	0.0010 ***	0.0470 ***	0.9520	0.5371	0.9533	0.0170	0.0310	0.9520
	50	S	1.0099	1.4088	0.0230	0.0240	0.9530	0.5225	1.0138	0.0310	0.0370 *	0.9320 **
	50	NT	1.0195	1.0000	0.0050 ***	0.0340	0.9610	0.5353	1.0000	0.0250	0.0280	0.9470
	50	P	1.0206	1.3331	0.0200	0.0250	0.9550	0.5371	1.0490	0.0320	0.0280	0.9400
	50	BC	1.0240	1.3652	0.0240	0.0250	0.9510	0.5371	1.0160	0.0320	0.0310	0.9370
	50	BC$_a$	1.0194	1.3249	0.0210	0.0250	0.9540	0.5370	1.0151	0.0320	0.0310	0.9370
Krinsky and Robb sampling	50	NS	1.0289	0.6946	0.0000 ***	0.0510 ***	0.9490	0.5407	0.9653	0.0110 **	0.0270	0.9620
	50	S	0.9849	1.4242	0.0260	0.0220	0.9520	0.5186	1.0272	0.0350 *	0.0350 *	0.9300 **
	50	NT	1.0323	1.0000	0.0020 ***	0.0310	0.9670 *	0.5387	1.0000	0.0210	0.0270	0.9520
	50	P	1.0289	1.4396	0.0240	0.0210	0.9550	0.5407	1.0359	0.0290	0.0290	0.9420
	50	BC	1.0305	1.4358	0.0230	0.0210	0.9560	0.5422	1.0284	0.0290	0.0280	0.9430
	50	BC$_a$	1.0258	1.3932	0.0210	0.0220	0.9570	0.5415	1.0282	0.0290	0.0280	0.9430

Table 4: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals. Significance codes: *** for p-value $<$ 0.001; ** for p-value $<$ 0.01; * for p-value $<$ 0.05. Model simulated: MNL model with non-orthogonal experimental design. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$.

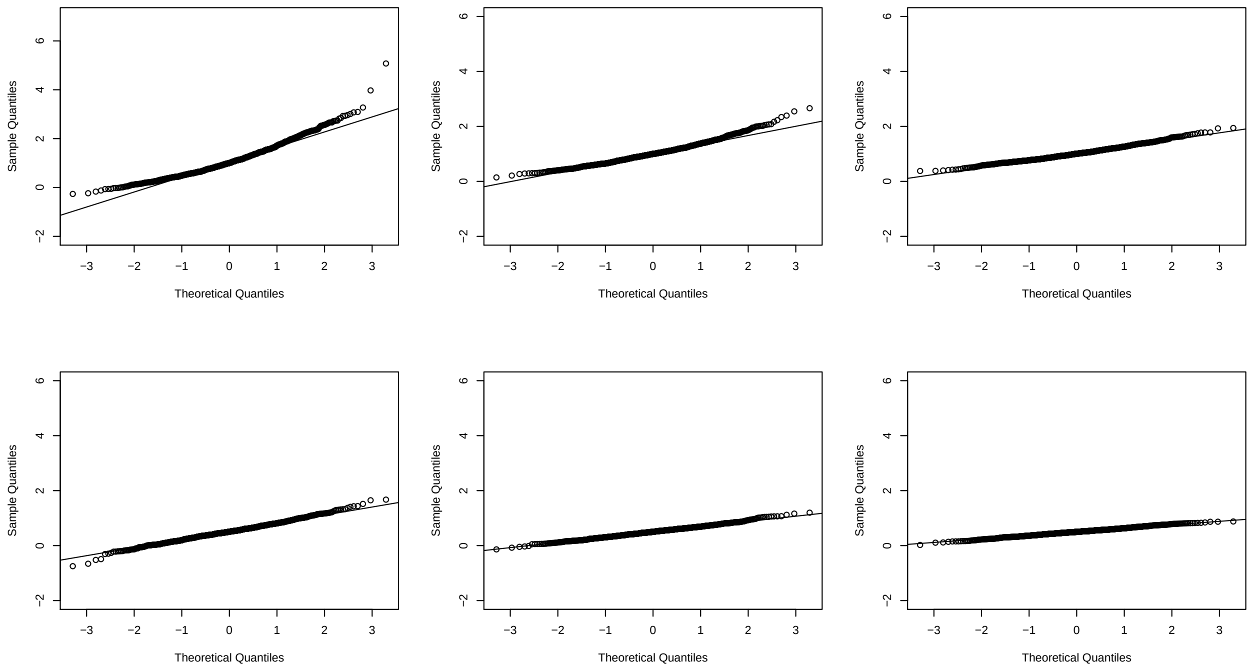

Figure 4: Quantile-quantile plots of the sample quantiles of $\widehat{WTP}_1$ (upper panel) and $\widehat{WTP}_2$ (bottom panel) for increasing values of N ($N = 10$, left panel; $N = 25$, central panel; $N = 50$, right panel). Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$. Non-orthogonal design.

cases, a percentage almost three times as large as that produced by using other methods. Notice that Delta yields unreliable CIs also for WTP_2 with $N = 10$. This is not due to the correlation between the numerator and denominator, which has, this time, different sign from WTP_2 (a situation in which Delta is expected to perform well). It is rather due to the diminished precision in parameter estimates, with respect to Table 1, as a consequence of the correlation induced among the attributes. This is in line with the results in Hole (2007), who found Delta intervals perform poorly as the precision of the estimates decreases. Unlike Hole (2007), however, there is no evidence of coverage rates for P being lower than the nominal level.

Figure 4 shows the effects on $\widehat{WTP}$ distribution of a correlation between the numerator and denominator having the same sign of $\widehat{WTP}$ (top panels) and of a lack of precision in parameter

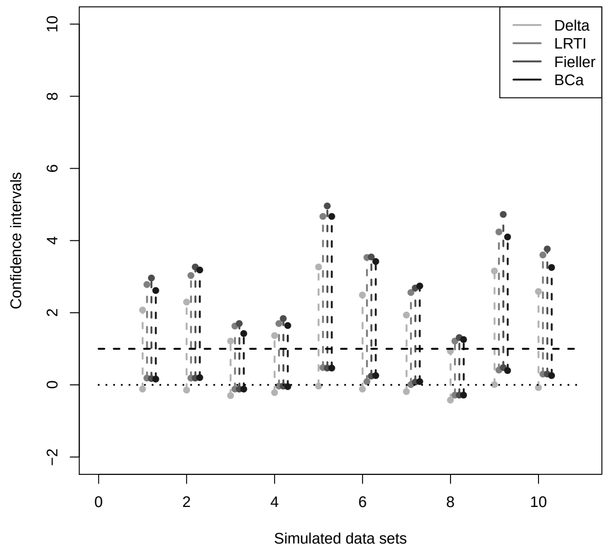

Figure 5: Confidence intervals for WTP_1 obtained through Delta, Fieller, LRTI and parametric BCa, for 10 different simulated data sets. $N = 10$. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$. Non-orthogonal design. Horizontal dashed line represents the true value of WTP_1. Dotted line corresponds to $WTP_1 = 0$.

estimates (bottom panels). While the first issue determines a positively skewed distribution, the second causes an overdispersed $\widehat{WTP}$ distribution with respect to the normal density[14]. Similar considerations emerge when looking at the shape index for the Monte Carlo intervals in Table 4. In fact, it is larger than 1 for $\widehat{WTP}_1$, decreasing as N increases, while it is always very close to 1 for $\widehat{WTP}_2$.

Figure 5 illustrates both the close agreement between CIs for WTP_1 produced through Fieller, LRTI and parametric BC$_a$, and the shift towards 0 for those obtained via Delta ($N = 10$). As noticed in the previous scenario, such a shift determines a higher inclusion of the 0 value.

[14]This happens for $N = 10$, in the first bottom panel, while heavy tails disappear as the estimates become more precise with the increase of N.

4.3. Incorrect model specification

This section investigates methods' performance under model mis-specification. Hole (2007) considers the case of neglected unobserved individual heterogeneity, finding P more reliable than Delta and Fieller. Here, the case of neglected heteroscedasticity, caused by hypothetical unobserved discrete population heterogeneity, is examined. This is accomplished by considering two groups and letting the scale parameter of the second (μ_2) fixed to 1 (i.e. $\text{var}(\epsilon) = \pi^2/6$), while that of the first (μ_1) takes values 2, 3 and 4. A MNL model is estimated *without* accounting for heteroscedasticity[15].

Table 5 shows a good global coverage for all the methods when μ_1 is not too far from μ_2 (i.e. μ_1=2). In this case, however, LRP and RRP are somewhat different from expected and this is more pronounced for Delta with respect to LRTI, Fieller and percentile methods. As the degree of heteroscedasticity increases the global coverage worsen for all methods, even if more slowly for LRTI, Fieller and percentile methods. Unlike Hole (2007), in this case bootstrap methods do not seem more robust than LRTI or Fieller.

5. Real data applications

A MNL model is estimated on two real data sets to compare the methods described in Section 3. The choice of the two data sets is motivated by their respective similarity with some of the test settings used in the simulation study. In fact, the first data set is characterized by a relatively small number of observations, potential correlation due to revealed preference data structure and high coefficient of variation for the cost parameter estimate. The second data set, not affected by such issues, should be less problematic.

[15] The scale parameter does not affect the ratio of any two coefficients, since it drops out in the ratio, so that WTP and other measures of marginal rates of substitution are not affected. Only the magnitudes of all coefficients are affected.

			$\widehat{WTP}_1$					$\widehat{WTP}_2$				
	μ_1	Method	Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
	2	Monte Carlo	0.4983	0.9798	0.0250	0.0250	0.9500	0.4950	1.1413	0.0250	0.0250	0.9500
	2	Delta	0.5130	1.0000	0.0120 **	0.0410 **	0.9470	0.4901	1.0000	0.0220	0.0250	0.9530
	2	LRTI	0.5176	1.0909	0.0140 *	0.0370 *	0.9490	0.5029	1.0513	0.0250	0.0230	0.9520
	2	Fieller	0.5203	1.1083	0.0140 *	0.0350 *	0.9510	0.4969	1.0590	0.0240	0.0220	0.9540
Parametric sampling	2	TIB	0.5223	1.0089	0.0120 **	0.0430 ***	0.9450	0.4972	1.0015	0.0230	0.0240	0.9530
	2	STIB	0.5269	1.1099	0.0110 **	0.0320	0.9570	0.5053	1.0550	0.0250	0.0190	0.9560
	2	NS	0.5162	0.9246	0.0090 **	0.0450 ***	0.9460	0.4921	0.9562	0.0200	0.0260	0.9540
	2	S	0.5151	1.0910	0.0150 *	0.0370 *	0.9480	0.4908	1.0529	0.0250	0.0230	0.9520
	2	NT	0.5170	1.0000	0.0100 **	0.0410 **	0.9490	0.4923	1.0000	0.0220	0.0240	0.9540
	2	P	0.5162	1.0815	0.0140 *	0.0380 **	0.9480	0.4921	1.0458	0.0240	0.0250	0.9510
	2	BC	0.5171	1.0802	0.0150 *	0.0370 *	0.9480	0.4931	1.0469	0.0230	0.0250	0.9520
	2	BC_a	0.5168	1.0672	0.0120 **	0.0380 **	0.9500	0.4933	1.0386	0.0230	0.0250	0.9520
Non param. sampling	2	NS	0.5157	0.9246	0.0070 ***	0.0460 ***	0.9470	0.4926	0.9558	0.0170	0.0260	0.9570
	2	S	0.5114	1.0957	0.0130 *	0.0330	0.9540	0.4900	1.0524	0.0250	0.0240	0.9510
	2	NT	0.5161	1.0000	0.0100 **	0.0390 **	0.9510	0.4912	1.0000	0.0220	0.0230	0.9550
	2	P	0.5157	1.0815	0.0130 *	0.0370 *	0.9500	0.4926	1.0463	0.0240	0.0250	0.9510
	2	BC	0.5168	1.0814	0.0130 *	0.0370 *	0.9500	0.4948	1.0503	0.0240	0.0240	0.9520
	2	BC_a	0.5166	1.0658	0.0130 *	0.0370 *	0.9500	0.4946	1.0424	0.0230	0.0240	0.9530
Krinsky and Robb sampling	2	NS	0.5196	0.8979	0.0050 ***	0.0470 ***	0.9480	0.4961	0.9387	0.0180	0.0270	0.9550
	2	S	0.5083	1.1038	0.0160	0.0390 **	0.9450	0.4848	1.0522	0.0260	0.0250	0.9490
	2	NT	0.5189	1.0000	0.0120 **	0.0410 **	0.9470	0.4950	1.0000	0.0210	0.0230	0.9560
	2	P	0.5196	1.1137	0.0150 *	0.0350 *	0.9500	0.4961	1.0653	0.0230	0.0220	0.9550
	2	BC	0.5204	1.1115	0.0150 *	0.0370 *	0.9480	0.4975	1.0581	0.0240	0.0220	0.9540
	2	BC_a	0.5198	1.0960	0.0150 *	0.0370 *	0.9480	0.4973	1.0514	0.0210	0.0220	0.9570
	3	Monte Carlo	0.6025	1.0958	0.0250	0.0250	0.9500	0.5336	0.9754	0.0250	0.0250	0.9500
	3	Delta	0.5939	1.0000	0.0040 ***	0.0590 ***	0.9370	0.5650	1.0000	0.0100 **	0.0380 **	0.9520
	3	LRTI	0.6055	1.1193	0.0120 **	0.0470 ***	0.9410	0.5640	1.0702	0.0230	0.0290	0.9480
	3	Fieller	0.6044	1.1326	0.0060 ***	0.0480 ***	0.9460	0.5742	1.0702	0.0160	0.0320	0.9520
Parametric sampling	3	TIB	0.6035	0.9925	0.0060 ***	0.0530 ***	0.9410	0.5735	1.0068	0.0120 **	0.0360 *	0.9520
	3	STIB	0.6113	1.1316	0.0130 *	0.0510 ***	0.9360 *	0.5779	1.0642	0.0130 *	0.0320	0.9550
	3	NS	0.6007	0.9043	0.0030 ***	0.0670 ***	0.9300 **	0.5683	0.9440	0.0070 ***	0.0380 **	0.9550
	3	S	0.5948	1.1218	0.0070 ***	0.0540 ***	0.9390	0.5637	1.0630	0.0180	0.0370 *	0.9450
	3	NT	0.6007	1.0000	0.0040 ***	0.0550 ***	0.9410	0.5691	1.0000	0.0090 **	0.0390 **	0.9520
	3	P	0.6007	1.1058	0.0060 ***	0.0520 ***	0.9420	0.5683	1.0593	0.0160	0.0330	0.9510
	3	BC	0.6036	1.1112	0.0070 ***	0.0510 ***	0.9420	0.5691	1.0572	0.0160	0.0330	0.9510
	3	BC_a	0.6031	1.0975	0.0070 ***	0.0510 ***	0.9420	0.5689	1.0518	0.0160	0.0340	0.9500
Non param. sampling	3	NS	0.5992	0.9072	0.0020 ***	0.0700 ***	0.9280 **	0.5670	0.9459	0.0070 ***	0.0400 **	0.9530
	3	S	0.5898	1.1225	0.0070 ***	0.0520 ***	0.9410	0.5609	1.0593	0.0180	0.0370 *	0.9450
	3	NT	0.5957	1.0000	0.0040 ***	0.0600 ***	0.9360 *	0.5667	1.0000	0.0080 ***	0.0380 **	0.9540
	3	P	0.5992	1.1022	0.0060 ***	0.0520 ***	0.9420	0.5670	1.0572	0.0160	0.0310	0.9530
	3	BC	0.6010	1.1146	0.0050 ***	0.0500 ***	0.9450	0.5687	1.0631	0.0150 *	0.0360 *	0.9490
	3	BC_a	0.6011	1.1010	0.0050 ***	0.0510 ***	0.9440	0.5685	1.0560	0.0150 *	0.0360 *	0.9490
Krinsky and Robb sampling	3	NS	0.6036	0.8760	0.0030 ***	0.0690 ***	0.9280 **	0.5757	0.9303	0.0070 ***	0.0400 **	0.9530
	3	S	0.5877	1.1274	0.0060 ***	0.0510 ***	0.9430	0.5592	1.0683	0.0170	0.0390 **	0.9440
	3	NT	0.6015	1.0000	0.0040 ***	0.0600 ***	0.9360 *	0.5733	1.0000	0.0100 **	0.0360 *	0.9540
	3	P	0.6036	1.1416	0.0060 ***	0.0470 ***	0.9470	0.5757	1.0750	0.0160	0.0310	0.9530
	3	BC	0.6047	1.1293	0.0070 ***	0.0490 ***	0.9440	0.5760	1.0699	0.0160	0.0300	0.9540
	3	BC_a	0.6043	1.1155	0.0070 ***	0.0520 ***	0.9410	0.5761	1.0630	0.0150 *	0.0310	0.9540
	4	Monte Carlo	0.6249	1.1421	0.0250	0.0250	0.9500	0.5482	1.0644	0.0250	0.0250	0.9500
	4	Delta	0.6494	1.0000	0.0030 ***	0.0770 ***	0.9200 ***	0.6166	1.0000	0.0100 **	0.0210	0.9690 **
	4	LRTI	0.6702	1.1311	0.0080 ***	0.0730 ***	0.9190 ***	0.6250	1.0720	0.0170	0.0270	0.9560
	4	Fieller	0.6627	1.1461	0.0040 ***	0.0570 ***	0.9390	0.6283	1.0782	0.0130 *	0.0170	0.9700 **
Parametric sampling	4	TIB	0.6591	1.0108	0.0030 ***	0.0740 ***	0.9230 ***	0.6241	1.0004	0.0060 ***	0.0200	0.9740 ***
	4	STIB	0.6725	1.1427	0.0070 ***	0.0580 ***	0.9350 *	0.6360	1.0806	0.0130 *	0.0200	0.9670 *
	4	NS	0.6603	0.8858	0.0010 ***	0.0810 ***	0.9180 ***	0.6230	0.9352	0.0050 ***	0.0180	0.9770 ***
	4	S	0.6481	1.1344	0.0040 ***	0.0630 ***	0.9330 *	0.6153	1.0678	0.0140 *	0.0180	0.9680 **
	4	NT	0.6583	1.0000	0.0030 ***	0.0730 ***	0.9240 ***	0.6239	1.0000	0.0070 ***	0.0190	0.9740 ***
	4	P	0.6603	1.1290	0.0050 ***	0.0630 ***	0.9320 **	0.6230	1.0692	0.0120 **	0.0190	0.9690 **
	4	BC	0.6611	1.1290	0.0040 ***	0.0610 ***	0.9350 *	0.6237	1.0678	0.0120 **	0.0190	0.9690 **
	4	BC_a	0.6598	1.1187	0.0030 ***	0.0630 ***	0.9340 *	0.6232	1.0613	0.0120 **	0.0190	0.9690 **
Non param. sampling	4	NS	0.6595	0.8904	0.0000 ***	0.0840 ***	0.9160 ***	0.6253	0.9339	0.0040 ***	0.0200	0.9760 ***
	4	S	0.6462	1.1343	0.0050 ***	0.0670 ***	0.9280 **	0.6146	1.0694	0.0160	0.0190	0.9650 *
	4	NT	0.6574	1.0000	0.0020 ***	0.0730 ***	0.9250 ***	0.6222	1.0000	0.0080 ***	0.0180	0.9740 ***
	4	P	0.6595	1.1231	0.0040 ***	0.0630 ***	0.9330 *	0.6253	1.0708	0.0130 *	0.0190	0.9680 **
	4	BC	0.6592	1.1261	0.0040 ***	0.0630 ***	0.9330 *	0.6270	1.0644	0.0120 **	0.0190	0.9690 **
	4	BC_a	0.6586	1.1151	0.0040 ***	0.0640 ***	0.9320 **	0.6265	1.0565	0.0120 **	0.0190	0.9690 **
Krinsky and Robb sampling	4	NS	0.6624	0.8655	0.0000 ***	0.0840 ***	0.9160 ***	0.6280	0.9198	0.0030 ***	0.0190	0.9780 ***
	4	S	0.6410	1.1394	0.0060 ***	0.0680 ***	0.9260 ***	0.6083	1.0729	0.0160	0.0200	0.9640 *
	4	NT	0.6597	1.0000	0.0030 ***	0.0730 ***	0.9240 ***	0.6256	1.0000	0.0080 ***	0.0200	0.9720 **
	4	P	0.6624	1.1554	0.0040 ***	0.0570 ***	0.9390	0.6280	1.0872	0.0120 **	0.0210	0.9670 *
	4	BC	0.6633	1.1494	0.0040 ***	0.0580 ***	0.9380	0.6291	1.0744	0.0120 **	0.0190	0.9690 **
	4	BC_a	0.6631	1.1388	0.0040 ***	0.0600 ***	0.9360 *	0.6292	1.0677	0.0120 **	0.0190	0.9690 **

Table 5: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. Model simulated: heteroscedastic MNL model arising from two populations with different scale parameters. $N = 50$. Parameter values: $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_C = -1$.

5.1. Data description

The first data set refers to a study of airport choice in a multi-airport region with the intent of exploring competition within a specific catchment area (Marcucci and Gatta, 2012). Data acquisition was based on a stated/revealed preference choice experiment describing a choice situation among four regional airports. Each interview included a revealed preference choice task and five hypothetical choice exercises in which respondents were asked to evaluate the four airports and choose the preferred one. The study area considered includes two regions in central Italy and four airports which are all located within the same catchment area. In order to detect the effect of correlation between numerator and denominator estimates, in the present study only revealed preference data are used, for a total of 176 binary responses[16]. The structural variables used are: A_MIN (access time in minutes); P_AIRL (1 for the preferred airline company and 0 otherwise); F_EURO (ticket cost in euros); NONSTOP (1 when the flight is non-stop and 0 otherwise); BAL_M_AV (absolute value of the difference between desired and actual departure time in minutes).

The second data set refers to a study focusing on local public transportation quality in five geographical areas of the Pesaro-Urbino province. The research produced quality indicators to be included in service contracts (Gatta and Marcucci, 2007). The interviewees had to choose, in eight stated preference exercises, among three options, the status quo and two hypothetical alternatives. The following five attributes were used to characterize service quality: COST (bus fare); DELAY (amount of delay at bus stop); TRIP LENGTH (bus travel time); FREQUENCY (number of buses per hour); AVAILABILITY (elapse of time between service inception and closure). An orthogonal fractional factorial design was developed, ensuring minimum attribute overlap. Overall, for the five geographical service segments, 2112 observations were gathered through paper-and-pencil interviews administered either on board or at the bus stops associated with the main routes.

[16]Stated preference data are excluded since based on an orthogonal fractional factorial experimental design.

5.2. WTP confidence intervals

Table 6 reports parameter estimates for the airport choice data set. All the coefficients are statistically significant at the 5% level, with the only exception being P_AIRL, and have the expected sign.

Attribute	Estimate	Std. Error	t-value	p-value
A_MIN	-0.0133	0.0055	-2.4121	0.0159
P_AIRL	0.9306	0.4938	1.8846	0.0595
NONSTOP	2.6298	0.4815	5.4612	0.0000
BAL_M_AV	-0.0038	0.0017	-2.2626	0.0237
COST	-0.0060	0.0019	-3.1428	0.0017

Table 6: Airport choice data: point estimate of attribute coefficients.

Table 7 reports CIs for the WTP obtained for all attributes, using the various methods.

		A_MIN	P_AIRL	NONSTOP	BAL_M_AV
	Delta	$[-4.516; 0.057]$	$[-15.041; 327.292]$	$[120.402; 762.024]$	$[-1.317; 0.036]$
	LRTI	$[-6.867; -0.476]$	$[-1.108; 429.885]$	$[232.286; 1202.873]$	$[-1.929; -0.100]$
	Fieller	$[-6.906; -0.400]$	$[-7.233; 449.617]$	$[226.466; 1225.493]$	$[-1.988; -0.084]$
Parametric sampling	TIB	$[-4.401; -0.197]$	$[-16.963; 304.926]$	$[118.441; 793.521]$	$[-1.252; 0.175]$
	STIB	$[-7.262; -0.384]$	$[-6.220; 376.290]$	$[275.885; 1066.060]$	$[-1.656; -0.150]$
	NS	$[-3.981; 1.972]$	$[-93.774; 318.680]$	$[-344.128; 646.919]$	$[-1.182; 0.623]$
	S	$[-5.450; -0.932]$	$[29.613; 362.006]$	$[294.323; 922.741]$	$[-1.540; -0.223]$
	NT	$[-6.243; 1.784]$	$[-69.231; 381.483]$	$[-231.720; 1114.146]$	$[-1.785; 0.503]$
	P	$[-6.432; -0.478]$	$[-6.429; 406.025]$	$[235.507; 1226.554]$	$[-1.904; -0.099]$
	BC	$[-6.703; -0.493]$	$[4.684; 426.746]$	$[238.254; 1272.613]$	$[-1.987; -0.104]$
	BC_a	$[-7.876; -0.643]$	$[5.530; 430.836]$	$[225.969; 1104.074]$	$[-1.806; -0.088]$
Non param. sampling	NS	$[-94.952; 66.013]$	$[-3511.334; 4599.439]$	$[-15942.351; 18826.957]$	$[-27.089; 22.905]$
	S	$[-19.210; -1.073]$	$[62.899; 479.804]$	$[295.250; 2445.603]$	$[-4.547; -0.320]$
	NT	$[-877.451; 872.992]$	$[-14528.753; 14841.004]$	$[-187404.206; 188286.632]$	$[-342.869; 341.588]$
	P	$[-70.472; 90.493]$	$[-4287.188; 3823.585]$	$[-17944.531; 16824.777]$	$[-24.186; 25.807]$
	BC	$[-232.835; 27.868]$	$[-172.023; 17291.169]$	$[-5302.960; 43624.263]$	$[-46.353; 10.631]$
	BC_a	$[-572.818; 0.284]$	$[-150.832; 17291.169]$	$[-7461.532; 36338.675]$	$[-34.700; 14.676]$
Krinsky and Robb sampling	NS	$[-3.970; 2.008]$	$[-113.276; 316.759]$	$[-279.565; 661.569]$	$[-1.183; 0.605]$
	S	$[-5.083; -0.899]$	$[42.793; 338.492]$	$[291.192; 931.942]$	$[-1.449; -0.249]$
	NT	$[-6.769; 2.310]$	$[-72.899; 385.150]$	$[-375.465; 1257.891]$	$[-2.149; 0.868]$
	P	$[-6.467; -0.489]$	$[-4.508; 425.528]$	$[220.857; 1161.991]$	$[-1.887; -0.098]$
	BC	$[-6.518; -0.489]$	$[1.256; 438.677]$	$[217.984; 1150.518]$	$[-1.861; -0.091]$
	BC_a	$[-7.449; -0.624]$	$[2.220; 446.446]$	$[204.190; 994.023]$	$[-1.709; -0.058]$

Table 7: Airport choice data: 95% confidence intervals of WPT for the attributes of the service.

It shows the poor performance of all bootstrap methods making use of non parametric sampling scheme. In fact, when a small sample size is involved, sampling some respondents many times can produce non statistically significant coefficients and unstable estimates, which, in turn, determines large CIs. These methods are, therefore, excluded from the graphical comparison in Figure 6. Pivotal bootstrap methods are also excluded, independently from the sampling scheme used, since they confirm the poor performance emerged in the simulation study. Fieller and LRTI produce, for all the attributes, very similar CIs, which reasonably include 0 only for the WTP of P_AIRL. Percentile methods perform similarly and deliver intervals that are only slightly different from those obtained via Fieller and LRTI. TIB and STIB produce, instead, quite different results in some cases, confirming the doubts already arisen in the simulation study. A final remark concerns the CIs produced by Delta. These are always shifted towards 0 and shorter than Fieller, LRTI and bootstrap methods. The shift also provokes the inclusion of 0 for the WTP of A_MIN and BAL_M_AV attributes, whose coefficients are significantly different from 0. Delta CIs are, thus, less informative, notwithstanding their shorter length. The shift observed could be due to a skewed $\widehat{WTP}$ distribution linked to a small sample size. The skewed $\widehat{WTP}$ distribution of NONSTOP attribute is strengthened by the correlation (approx. equal to 0.025) between F_EURO and NONSTOP estimates having the same sign of WTP.

On the basis of the simulation study, taking LRTI as benchmark, one can evaluate the variations in CIs depending on the method used. The comparison is performed using three indexes: 1) *the good* (G), 2) *the bad* (B), and 3) *the ugly* (U). G index measures the percentage overlapping between the CI produced by the benchmark and the alternative[17]. B index is calculated by adding the absolute value of the difference between the lower bounds to that between the upper bounds of the benchmark and the alternative method. This represents the total over- and under-estimation bias that is, for comparison purposes, normalized using the length of the benchmark CI[18]. U index is binary. It is equal to 1 when the 0 value is

[17]For example, for the NON-STOP attribute one notices that the benchmark CI length is equal to 970 = 1202 (upper bound) - 232 (lower bound); while the corresponding result when using Delta is 642 = 762 (upper bound) - 120 (lower bound). The absolute overlapping is 530 = 762 - 232 which represents 55% overlapping between the two CIs.

[18]For example, for the NON-STOP attribute and for Delta, the difference between the two lower bounds is

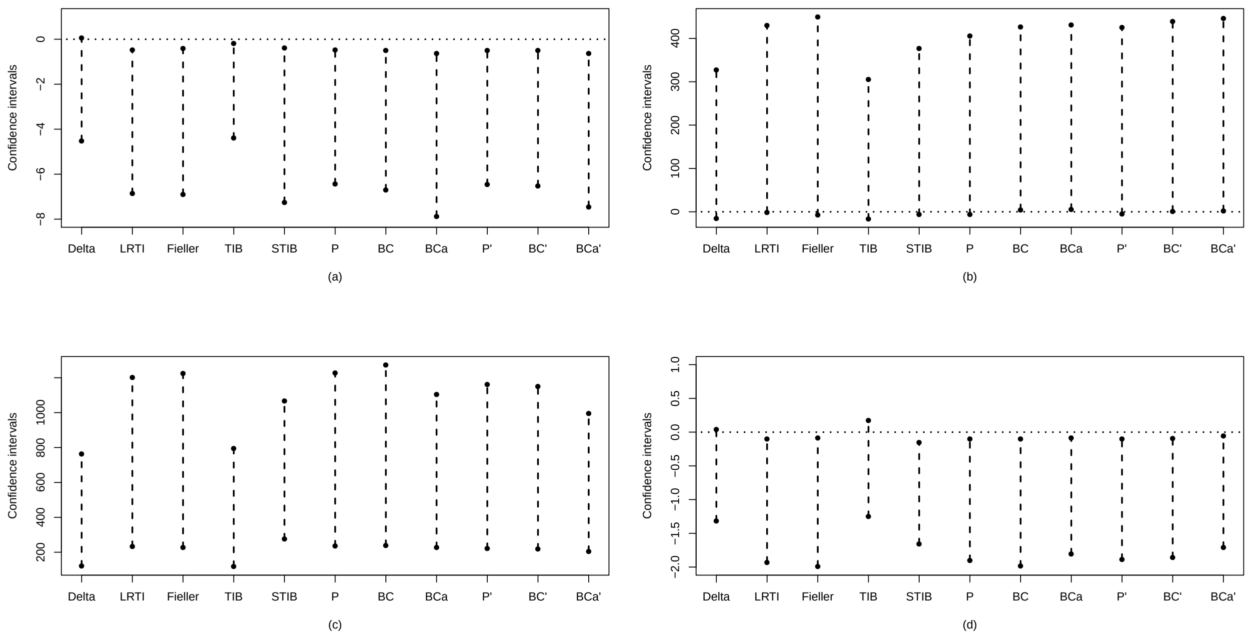

Figure 6: Airport choice data. Confidence intervals for WTP for the attributes (a) access time, (b) preferred airline company, (c) non-stop flight, (d) departure time. Percentile bootstrap intervals are obtained through parametric or Krinsky and Robb sampling (denoted with a prime). Dotted line corresponds to WTP=0.

included in only one of the two CIs (benchmark and alternative method). It is equal to 0 if the 0 value either falls within the two CIs or in none of them. The joint consideration of the three indexes is needed for a performance evaluation of the methods considered.

The indexes calculated confirm the intuitions derived from simulations. Considering the results for the three indexes, averaged over the four attributes, Fieller is the best performer (G index=100%, B index=4%, U index=0), while also percentile bootstrap methods perform well (G index=97%, B index=6%, U index=0.17). STIB (G index=88%, B index=14%, U index=0) performs on average better than TIB, which produces unsatisfactory results (G

equal to 112 = 232 (benchmark) - 120 (Delta), while the difference between the two upper bounds is equal to 440 = 1202 (benchmark) - 762 (Delta). The sum is equal to 552 and the B index is 57% = 552 / 970.

index=63%, B index=45%, U index=0.25). Finally, Delta, given the data characteristics, performs poorly (G index=65%, B index=42%, U index=0.5). As an aside, please note that the correlation between NONSTOP and F_EURO provokes, on average, a lower G index and a higher B index with respect to the other attributes where this correlation is not relevant.

Attribute	Estimate	Std. Error	t-value	p-value
DELAY	-0.1317	0.0164	-8.0114	0.0000
TRIP LENGTH	-0.0241	0.0035	-6.8991	0.0000
FREQUENCY	0.4015	0.0402	9.9756	0.0000
AVAILABILITY	0.0037	0.0003	11.4848	0.0000
COST	-1.4651	0.0889	-16.4765	0.0000

Table 8: Local public transport data: point estimate of attribute coefficients.

In the second empirical example, with the data set characterized by far less problematic features (e.g. large sample size, no attribute correlated with cost, low coefficient of variation for the cost parameter estimate), all various methods produce very similar CIs for WTP. Table 8 provides attribute coefficient estimates for the local public transport data. All the coefficients are highly significant and have the expected sign. Table 9 reports the upper and lower bounds of the WTP intervals for all attributes. The three indexes indicate a generalized overall good performance for all the methods. In fact, BC, the relative worst performer, is characterized by G index=87%, B index=15%, U index=0, that is comparable to the higher performing methods for the first data set. This suggests that, whenever confronted with potentially problematic data sets, the analyst should carefully consider which method to use when calculating CIs for WTP. Such cautiousness is not needed when using large and well-behaved data sets.

6. Conclusions

This paper compares alternative methods to compute CIs for WTP. Monte Carlo simulations are used to assess the performance of the methods considered under different scenarios. More in detail, the paper investigates: 1) correct model specification with cost coefficient

		DELAY	TRIP LENGTH	FREQUENCY	AVAILABILITY
	Delta	[−0.1140; −0.0658]	[−0.0213; −0.0115]	[0.2118; 0.3363]	[0.0020; 0.0030]
	LRTI	[−0.1161; −0.0670]	[−0.0215; −0.0116]	[0.2157; 0.3412]	[0.0020; 0.0031]
	Fieller	[−0.1154; −0.0668]	[−0.0216; −0.0117]	[0.2150; 0.3406]	[0.0021; 0.0031]
Parametric sampling	TIB	[−0.1155; −0.0638]	[−0.0210; −0.0130]	[0.2038; 0.3349]	[0.0022; 0.0030]
	STIB	[−0.1184; −0.0650]	[−0.0206; −0.0122]	[0.2009; 0.3430]	[0.0021; 0.0032]
	NS	[−0.1111; −0.0623]	[−0.0213; −0.0113]	[0.2062; 0.3321]	[0.0020; 0.0030]
	S	[−0.1137; −0.0661]	[−0.0216; −0.0117]	[0.2153; 0.3396]	[0.0021; 0.0031]
	NT	[−0.1140; −0.0658]	[−0.0213; −0.0116]	[0.2115; 0.3366]	[0.0020; 0.0030]
	P	[−0.1175; −0.0686]	[−0.0216; −0.0116]	[0.2160; 0.3419]	[0.0021; 0.0031]
	BC	[−0.1175; −0.0686]	[−0.0216; −0.0116]	[0.2143; 0.3383]	[0.0021; 0.0031]
	BC_a	[−0.1176; −0.0688]	[−0.0216; −0.0116]	[0.2143; 0.3383]	[0.0021; 0.0031]
Non param. sampling	NS	[−0.1127; −0.0651]	[−0.0214; −0.0113]	[0.2060; 0.3327]	[0.0020; 0.0030]
	S	[−0.1153; −0.0683]	[−0.0217; −0.0115]	[0.2163; 0.3400]	[0.0021; 0.0031]
	NT	[−0.1136; −0.0662]	[−0.0214; −0.0115]	[0.2114; 0.3367]	[0.0020; 0.0030]
	P	[−0.1147; −0.0671]	[−0.0215; −0.0115]	[0.2154; 0.3421]	[0.0021; 0.0031]
	BC	[−0.1139; −0.0664]	[−0.0216; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
	BC_a	[−0.1140; −0.0665]	[−0.0215; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
Krinsky and Robb sampling	NS	[−0.1127; −0.0651]	[−0.0214; −0.0113]	[0.2060; 0.3327]	[0.0020; 0.0030]
	S	[−0.1153; −0.0683]	[−0.0217; −0.0115]	[0.2163; 0.3400]	[0.0021; 0.0031]
	NT	[−0.1136; −0.0662]	[−0.0214; −0.0115]	[0.2114; 0.3367]	[0.0020; 0.0030]
	P	[−0.1147; −0.0671]	[−0.0215; −0.0115]	[0.2154; 0.3421]	[0.0021; 0.0031]
	BC	[−0.1139; −0.0664]	[−0.0216; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
	BC_a	[−0.1140; −0.0665]	[−0.0215; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]

Table 9: Public transport data: 95% confidence intervals of WPT for the attributes of the service.

approaching 0; 2) correct model specification with correlation between attribute and cost coefficient estimates having the same sign of $\widehat{WTP}$ and 3) incorrect model specification due to neglected heteroscedasticity. The main findings are summarized below.

1. Most of the scenarios considered reveal some skewness in $\widehat{WTP}$ distribution which should result in asymmetric CIs, especially for small sample sizes. Delta and NT produce, by construction, symmetric CIs thus failing to account for skewness. This translates in a WTP undervaluation. In fact, as suggested by Armstrong et al. (2001), $\widehat{WTP}$ distribution is generally positively skewed and thus the CI's mid-point should be greater than WTP point estimate.
2. $\widehat{WTP}$ skewness is particularly relevant in case of correlation between attribute and cost coefficient estimates having the same sign as $\widehat{WTP}$, or for values of the cost parameter approaching 0. This phenomenon decreases as the sample size increases, so that using symmetric CIs becomes less problematic if the sample is of a reasonable size. Bolduc et al. (2010) underline that very large sample sizes are needed to compensate for cost

parameter estimates approaching 0.

3. Bootstrap methods belonging to the pivotal family are not too accurate, often producing, CIs with poor coverage rates in comparison to nominal levels. Additionally, NT sometimes produces large CIs since it relies on a bootstrap sample estimate of $\widehat{WTP}$ standard error, whose adoption might be very misleading (Daly et al., 2012a).
4. Percentile bootstrap methods prove more accurate and generally perform well. In particular, BC and BC_a relax some assumptions of P and seem more reliable given their ability to account for asymmetric $\widehat{WTP}$ distributions. However, percentile bootstrap methods are characterized by coverage rates significantly different from what expected when the cost parameter approaches 0. Nevertheless, as also shown by Bolduc et al. (2010), smaller cost parameter estimates are necessary to make percentile bootstrap methods unreliable than those sufficient to undermine Delta ones.
5. Bootstrap methods belonging to the test inversion family require careful convergence monitoring, which is not easy to guarantee in a simulation context, thus explaining their sometimes unsatisfactory performance. Nevertheless, positive results encourage future research aimed at determining an appropriate stopping rule.
6. Parametric, non-parametric and Krinsky and Robb resampling schemes do not produce substantially different results. However, non-parametric sampling shows its limits when dealing with small sample sizes and, in general, produces slightly larger CIs, due to the efficiency loss ascribable to repeated sampling of the same individuals. When using bootstrap methods, the smaller the sample size the wiser it is to resample parametrically.
7. Approximation methods belonging to the test inversion family have good performances, are robust to cost parameter approaching 0, simple to calculate and not particularly time-consuming. Monte Carlo simulations confirm the intuition in Armstrong et al. (2001) concerning the inclusion of LRTI CIs in Fieller ones. In fact, for $N = 10$, depending on the scenario considered, this happens between 33% and 74% of the times. These percentages shrink as N increases and, for all scenarios, they get close to 25% for

$N = 50$ suggesting an asymptotic convergence. LRTI thus seems preferable to Fieller, at least when small sample sizes are involved. Additionally, LRTI performs slightly better than Fieller when the cost parameter approaches 0, since its LRP moves faster towards its nominal value as N rises.

In summary, the simulation study suggests using LRTI since it: 1) produces not necessarily symmetric CIs; 2) is not affected by cost parameter close to 0; 3) provides good coverage rates with a correctly specified model; 4) is robust to small departures from correct specification. Fieller represents a valid alternative but when sample sizes are small, might render larger CIs. One could use percentile bootstrap methods, given a cost parameter not too close to 0, since they produce the entire simulated $\widehat{WTP}$ distribution as a byproduct. This might be of interest for policy evaluation, notwithstanding the higher computational time required. On the other hand, Delta, despite its simplicity, rapidity of calculation and diffusion in the literature, proved very sensitive to any departure from normality, either due to skewness or kurtosis. Moreover, due to symmetry, Delta produces CIs systematically shifted towards 0.

The conclusions drawn in the simulation study are pertinent to the real applications investigated. In the first data set, characterized by a small sample size, Delta confirms its limits, while LRTI, Fieller and, to a less extent, percentile bootstrap methods produce similar results. In the second, less problematic data set, all the methods produce fairly similar CIs. In this case, the choice of the method has no substantial implications.

To conclude, notice that some of the considerations emerged from the simulation study retain their validity when building CIs for mixed logit models, in particular the possible shortcomings of Delta. These are evident in the real data comparison between Delta and P, based on Krinsky and Robb resampling, illustrated in Bliemer and Rose (2013), in the context of mixed logit model. They fit seven different models to the same data assuming the following combinations of distributions for the two attribute parameters intervening in the calculation of WTP: 1) fixed divided by fixed (i.e., an MNL model), 2) normal divided by fixed, 3) normal divided by normal (independent), 4) normal divided by normal (dependent), 5) fixed divided by lognormal, 6) triangular divided by fixed, and 7) normal distribution estimated in WTP space. While the two methods provide very similar CIs when the denominator is fixed, results change considerably in the other cases. Having the cost parameter normally distributed is problematic, as a normal distribution has a positive probability mass at zero and Delta performs poorly when β_C approaches 0. As a result, Bliemer and Rose (2013) obtain a suspiciously small Delta CI for the expected WTP, even smaller than the one obtained in case 2), a counterintuitive finding due to the greater uncertainty induced by a random cost coefficient in situations 3) and 4). In these two cases P correctly renders CIs larger than the one resulting from case 2). Delta keeps showing its weakness also in cases 5) and 6), in which the occurrence $\beta_C = 0$ is given a null probability. In case 5), the lognormal probably induces a skewed distribution for the expected WTP estimator, which the symmetric Delta CI cannot capture. This determines a relevant shift of the Delta CI towards 0, compared to the CI calculated through P, and even the inclusion of 0 within the interval. A similar shift can be noticed also in case 6). The two methods produce, instead, close results in WTP space. In summary, the examples in Bliemer and Rose (2013) show evidence in support of P, based on Krinsky and Robb resampling, with respect to Delta, also in the mixed logit framework. It would be interesting to evaluate the extensibility to mixed logit models of Fieller and LRTI, which in our study outperformed all the other methods.

Armstrong, P., Garrido, R., Ortúzar, J. D., 2001. Confidence intervals to bound the value of time. Transportation Research, E 37, 143–161.

Ben-Akiva, M., Lerman, S. R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA.

Bliemer, M. C. J., Rose, J. M., 2013. Confidence intervals of willingness-to-pay for random coefficient logit models. Transportation Research, B 58, 199–214.

Bolduc, D., Khalaf, L., Yélou, C., 2010. Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. Journal of Econometrics 157, 317–327.

Carpenter, J. R., 1999. Test-inversion bootstrap confidence intervals. Journal of the Royal Statistical Society, B 61, 159–172.

Carpenter, J. R., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. Statistics in Medicine 19, 1141–1164.

Daly, A. J., Hess, S., de Jong, G., 2012a. Calculating errors for measures derived from choice modelling estimates. Transportation Research, B 46, 333–341.

Daly, A. J., Hess, S., Train, K. E., 2012b. Assuring finite moments for willingness to pay in random coefficients models. Transportation 39, 19–31.

Davison, A. C., Hinkley, D. V., 1997. Bootstrap Methods and their Application. Cambridge University Press: Cambridge.

DiCiccio, T. J., Efron, B., 1996. Bootstrap confidence intervals. Statistical Science 11, 189–212.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

Efron, B., 1987. Better bootstrap confidence intervals. Journal of the American Statistical Association 82, 171–185.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Chapman and Hall: London.

Fieller, E. C., 1932. The distribution of the index in a normal bivariate population. Biometrika 24, 428–440.

Fieller, E. C., 1940. The biological standardization of insulin. Supplement to the Journal of the Royal Statistical Society 7, 1–64.

Fieller, E. C., 1954. Some problems in interval estimation. Journal of the Royal Statistical Society, B 16, 175–185.

Finney, D. J., 1971. Probit Analysis. Cambridge University Press: London.

Garrido, R. A., Ortúzar, J. d. D., 1993. The chilean value of time study: methodological developments. In: Proceedings of the 21st PTRC Summer Annual Meeting. van Montfort, K. and Oud, H. and Satorra, A.

Garthwaite, P. H., Buckland, S. T., 1992. Generating monte carlo confidence intervals by the robbins-monro process. Journal of the Royal Statistical Society, C 41, 159–171.

Gatta, V., Marcucci, E., 2007. Quality and public transport service contracts. European Transport Trasporti Europei 36, 92–106.

Guria, J., Leung, J., Jones-Lee, M., Loomes, G., 2005. The willingness to accept value of statistical life relative to the willingness to pay value: Evidence and policy implications. Environmental & Resource Economics 32, 113–127.

Hall, P., 1988. Theoretical comparison of bootstrap confidence intervals (with discussion). Annals of Statistics 16, 927–985.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer-Verlag: London.

Hensher, D. A., 2010. Hypothetical bias, choice experiments and willingness to pay. Transportation Research, B 44, 735–752.

Hensher, D. A., Greene, W. H., 2003. The mixed logit model: the state of the practice. Transportation 30, 133–176.

Hinkley, D. V., 1969. On the ratio of two correlated normal variables. Biometrika 56, 635–639.

Hirschberg, J., Lye, J., 2010. A geometric comparison of the delta and fieller confidence intervals. The American Statistician 64, 234–241.

Hole, A. R., 2007. A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Economics 16, 827–840.

Iraguen, P., Ortúzar, J. D., 2004. Willingness-to-pay for reducing fatal accident risk in urban areas: an internet-based web page stated preference survey. Accident Analysis and Prevention 36, 513–524.

Jara-Diaz, S. R., Martinez, F. J., 1999. On the specification of indirect utility and willingness to pay for discrete residential location models. Journal of Regional Science 39, 675–688.

Kabaila, P., 1993. Some properties of profile bootstrap confidence intervals. Australian Journal of Statistics 35, 205–214.

Krinsky, I., Robb, A. L., 1986. On approximating the statistical properties of elasticities. The Review of Economics and Statistics 68, 715–719.

Krinsky, I., Robb, A. L., 1990. On approximating the statistical properties of elasticities: A correction. The Review of Economics and Statistics 72, 189–190.

Li, Z., Hensher, D., Rose, J. M., 2010. Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. Transportation Research, E 46, 384–403.

Marcucci, E., Gatta, V., 2012. Dissecting preference heterogeneity in consumer stated choices. Transportation Research, E 48, 331–339.

Marsaglia, G., 2006. Ratios of normal variables. Journal of Statistical Software 16, 1–10.

Molin, E. J. E., Timmermans, H. J. P., 2006. Traveler expectations and willingness-to-pay for web-enabled public transport information services. Transportation Research, C 14, 57–67.

Ortúzar, J. D., Cifuentes, L. A., Williams, H. C. W. L., 2000. Application of willingness-to-pay methods to value transport externalities in less developed countries. Environmental and Planning, A 32, 2007–2018.