The Australian Journal of
Agricultural and
Resource Economics

Journal of the Australian
Agricultural and Resource
Economics Society

The Australian Journal of Agricultural and Resource Economics, 55, pp. 429–450

Predicting versus testing: a conditional cross-forecasting accuracy measure for hypothetical bias*

Dmitriy Volinskiy, Wiktor Adamowicz and
Michele Veeman[†]

A measure of hypothetical bias, or the divergence between stated and revealed preferences, based on conditional cross-forecasting accuracy is suggested, based on out-of-sample prediction accuracy when estimates from stated preference data are used in place of those from actual choices, and vice versa. We describe an application of this measure to assess hypothetical bias in the context of an inquiry into people's willingness to pay to avoid canola oil produced from genetically modified plants. The analysis suggests the presence of groupwise hypothetical bias in these choice data.

Key words: conditional logit, hypothetical bias, mixed logit, out-of-sample prediction.

1. Background and objectives

Hypothetical bias (HB), which is manifested as systematic divergence between welfare estimates obtained through stated preference (SP) and revealed preference (RP) choice instruments, has long been a troublesome issue for non-market valuation and welfare analysis that uses SP methods.[1] The absence of economic commitment of research subjects that underlie the use of hypothetical scenarios in stated preference methods to obtain willingness to pay (WTP) estimates for particular goods/attributes leads to the concern that respondents may discount financial consequences of their choices

* The authors gratefully acknowledge funding support by Genome Canada, Genome Alberta, Alberta Agriculture Research Institute and the Alberta Crop Industry Development Fund.

[†] Dmitriy Volinskiy (email: dmitriy.volinskiy@ualberta.ca) is Postdoctoral Fellow, Wiktor Adamowicz and Michele Veeman are professors with the Department of Resource Economics and Environmental Sociology (formerly Rural Economy), University of Alberta, Edmonton, Alberta, Canada.

[1] The term 'hypothetical bias', although common in the literature, is not completely accurate in the sense that a purely hypothetical stated preference choice task provides no incentive for the respondent to answer truthfully and may provide little economic information (Carson and Groves 2007). Carson and Groves (2007) suggest that stated preference methods be assessed in terms of the extent to which the surveys are consequential or incentive compatible. We continue to use the term 'hypothetical bias' to describe the difference between stated and revealed preference, but we are interested in assessing stated preference questions designed to be consequential against revealed preferences.

doi: 10.1111/j.1467-8489.2011.00552.x

or become emotionally invested in the simulated transaction, with consequent HB. The majority position on the bias is that, unless some form of calibration or adjustment is undertaken, such as cheap talk scripts (List 2001), HB probably exists (Murphy *et al.* 2005; Ehmke *et al.* 2008). If it is positive, attribute values inferred from an RP choice experiment will be lower than the respective values elicited in its hypothetical counterpart (Cummings *et al.* 1995; Loomis *et al.* 1997; List and Shogren 1998). HB has been of scholarly interest *per se* and is of concern, naturally, when designing more bias-robust experiments (Johnson 2006) or when combining SP and RP data or transferring SP estimates is considered (Haener *et al.* 2001, von Haefen and Phaneuf 2008).

In an ideal, yet unrealistic, setting, HB is best detected when the good is clearly defined as a unified whole and all subjects queried are identical. However, modern valuation experiments commonly deal with multi-attribute goods and situations where the population of subjects vary in their substitution patterns. Constructing an unambiguous and feasible HB test in these circumstances poses challenges.

We argue that treating HB as a dichotomous occurrence (i.e. either present or not) is not appropriate in certain situations. Building on the out-of-sample prediction approach (Haener *et al.* 2001; Chang *et al.* 2009), we assess HB in terms of what we describe as 'conditional cross-forecasting accuracy'. Specifically, we cast HB in terms of the out-of-sample prediction accuracy when estimates from SP data are used in place of those from actual choices, and vice versa. A test for the existence of HB as a dichotomous event would either lead to rejection of the null hypothesis of no bias or not. In contrast, in our approach, the probability measure of prediction accuracy becomes a continuous metric for the bias.

Advantages of the approach we develop include that the proposed measure of conditional cross-forecasting accuracy is based on out-of-sample information and retains its meaning even if the fitted choice model is misspecified. Further, the measure is model-independent, allowing comparisons between different types of models, and it should be computable, regardless of the nature and complexity of the model. The rationale for the use of the conditional cross-forecasting accuracy measure, together with its mechanics and merits, is explained in detail in Section 2. In Section 3, we provide an example of the application of the conditional cross-forecasting accuracy approach to assess HB in the context of an inquiry into whether or not people's WTP to avoid genetically modified (GM) ingredients is systematically affected by the hypothetical nature of a choice situation. This example uses data from a RP choice experiment in which participants had the opportunity to trade an endowed bottle of GM canola oil for an alternative canola oil and a SP replica of this experiment. Concluding the paper, Section 4 offers a review and interpretation of the empirical results and provides directions for further research regarding refinement of the suggested bias measure.

2. Conditional cross-forecasting accuracy measure for hypothetical bias

Initially, consider an experiment to assess the existence of HB in people's values for a single unit of a good. Let μ^{RP} be the desired welfare measure, which corresponds to an actual transaction in which a person obtains the good. Let μ^{SP} be the same type of value for this good as stated by that person (i.e. the hypothetical value). A statistical test of the presence of HB of unknown sign and magnitude can then be constructed as follows:

$$\mathrm{H}_0: \mu^{SP} - \mu^{RP} = 0 \text{ versus } \mathrm{H_A}: \mu^{SP} - \mu^{RP} \neq 0, \tag{1}$$

where $\mu^{SP} - \mu^{RP}$ is the bias. Many studies dealing with HB (see surveys by Harrison and Rutstrom (2008), Murphy *et al.* (2005), List and Gallet (2001)) use the test in Equation (1) in many of its possible forms to assess the presence of HB in their value estimates.

Modern valuation experiments, however, increasingly deal with multi-attribute goods and services, and socio-economic characteristics of the respondents should be taken into account. In these circumstances, let $\mathbf{x}$ be a set of attributes of the good, let $\mathbf{z}$ be a set of relevant socio-economic characteristics of experiment subjects and let $\boldsymbol{\theta}^{RP}$, $\boldsymbol{\theta}^{SP}$ be model parameters to be estimated. The respective value functions are $\mu(\boldsymbol{\theta}^{RP}, \mathbf{x}, \mathbf{z})$ and $\mu(\boldsymbol{\theta}^{SP}, \mathbf{x}, \mathbf{z})$. The strict equivalent of the test in Equation (1) is given by:

$$\mathrm{H}_0: \mu(\boldsymbol{\theta}^{SP},\mathbf{x},\mathbf{z}) - \mu(\boldsymbol{\theta}^{RP},\mathbf{x},\mathbf{z}) = 0, \forall\, (\mathbf{x},\mathbf{z}) \text{ versus } \mathrm{H}_A: \mu(\boldsymbol{\theta}^{SP},\mathbf{x},\mathbf{z}) - \mu(\boldsymbol{\theta}^{RP},\mathbf{x},\mathbf{z}) \neq 0 \tag{2}$$

for at least one $(\mathbf{x}, \mathbf{z})$ pair.

Now, the HB value is $\mu(\boldsymbol{\theta}^{SP}, \mathbf{x}, \mathbf{z})$–$\mu(\boldsymbol{\theta}^{RP}, \mathbf{x}, \mathbf{z})$, which no longer gives an immediate sense of the bias because it is unclear which $(\mathbf{x}, \mathbf{z})$ should be considered. Indeed, no single $(\mathbf{x}, \mathbf{z})$ pair represents the entire population of people and all combinations of the attributes of the good or service. If H_0 fails to hold for a single pair of the many possible pairs, it seems unreasonable to conclude that HB is generally present. A significant problem in finding a suitable test for the expression given in Equation (2) is that this involves unspecified values of $\mathbf{x}$ and $\mathbf{z}$, while statistical tests are normally formulated using specific functions of the parameters. In certain applications, researchers have used a less stringent alternative hypothesis, e.g. $\mathrm{H_A}$: $\mu(\boldsymbol{\theta}^{\mathrm{SP}}, \mathbf{x}_1, \mathbf{z}_1) - \mu(\boldsymbol{\theta}^{RP}, \mathbf{x}_1, \mathbf{z}_1) \neq 0$ for at least one $(\mathbf{x}_{1,}\ \mathbf{z}_1)$ pair, where $\mathbf{x}_{1,}\ \mathbf{z}_1$ are certain subsets of $\mathbf{x}$, $\mathbf{z}$ (as an example, Brown and Taylor (2000) investigate the role of gender as a possible explanation of HB; Ehmke *et al.* (2008) address geographical variation of the bias). This aspect of the problem of finding a suitable test can be abated in part by taking expectations over all possible values of $\mathbf{z}$, so that the expression for the bias becomes $E_{\mathbf{z}}\left\{\mu(\boldsymbol{\theta}^{SP}, \mathbf{x}, \mathbf{z}) - \mu(\boldsymbol{\theta}^{RP}, \mathbf{x}, \mathbf{z})\right\}$. However, the latter test still involves attributes $\mathbf{x}$, and uncertainties about the distribution of $\mathbf{z}$ may make inference unreliable. Furthermore, commodity attributes are not necessarily ‘vertical’ where more of an attribute is better, as when based on quality features, so that consumers are expected to prefer more to less, and preferences for both horizontal (where preference is a pure matter of taste, like colour) and vertical attributes may differ from person to person. If a mixed model (i.e. a model where parameters $\boldsymbol{\theta}^{RP}$, $\boldsymbol{\theta}^{SP}$ are allowed to be randomly individual-specific) is used to account for such preference heterogeneity, testing becomes further complicated, because the investigator will need to deal with distributions of parameters that, in turn, come from distributions. Consequently, unless only a single point from the (**x**, **z**) space is considered, there does not seem to be any generally applicable approach to take into account all possible values of **x** and **z** when testing for the presence of bias.

An intuitive way to bypass the difficulty of dealing with a multitude of (**x**, **z**) pairs is not to calculate the welfare change measure but, instead, to test for HB by proxy, testing the two sets of model parameters (or any relevant subsets of these) for equality. Thus,

$$\mathrm{H}_0 : \boldsymbol{\theta}^{SP} = \boldsymbol{\theta}^{RP} \text{versus } \mathrm{H}_\mathrm{A} : \boldsymbol{\theta}^{SP} \neq \boldsymbol{\theta}^{RP}. \tag{3}$$

While an expression for HB does not appear in the hypothesis statement of Equation (3), intuition suggests that if there is a structural change between SP and RP, and accordingly $\boldsymbol{\theta}^{RP}$ is different from $\boldsymbol{\theta}^{SP}$ as revealed by a test, then μ^{RP} and μ^{SP} cannot be the same for all (**x**, **z**), indicating HB. It is expected that although this will hold true in many cases, generalisation may be misleading, because whether μ^{RP} differs from μ^{SP} will depend on how choice model parameters enter the welfare calculation formula. As an illustration, consider the issue of scale when using the common linear utility specification: $u_1 = s\boldsymbol{\beta}\mathbf{x}_i$, where $s\boldsymbol{\beta}\mathbf{x}_i$ is the indirect utility function, consisting of $\mathbf{x}_i$, a vector of attributes associated with alternative i, coefficients $\boldsymbol{\beta}$, and an unknown positive scale parameter s, and ε_i. When a single set of data is used to estimate a model, s is confounded with the parameter vector and cannot be identified (Haener *et al*. 2001). This implies that, even though s does not affect the rates of substitution between attributes and has no role in WTP calculation, if s^{RP} from the revealed choice data is different from the hypothetical choice s^{SP}, then a test based on Equation (3) may lead the investigator to wrongfully assume the presence of HB where there is actually none.

The confounding problem serves to illustrate a more general difficulty. In fact, whether or not the confounding takes place can be tested by conducting a likelihood ratio test or applying the Wald test, which has the same asymptotic properties, to the ratios of coefficients. More generally, if the WTP function (or functions in the context of the paper, as we are dealing with partworths) is twice continuously differentiable, does not depend on (**x, z**), has a closed-form solution and coefficient estimators are asymptotically normal, the two sets of parameters can be tested for WTP equality. However, if at least one of the above-mentioned conditions does not hold, then no test appears to be generally applicable and testing, if possible at all, can only be carried out on a case-specific basis.

Another issue with the testing of model parameters is the appropriateness of a selected model. If the selected specification is not correct in terms of describing people's choosing behaviour (something that the investigator ultimately has no way of knowing), then any test involving estimated model parameters will be wrong. If the model fails to explain most of the variation in subjects' choices on at least one of the data sets, testing for coefficient equality will be of little value. Lastly, testing for HB using estimated coefficients does not provide any information about the magnitude or severity of bias.

The existence of HB has typically been viewed as a dichotomous occurrence although some studies consider varying levels of HB (Ehmke *et al.* 2008). We consider an alternative approach to detecting the existence of HB, which uses the out-of-sample predictive ability of the model, as discussed in Haener *et al.* (2001). We extend this type of approach to use cross-prediction between SP and RP to measure HB. Instead of a test, we suggest use of a continuous, probability-based HB metric for the measurement of HB. The rationale for using such a metric is the interchangeability of SP and RP data, should HB be absent. If experiment subjects exhibit no HB and the model selected to describe the subjects' behaviour has some explanatory power, it should not matter which parameter estimates, whether SP or RP, are used to obtain the desired value estimates. Similarly, in the absence of HB, when predicting choices on an out-of-sample basis, the model with parameter estimates obtained on the SP data should demonstrate the same prediction ability for an actual choice holdout sample as the model with the parameters estimated on the RP data set, and vice versa. Let P^A be a choice model estimated on either SP or RP data such that, when the estimates from the model are applied to a choice set with characteristics d, it is capable of producing a prediction $\hat{y}^A = P^A(d)$ for the subject's actual choice y.

Let P^B, $P^B(d) = \hat{y}^B$, be its SP or RP 'counterpart'; specifically, if P^A is estimated on the SP data, then P^B points to the same model estimated on the RP data set, and vice versa. Also, let $H^A = \{\{d_1^A, d_1^A, \ldots, d_h^A\}, \{y_1^A, y_2^A, \ldots, y_h^A\}\}$ be a holdout sample left out when estimating P^A, and let H^B be a holdout sample (of the same structure) for P^B. Finally, let H denote the pooled holdout sample, i.e. $H = H^A \cup H^B$. Then, define conditional cross-forecasting accuracy (CCFA) as the (conditional) probability of P^B matching every correct prediction made by P^A:

$$CCFA = \Pr[\hat{y}^B = y | \hat{y}^A = y], (d, y) \in H. \tag{4}$$

As CCFA is a probability, it is bound between zero and unity. Consequently, if HB is absent, the rate at which P^B would match correct forecasts by P^A is expected to be high, approaching unity in the limit. Alternatively, if HB occurs, this conditional probability should be lower, and the more pronounced is the bias, the lower will be the probability.

Predictions made with discrete choice models are typically expressed as expected choice probabilities, not outcomes. To operationalise the CCFA expression in Equation (4), we consider the limits of these probabilities as the researcher's uncertainty about the outcome of a choice decreases to zero. That is, if the *most likely* choice outcome, according to the prediction, coincides with the actual choice outcome, the forecast is considered to be correct. This safeguards the CCFA metric against the scale effect and, in general, makes it insensitive to effects of any monotone, positive transformation of the utility function used with the choice model. Specifically, if $V_j > V_k$, where V_j and V_k are the values of an indirect utility function evaluated for choice options j and k, the random utility is given by $u = u(V, \varepsilon)$, $E_\varepsilon[u] = V$, and T is any strictly monotone, positive transformation of u, then given that $\mathrm{var}(\varepsilon) = 0$, $T(u_j)$ is greater than $T(u_k)$ with probability one: $\Pr[u_\mathrm{j} > u_k] = \Pr[T(u_j) > T(u_k)] = 1$. The CCFA thus reflects a diversion between people's stated and revealed ordering of choice options rather than addressing SP/RP differences in cardinal utility, which makes the measure applicable to cases where using cardinal utility is not appropriate (Slovic *et al.* 1979).

The CCFA has several desirable properties as a measure of HB. First, it does not depend on the nature and/or complexity of the model and does not even need to rely on conventional utility theory. All that is needed from the model is the ability to forecast. Thus, the CCFA of any number of competing models can be directly compared, notwithstanding the extent to which the models differ from each other. Second, the measure does not rely on the explanatory ability of the model because conditioning on the 'own' forecasting accuracy, i.e. the probability of correctly forecasting choices from the data set the model has been estimated on, omits all incorrect predictions. Consequently, the approach does not rest on how many choices both estimates (i.e. both P^A and P^B) correctly predict. Third, unlike the test statistic approach, the CCFA uses out-of-sample information, which becomes a significant advantage when subjects are heterogeneous and outliers are likely to be present. In addition, the CCFA indicates the degree of divergence between the SP and RP estimates, not just the fact of the divergence. Fourth, as the CCFA is related to the model fit, but not its nature, it cannot be wrong even if the model is misspecified. Last, but not least, the CCFA is a probability; therefore, it can be used in decision analysis and for policy-making purposes as such, i.e. as the probability of any event whose realisation is uncertain.

The point estimate of CCFA is obtained by first counting the number of correctly forecast choices by P^A and then determining the number of those choices that P^B, the 'counterpart', has matched in its forecasts:

$$\hat{C}CFA = \frac{\sum 1[(\hat{y}^A = y) \wedge (\hat{y}^B = y)]}{\sum 1[\hat{y}^A = y]}, \ (d, y) \in H. \qquad (5)$$

For example, let us suppose that the number of correctly forecast choices is 5 for SP choices (i.e. when SP estimates are plugged in the model, five choices from the SP part of the pooled *H* are correctly forecast), three for RP (when RP estimates are used, three choices from the RP part of the pooled *H* are correctly forecast). As P^A points to the 'own' model, the number of correctly forecast choices by P^A is $5 + 3 = 8$, which is the denominator in the CCFA expression. To calculate the numerator, let us additionally assume that, of the five correctly forecast choices from the SP part, three forecasts are matched when RP estimates are plugged in, instead of the 'own' SP estimates. Also assume that of the three correctly forecast choices from the RP part, two forecasts are matched when SP estimates are plugged in, instead of the 'own' RP estimates. As a result, the number of choices that P^B, the 'counterpart', has matched in its forecasts is $3 + 2 = 5$. This is the numerator in the CCFA expression. The value of the CCFA measure is $5/8 = 0.625$. The number in the denominator is the probability of the conditioning event, which is a choice from *H* being correctly forecast by P^A; the number in the numerator is the probability of the joint event occurring when both P^A and P^B yield the same correct forecast.

The value of any point CCFA estimate will depend on which observations from the entire available data go into the estimation subset and which are left out to form the holdout sample. To mitigate the effect of this sample partitioning and to obtain the sampling distribution of the CCFA, we suggest the following resampling-based procedure, of which one replication is illustrated in Figure 1:

1. Split each sample (i.e. the SP and RP choice data) randomly into two parts: an estimation subset and a holdout sample of a predetermined size. Pool the two holdout samples to form holdout sample *H*.

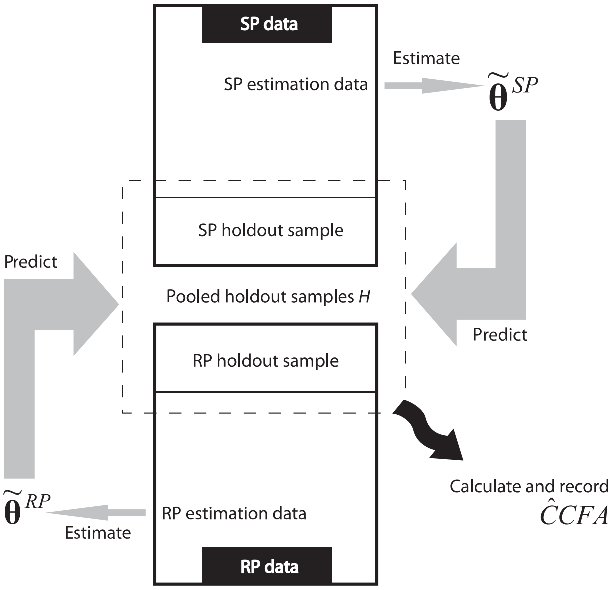

Figure 1 Obtaining empirical distribution of the conditional cross-forecasting accuracy.

2. Estimate the model on each estimation subset and produce forecasts for the data in H.
3. Count all correctly forecast choices for the own subsample and see how many of these were also correctly forecast by the counterpart. Their ratio will be the point estimate of the CCFA. Record the value obtained.
4. Repeat steps 1–3 R times to obtain R values from the empirical distribution of the CCFA.

Selecting the size of the holdout sample involves a trade-off to be made regarding the sampling variation of parameters that shape the distribution of the CCFA. If the size of the holdout sample is set to be relatively large, less information will be available to estimate the model (total sample–holdout sample = estimation sample). This will increase the sampling variation of the model parameter estimates, which will lead to less precise CCFA values. At the same time, increasing the size of the holdout sample will work to increase the precision of CCFA values as more information will be available to calculate the measure. Setting the holdout sample to be small will reverse the action of the two effects.

It may happen that the sizes of the SP and RP data sets are starkly different (e.g. if the SP data come from an Internet-administered survey, while the RP data come from an in-laboratory experiment). In this case, the disparity should be taken into account. One approach would discard a certain number of observational units from the large sample at each replication. Another is to sample with replacement the estimation and holdout sub-samples of a fixed size, using the classical bootstrapping procedure by Efron (1979).

The suggested measure comes with a caveat. If the size of both the estimation and the holdout sample approaches infinity, the moment in the denominator in Equation (5) will converge in probability to a number characterising the predictive ability of the model. The numerator's moment should converge to the same number. This is essentially the same as dividing the data into two parts, estimating the same model on both, and then considering the probability limit of the difference of the two sets of estimates. If the estimator is consistent, then the probability limit will be zero (provided the size of either of the parts is not fixed). This, of course, applies to the estimator of the measure before the bootstrapping procedure enters the picture. Singh (1981) addresses the consistency of some basic random quantities with the classical bootstrap, but his findings (notably Theorem 1, part D) are not applicable in our case, first and foremost because our set-up is not compatible with that used by Singh (ibid.). As we do not formally address large sample properties of the CCFA, its consistency in a general case is unknown.

In the following section, we apply the CCFA measure to assess HB in the context of an inquiry into whether or not people's WTP to avoid GM food ingredients is systematically affected by the hypothetical nature of a choice situation.

3. Testing conditional cross-forecasting accuracy on canola oil labelling data

3.1. Introduction to the data collection instruments and data

Valuations of private goods that have attributes linked to issues of social concern are likely to reflect the views of individuals who exhibit a high degree of emotional involvement and the influence of culture and ethics. For many people, this appears to be the case with GM food products (e.g. Noussair *et al.* 2004; Hu *et al.*, 2004). Much of the commonly consumed food product, canola oil, is derived from herbicide-resistant GM canola plants. However, as the GM-modified protein of modified canola plants is removed from and not discernible in the processed oil, in most countries that require labelling of foods that contain GM ingredients, GM-derived canola generally does not need to be identified by labelling, although this is debated by some people. Emotional involvement and the influence of culture and ethics are also likely to influence many people's views of the country of origin of food; labelling of food origin is another contentious issue. Both attributes are considered in the two data sets used in this application of our proposed method to assess HB.

One set of data used in this study is from an RP choice experiment in which participants were given a series of choice tasks in which they could trade an endowed bottle of GM-derived canola oil that was produced in a specified country for an alternative canola oil, which could possess different attributes. This revealed choice experiment was conducted in Edmonton, Alberta, Canada, in the fall of 2005. Data collection, analysis and WTP estimates from this experiment are reported in Volinskiy *et al.* (2009). With the purpose of assessing the existence and impact of HB that might be associated with the more commonly pursued stated choice approach to valuation, an SP choice replica of this experiment was conducted in summer 2007, also in Edmonton, Alberta. This experiment replicated all details of the RP choice experiment, except that participants made a series of stated choices, rather than actual choices, of their preferred canola oil product, relative to their hypothetical endowment of a bottle of GM-derived canola oil.

Experiment participants were given (actually or hypothetically, as relevant) a 1 L bottle of canola oil and the opportunity to acquire a different type of canola oil through completion of a series of computer-based tasks in which they were shown the labels of different available canola oils. In each task, participants could trade (actually or hypothetically) the endowed oil for an alternative canola oil. In some instances, a premium (implying a cost relative to the endowment) or a discount (implying a refund relative to the endowment) was applied (in the RP experiment) or stated (in the SP experiment) relative to the endowed oil. Respondents always had the option of not choosing the alternative oil. In each choice task, discounts or refunds for the different oil products were specified relative to the endowment, and the choice could be for the endowment or an alternative oil product. As had been indicated to respondents prior to the elicitation process, in the RP experiment, one of the various tasks was selected at random to be 'binding' upon the individual's completion of the task sequence, that is, the type of oil and the price specified in the binding task was the oil product that they actually received and the sum that they actually paid or received. This was not necessary in the SP version of this experiment. The SP experiment participants were repeatedly reminded to make their choices as if choosing between the oils was a real task; as a form of cheap talk script, it was pointed out to the participants that, if each choice task had been an actual exchange transaction, they would have had to pay for/receive a discount for their choice of premium/lower-priced oils.

The on-screen label information that was provided to participants for each product in each of the two experiments indicated '100% Canola Oil', the required nutrition labelling that applies in practice, and the three other items of information that relate to the attributes of the canola oils: (i) country of origin, (ii) type of oil and (iii) price. These features were systematically varied across the various choices. The attributes of the canola oil products and their levels are shown in Table 1.

To enable testing of the potential influence of format effects, a split-sample design was applied in each experiment. Relative to the endowment canola oil product, respondents were randomly assigned either to a format with only one alternative oil in each choice set (*SP1* denotes this format for the SP version of the experiment, and *RP1* denotes this format for the RP version) or to a format in which the various choice tasks included the endowed oil and two other alternatives (*SP2* and *RP2*, respectively). In each of the revealed and stated choice experiments, respondents in the *SP1*/*RP1* group were presented with a sequence of 20 choice sets, while each member of the *SP2*/*RP2* group was presented with 10 choice sets. In each case, the endowed oil was derived from GM ingredients. The endowment (whether real or hypothetical) for some half of the respondents within each format group was a US-made canola oil, while the endowment for the balance of respondents was a Canadian-made GM canola oil product. An effort was made to recruit a

Table 1 Attributes of canola oil products and their levels

Attribute	Levels
Ingredients	'Canola oil from genetically modified canola' 'Canola oil from non-genetically modified canola' 'Canola oil'
Country of origin	'Made in Canada' 'Made in USA'
Price	'Get a refund of 1 dollar with this oil' 'Get a refund of 50 cents with this oil' '$0.00. You already have this oil'* 'Pay 50 cents for this oil' 'Pay 1 dollar for this oil'

*Only endowed oils had this price level.

group of respondents that exhibited socio-economic characteristics that were representative of the population in the first experiment and to recruit a similar group of respondents in the second experiment. A summary of socio-economic characteristics of respondents from the four formats/groups is presented in Table 2.

To see whether there are differences in the distributions of the socio-economic characteristics between SP and RP data, Kolmogorov–Smirnov (K–S) tests are conducted. The distributions of gender are found to be significantly different at 10% for both *SP1*/*RP1* and *SP2*/*RP2* (the K–S statistic values are 0.08 and 0.09, respectively). The distributions of the household income are also significantly different at 10% for the *SP1*/*RP1* groups (the K–S statistic value is 0.07). No distribution differences are found for age and education level.

To keep the present study to the point and within the allowed limits, many details pertaining to the experiment design and pre-testing (focus groups, sampling, test runs) are not included here but are available from the authors upon request.

3.2. Estimation

We model the utility that an individual derives from an oil choice option as a linear function of the attributes of the various canola oils:

$$u_{ijt} = V_{ijt} + \varepsilon_{ijt} = \mathbf{x}_{jt}\boldsymbol{\beta}_i + \varepsilon_{ijt}, \tag{6}$$

where i indexes respondents, j = 0, 1 and t = 1, 2, …, 20 for the formats with one alternative oil in each choice set (*SP1*/*RP1*) and j = 0, 1, 2; t = 1, 2, …, 10 for the *SP2*/*RP2* group. Explanatory variables $\mathbf{x}_{jt}$, which are

Table 2 Respondent characteristics

Characteristic	Format/group	
	SP1	*RP1*
Number of respondents	101	110
% of women	55	47
Median age	43	45
Median education level	College diploma/degree	College diploma/degree
Median household income	C$60–69	C$50–59
	SP2	*RP2*
Number of respondents	100	120
% of women	47	56
Median age	43	46
Median education level	College diploma/degree	College diploma/degree
Median household income	C$60–69	C$60–69

dummy-coded canola oil product attributes and the linear price relative to the endowment (see Table 1), are summarised in Table 3; $\boldsymbol{\beta}_i$ are the associated coefficients, which may be allowed to be individual-specific. All error terms are assumed to be independently and identically distributed standard Gumbel variates.

The first model fitted is the conventional conditional multinomial logit (CL), which restrains preference heterogeneity in the population of subjects by placing a restriction on coefficients to be equal for all subjects, $\boldsymbol{\beta}_{\mathrm{i}} = \boldsymbol{\beta}, \forall \mathrm{i}$. The CL model also maintains the property of independence of irrelevant alternatives (IIA). Estimates from the model are reported in Table 4.

The estimates in the *SP1*/*RP1* and *SP2*/*RP2* pairs are starkly dissimilar, with the consequences of this dissimilarity apparent in terms of the attribute WTP values. The latter are calculated as attribute partworths, i.e. by dividing the estimated attribute coefficients by the negative of the estimated price coefficient. The WTP estimates for the hypothetical choice data tend to be much larger than their RP counterparts, several times larger in certain cases. The indicator of Canada as the country of origin, *CAN*, stands out in this respect. The estimated WTP for *CAN* is eight times larger in the *SP1* format (C$2.69) compared with its *RP1* value of C$0.34 and 2.5 times larger in the *SP2* format: C$1.27 versus C$0.49. The values of the alternative-specific constant (ASC) attribute, which can be interpreted as a preference to keep the endowed oil for its own sake, vary. The WTP estimates of this attribute range from C−$1.23 in *SP1*, indicating a strong tendency to part with the endowment, to C$0.83 (*SP2*), which indicates the opposite tendency. In general, RP estimates almost never exceed one dollar (which was the maximum value of the premium/discount in the experiment), while SP estimates almost always exceed the one dollar value.

In both cases, likelihood ratio tests conducted on the SP and RP sets of parameter estimates lead to the clear rejection of the null hypothesis of coefficient equality (and, accordingly, rejection of the absence of HB), with the relevant *P*-values near zero. In contrast, the average CCFA values presented in Table 5 are quite high, falling in the range of 0.85–0.95. Three selected quantiles from the distribution of the measure, reported in part (b) of the table, also show that the CCFA is generally high: its value only drops below

Table 3 Explanatory variables for indirect utility specification

Mnemonic	Description
ASC	Alternative-specific constant for the endowed oil; 1 if the option is the endowed oil, 0 otherwise
CAN	1 if oil is made in Canada, 0 otherwise
Unspec	1 if GM content is not stated on the label, 0 otherwise
NonGM	1 if the oil is non-GM, 0 otherwise
Price	Premium/refund value, C$. Negative if refund

GM, genetically modified.

Table 4 Parameter estimates from CL model specification

Coefficient	Estimate*	WTP, C$	Coefficient	Estimate	WTP, C$
SP1			*RP1*		
ASC	−0.852 (0.147)	−1.23	*ASC*	0.266 (0.115)†	0.17
CAN	1.852 (0.099)	2.69	*CAN*	0.531 (0.069)	0.34
Unspec	0.907 (0.171)	1.32	*Unspec*	1.108 (0.144)	0.70
NonGM	1.189 (0.174)	1.73	*NonGM*	1.315 (0.146)	0.83
Price	−0.688 (0.084)		*Price*	−1.577 (0.075)	
SBIC at maximum		0.802	SBIC at maximum		1.023
McFadden's R^2		0.26	McFadden's R^2		0.23
SP2			*RP2*		
ASC	0.711 (0.164)	0.83	*ASC*	0.852 (0.147)	0.67
CAN	1.087 (0.094)	1.27	*CAN*	0.622 (0.083)	0.49
Unspec	1.288 (0.158)	1.50	*Unspec*	1.321 (0.149)	1.03
NonGM	1.823 (0.170)	2.12	*NonGM*	1.824 (0.157)	1.42
Price	−0.859 (0.074)		*Price*	−1.281 (0.072)	
SBIC at maximum		1.688	SBIC at maximum		1.649
McFadden's R^2		0.16	McFadden's R^2		0.17

*Standard errors are in parentheses. †P-value of this estimate is 0.02; P-values of all other estimates are <0.01. CL, conditional multinomial logit; SBIC, Schwarz Bayesian information criterion; WTP, willingness to pay.

0.5 in about 5% of cases, while the 95th percentile is universally equal to one (standard errors of the quantiles are calculated using the procedure described by Krinsky and Robb (1991) with 100 draws from the estimated covariance matrix). While this seemingly puzzling outcome might initially be taken as implying failure on the part of the CCFA measure, it actually illustrates a strong point of this measure.

Note that the own-predictive ability of the CL model is extremely low. For *SP1*/*RP1*, the forecasting performance of the CL model is consistently worse than a random guess (which would have given an average value of 0.5 for this choice format, while the maximum rate attained here is 0.48 (*SP1*, observations on five subjects withheld)). Forecasting for the formats with two alternative oils yields averages close to 0.33, which random guess is expected to yield. The predictive ability of the CL model specification is slightly higher when the holdout samples include five subjects (5 × 20 = 100 choice results left out for *SP1*/*RP1* and 5 × 10 = 50 for *SP2*/*RP2*). However, as the holdout sample size increases to 10 and, finally, to 20 subjects, the quality of the prediction is poorer. Although the own-predictive ability of the specification appears to be somewhat better for the RP data, it is much poorer for the SP data.

The CCFA measure only counts cases where the ordering of choice options implied by the expected choice probabilities agrees with the observed choices. Consequently, obtaining such high CCFA values with the CL model means that, while the vast majority of the predictions are wrong, the SP and RP estimates tend to explain the subject's behaviour in the few cases, which they

Table 5 Out-of-sample predictive ability of CL model specification

(a) Summary

Holdout size, # of subjects	Alternative oils, # in choice set	Predictive ability*		
		SP, own	RP, own	CCFA
5	1	0.37	0.48	0.85
5	2	0.31	0.33	0.86
10	1	0.28	0.39	0.95
10	2	0.27	0.34	0.88
20	1	0.28	0.39	0.94
20	2	0.30	0.33	0.86

*Values are averages from $R = 100$ replications. CCFA, conditional cross-forecasting accuracy; CL, conditional multinomial logit; RP, revealed preference; SP, stated preference.

(b) Select quantiles from CCFA distribution

Holdout size, # of subjects	Alternative oils, # in choice set	Select quantiles*		
		5%	50%	95%
5	1	0.51 (0.05)	0.87 (0.06)	1.00 (0.06)
5	2	0.54 (0.08)	0.86 (0.07)	1.00 (0.08)
10	1	0.48 (0.03)	0.92 (0.05)	1.00 (0.02)
10	2	0.56 (0.04)	0.90 (0.03)	1.00 (0.03)
20	1	0.50 (0.03)	0.95 (0.03)	1.00 (0.01)
20	2	0.47 (0.02)	0.88 (0.04)	1.00 (0.01)

*Standard errors are in parentheses. CCFA, conditional cross-forecasting accuracy.

correctly forecast. In other words, WTP values that the CL model predicts are mostly not correct, but when they are correct, there is little systematic divergence between the SP and RP cases, which the testing of coefficients misses completely. To investigate this issue, we re-run the CL specification while recording the identifier of each respondent for which the SP and RP estimates are interchangeable and then calculate the intersection of the obtained sets. We find that about 15% of the combined sample figure prominently at this intersection. This appears to be the small group of people, noted previously, for which the CL model correctly predicts WTP values.

As CL does not seem to be an acceptable model choice for the data, the second model we fit is a mixed logit (ML) model. Unlike CL, the ML model does not impose the taste variation/constant coefficient restriction and is IIA free: individual-specific parameter values are assumed to come from their normal population distribution, $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. A summary of the estimates from the ML model is in Table 6; a complete set of estimates is provided in Table A1 in the Appendix. The ML model was estimated with the use of NLOGIT software, employing quasi-random Halton sequences of 250 points to perform numerical integration.

For the data format with one alternative oil per choice set, the ML estimates show a relatively high extent of dissimilarity between the SP and RP

Table 6 Parameter estimates from ML model specification

Coefficient*	Estimate†	WTP, C$‡	Coefficient	Estimate	WTP, C$
SP1			*RP1*		
ASC	−2.660 (0.728)	−0.54	*ASC*	−0.251 (0.469)§	0.04
CAN	4.461 (0.975)	0.99	*CAN*	3.056 (0.436)	0.28
Unspec	4.966 (1.121)	0.72	*Unspec*	3.189 (0.581)	0.29
NonGM	6.375 (1.271)	1.33	*NonGM*	3.840 (0.649)	0.32
Price	−2.817 (0.793)		*Price*	−6.388 (0.728)	
SBIC at maximum¶		0.543	SBIC at maximum		0.619
McFadden's R^2		0.48	McFadden's R^2		0.46
SP2			*RP2*		
ASC	0.637 (0.505)§	0.35	*ASC*	0.959 (0.340)	0.12
CAN	2.935 (0.443)	0.77	*CAN*	2.267 (0.313)	0.39
Unspec	2.462 (0.429)	0.42	*Unspec*	2.204 (0.453)	0.60
NonGM	3.777 (0.692)	0.69	*NonGM*	3.499 (0.435)	0.62
Price	−2.539 (0.357)		*Price*	−3.571 (0.407)	
SBIC at maximum		1.319	SBIC at maximum		1.216
McFadden's R^2		0.41	McFadden's R^2		0.44

*Only the estimated population means *P* are presented; see Table A1 in Appendix for all estimates. †Standard errors are in parentheses. ‡Median WTP values are reported; see Table A2 in Appendix for more quantiles from the WTP distribution. §*P*-values of these estimates are >0.1; *P*-values of all other estimates are <0.01. ML, mixed logit; SBIC, Schwarz Bayesian information criterion; WTP, willingness to pay.

data, although the discrepancy is no longer as high as with the CL model. There is a large, highly significant negative *ASC* estimate in *SP1*, but the same coefficient estimate is statistically zero in *RP1*. The absolute value of the marginal utility of money (price coefficient) is more than twice larger in *SP1* than in *RP1* (6.388 versus 2.817). Likelihood ratio tests for the equality of SP and RP estimates still result in the rejection of the null hypothesis of no HB, with confidence exceeding 99%, indicating the presence of HB. However, the estimated means of the coefficient distributions do not appear to be dramatically different in the *SP2*/*RP2* pairs. This is reflected in the WTP values. As the ML specification features random parameter values, attribute WTP, in turn, comes from distributions of ratios of normal variables. We obtain quantiles from these distributions based on 1000 draws; median WTP values are reported in Table 6, while more quantile information is presented in Table A2 in the Appendix.

Attribute WTP values in the *SP1*/*RP1* pairs are considerably dissimilar, yet those in the *SP2*/*RP2* pair are similar (except, again, for the value of the 'Made in Canada' attribute, which tends to be significantly higher in hypothetical choices across both choice formats and model specifications).[2] The ML specification, in general, provides a better fit for the data than the CL

[2] Note that this is consistent with Carson and Groves (2007) theoretical prediction that a multiple alternative choice framework in a multi-attribute setting will be incentive compatible, at least in terms of the marginal values, for private or quasi-public goods.

model, as evidenced by lower values of the Schwarz Bayesian information criterion (SBIC). There appears to be a significant amount of random taste variation in the population of canola oil shoppers, because the estimated standard deviations of model parameters are no smaller than the estimated mean values (refer to Table A1 in the Appendix). The better model fit, together with preference heterogeneity being taken into account, leads to changes in the out-of-sample predictive ability, which are seen in the values presented in Table 7. Because of the large amount of computation involved in the estimation of the ML models, only $R = 10$ replications were made to estimate the distribution of the CCFA and the 'own' predictive ability.

We observe a remarkable improvement in the 'own' predictive ability of the ML choice model (and note that Chang *et al.* 2009 also find that the ML tends to outperform the CL in prediction success in most of the cases they examined). The own predictive ability of the CL model was very poor, even falling short of random guess prediction rates. Moving to the ML specification raises prediction rates to range from 0.69 to 0.81% of correct predictions, which corresponds to an improvement in the average predictive ability (over a random guess) of 48% for *SP1*/*RP1* pairs and 101% for the *SP2*/*RP2* pairs. The ML CCFA values are slightly lower than those from the CL model, but still quite high, especially for the *SP2*/*RP2* formats, where CCFA values in the range of 0.73–0.86% correct predictions are more than two

Table 7 Out-of-sample predictive ability of ML model specification

Holdout size, # of subjects	Alternative oils, # in choice set	Predictive ability*		
		SP, own	RP, own	CCFA
5	1	0.81	0.75	0.74
5	2	0.73	0.78	0.86
10	1	0.73	0.74	0.70
10	2	0.71	0.72	0.77
20	1	0.71	0.72	0.69
20	2	0.67	0.69	0.73

*Values are averages from $R = 10$ replications. CCFA, conditional cross-forecasting accuracy. ML, mixed logit; RP, revealed preference; SP, stated preference.

(b) Select quantiles from CCFA distribution

Holdout size, # of subjects	Alternative oils, # in choice set	Select quantiles*		
		5%	50%	95%
5	1	0.42 (0.09)	0.76 (0.08)	1.00 (0.12)
5	2	0.44 (0.11)	0.75 (0.12)	1.00 (0.14)
10	1	0.45 (0.10)	0.75 (0.09)	1.00 (0.11)
10	2	0.43 (0.08)	0.74 (0.08)	1.00 (0.12)
20	1	0.40 (0.06)	0.74 (0.05)	1.00 (0.09)
20	2	0.41 (0.09)	0.72 (0.10)	1.00 (0.07)

*Standard errors are in parentheses. CCFA, conditional cross-forecasting accuracy.

times higher than the random guess reference point of 0.33. The pattern of the CCFA distribution spread remains largely unchanged, compared with the CL: the 5th percentile is at about 0.5, while the 95th percentile is, as before, unity.

We also observe that, as the holdout sample size grows from 5 to 10 individuals (about 5% and 10% of the total sample, respectively), the sampling variation of the estimated CCFA quantiles appears to be decreasing, whereas the average quantile values remain unaffected. This appears to be true with both CL and ML model specifications. The implication is that the holdout sample size of 10 appears to be preferable to 5. Going further from 10 to 20 individuals does not seem to produce the same effect.

4. Discussion and concluding remarks

A question to be answered at this point is: does the evidence indicate that HB is present in the data on canola oil choices? The answer is yes, but it also appears that not every SP experiment participant was prone to this bias. Two principal observations from the study underlie this reasoning. First, the CCFA estimates show that correct predictions from the SP and RP estimates tend to overlap to a significantly high degree, which indicates the existence of a group of subjects who would demonstrate the same choosing behaviour regardless of whether the choice is hypothetical or actual (thus, these individuals exhibit no HB). Second, the CL specification produces dramatically larger attribute WTP values for the SP data than for the RP data but has a very poor out-of-sample predictive ability, while the WTP inferred from the ML estimates for the RP and SP data differ less (the estimated parameter means are notably close in the *SP2*/*RP2* pairs – an outcome predicted by Carson and Groves' (2007) theory surrounding multi-attribute model incentive compatibility), and the model predicts fairly well.

Suppose that subjects' preferences are heterogeneous and that while one group of subjects does not show any effects of HB in their choices, another group is responsible for extreme cases of HB by attaching little (or no) weight to the price attribute. This idea of groupwise HB has been mentioned in literature (Champ *et al.* 1997; Champ and Bishop 2001). In this situation, the constant coefficient CL model is clearly misspecified for the SP data. With the CL model, in addition to ignoring preference heterogeneity, the presence of individuals who exhibit HB, and for whom attribute partworths are sufficiently large to effectively approach infinity, is an issue that is also ignored. This increases the absolute values of estimated attribute coefficients and decreases the estimated price coefficient. In this situation, when the CL model is used to predict hypothetical choices, the predictions are not useful because, *inter alia*, the estimates are invalid for both HB-prone and HB-free groups of subjects. Predicting RP choices is equally problematic, because the WTP values for attributes that are inferred from the SP model are too high for the HB-free RP data. Consequently, the SP-estimated model performs the worst: it accounts for neither taste variation nor the HB-prone/HB-free split. On these types of data, the RP-estimated CL model performs somewhat better (it does not need to address HB), but failing to accommodate preference heterogeneity still fails to explain the majority of choices. In contrast, the ML model specification does accommodate random taste variation and substitution patterns, which dramatically improves the predictive ability of the choice model. While the best model choice for the SP data could possibly be a random effects mixture model (e.g. a latent class model), ML can also accommodate the extreme HB cases, at least in part. However, with the ML specification, groupwise HB will not exclusively affect mean parameter values: the variance of model parameters may partly subsume the HB effect. As a result, the ML model provides much improved prediction quality and less divergence between the WTP values on the data sets in question.

Suspected groupwise HB is not the only feature that characterises divergence between the SP and RP values in the data sets used here. The WTP estimates for Canadian-made canola oils (*CAN* attribute partworth) are consistently higher in the hypothetical choice data set relative to the RP data set. Volinskiy *et al.* (2009) found evidence of extensive choice-variety format effects in the RP data: based on the RP data alone, the attribute WTP in the format with one alternative oil per choice set was significantly different from those with two alternative oil products. The same conclusion evidently holds for the data from the SP experiment.

The experimental nature of the RP data collection, which applied a laboratory simulation of a shop, could lead some level of HB also to be associated with the RP data. Despite the use of focus groups and pretesting, the generated RP data may still have been different from real choices made in markets and may explain why the RP models for the two split-sample experiments are not fully consistent. Considering also the small sample sizes in each experiment, there may be other reasons for the observed differences.

The empirical application of the CCFA given here illustrates both strengths and limitations of this proposed measure of HB. If some people are prone to HB while others are not, the existence of HB in people's choices becomes a relative judgment: how many respondents should be in the HB group for the investigator to conclude that HB is present? The existence of this bias, considered at the population level, is not a dichotomous occurrence, so measuring it requires a continuous, probabilistic metric, which the CCFA conveniently offers. The CCFA is robust by construction to the quality of model fit, demonstrated in the empirical application by the small change in the CCFA estimates for the much less useful CL model versus the superior ML model. Even so, the CCFA may be misleading in the case that it is used to evaluate the model rather than the model's predictive ability. The CCFA takes the model as supplied by the researcher. As a thought experiment, suppose there is no ML and the CL specification is the best the researcher can come up with. The specification produces high CCFA values, so the researcher concludes there is no HB. Although this is based on an incorrect model, the conclusion still holds because, *given the model*, HB cannot be detected insomuch as the model explains the choices. On the other hand, the ML specification is considerably more sensitive; more changes can be detected with the ML, hence the lower CCFA values.

It should be noted that the CCFA measure may not be sufficient in itself to give a nuanced assessment of the nature of hypothetical bias in a particular set of data. In this study, we use the measure in conjunction with the 'own' out-of-sample predictive abilities and coefficient/WTP estimates of two models to arrive at a conclusion with respect to HB in the valuation of attributes of canola oil products. It is possible that without knowledge of how well an estimated model based on SP data actually predicts SP choices, or how well RP estimates predict RP choices, there could be misinterpretations of the CCFA value. For example, if SP estimates happened to correctly predict all choices in the SP holdout sample, and RP estimates predicted all choices in the RP sample, and the correct predictions overlapped only in a single observation, then the estimated CCFA would be unity, which would misrepresent that situation. To avoid such mistakes, we propose estimation of the model's 'own' predictive ability together with the measure of the bias. An alternative would be to calculate the CCFA values for SP and RP separately. For this purpose, the expression for the CCFA (Eqn (4)) may be further decomposed into SP- and RP-specific parts:

$$\begin{aligned}CCFA = \Pr[\hat{y}^B = y|\hat{y}^A = y, A = SP] \times \Pr[A = SP] + \Pr[\hat{y}^B = y|\hat{y}^A \\ = y, A = RP] \times \Pr[A = RP],\end{aligned}$$

and the two conditional probabilities can then be used as two CCFA values for SP and RP, respectively. Finally, the CCFA approach does not provide a 'convenient' measure that indicates the size of the SP welfare measure relative to the size of the RP welfare measure (e.g. WTP 1.5 times as much), but as we have argued earlier, this is difficult to do in a multi-attribute, heterogeneous sample case.

The CCFA measure can readily be used with ranking data and, by virtue of being a measure of forecasting accuracy, the measure can also be applied outside of the HB realm to compare the predictive ability of two competing models on a single discrete choice data set. Ways to refine the suggested conditional cross-forecasting accuracy measure of hypothetical bias are open for further discussion. We believe that this issue makes an interesting research agenda for those interested in empirical assessment of stated versus revealed preference data.

References

Brown, K. and Taylor, L. (2000). Do as you say, say as you do: evidence on gender differences in actual and stated contributions to public goods, *Journal of Economic Behavior and Organization* 43(1), 127–139.

Carson, R. and Groves, T. (2007). Incentive and informational properties of preference questions, *Environmental and Resource Economics* 37, 181–210.

Champ, P. and Bishop, R. (2001). Donation payment mechanisms and contingent valuation: an empirical study of hypothetical bias, *Environmental and Resource Economics* 19, 383–402.

Champ, P., Bishop, R., Brown, T. and McCollum, D. (1997). Using donation mechanisms to value nonuse benefits from public goods, *Journal of Environmental Economics and Management* 33, 151–162.

Chang, J., Lusk, J. and Norwood, B. (2009). How closely do hypothetical surveys and laboratory experiments predict field behavior? *American Journal of Agricultural Economics* 91(2), 518–534.

Cummings, R., Harrison, G. and Rutstrom, E.E. (1995). Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive-compatible? *American Economic Review* 85(1), 260–266.

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

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

von Haefen, R. and Phaneuf, D. (2008). Identifying demand parameters in the presence of unobservables: a combined revealed and stated preference approach, *Journal of Environmental Economics and Management* 56, 19–32.

Haener, M., Boxall, P. and Adamowicz, W. (2001). Modeling recreation site choice: do hypothetical choices reflect actual behavior? *American Journal of Agricultural Economics* 83(3), 629–642.

Harrison, G. and Rutstrom, E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods, in Plott, C. and Smith, V. (eds), *Handbook of Experimental Economics Results*. Elsevier Press, New York, pp. 752–767.

Hu, W., Veeman, M. and Adamowicz, W. (2004). Labelling genetically modified food: Heterogeneous consumer preferences and the value of information *Canadian Journal of Agricultural Economics* 52(3), 79–99.

Johnson, R. (2006). Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum, *Journal of Environmental Economics and Management* 52(1), 469–481.

Krinsky, I. and Robb, A. (1991). Three methods for calculating the statistical properties of estimators: a comparison, *Empirical Economics* 16(2), 199–209.

List, J. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportcards, *American Economic Review* 91(5), 1498–1507.

List, J. and Gallet, C. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? Evidence from a meta-analysis, *Environmental and Resource Economics* 20(3), 241–254.

List, J. and Shogren, J.F. (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment, *Journal of Economic Behavior and Organization* 37(2), 193–205.

Loomis, J., Brown, T., Lucero, B. and Peterson, G. (1997). Evaluating the validity of the dichotomous choice question format in contingent valuation, *Environmental and Resource Economics* 10(2), 109–123.

Murphy, J., Allen, P., Stevens, T. and Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation, *Environmental and Resource Economics* 30(3), 313–325.

Noussair, C., Robin, S. and Ruffieux, B. (2004). Do consumers really refuse to buy genetically modified food? *The Economic Journal* 114(January), 102–120.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Annals of Statistics* 9(6), 1187–1195.

Slovic, P., Lichtenstein, S. and Fischhoff, B. (1979). Images of disaster: perception and acceptance of risks from nuclear power, in Goodman, G. and Rowe, W. (eds), *Energy Risk Management*. Academic Press, London, pp. 223–245.

Volinskiy, D., Adamowicz, W.L., Veeman, M. and Srivastava, L. (2009). Does choice context affect the results from incentive-compatible experiments? The case of non-GM and country-of-origin premia in canola oil, *Canadian Journal of Agricultural Economics* 57(2), 205–221.

Appendices

Parameter and WTP estimates from ML model specification

Table A1 Parameter estimates

Coefficient	Estimates (SE)			
	SP1	*RP1*	*SP2*	*RP2*
	Means ***β***			
ASC	−2.660 (0.728)***	−0.251 (0.469)	0.637 (0.505)	0.959 (0.340)***
CAN	4.461 (0.975)***	3.056 (0.436)***	2.935 (0.443)***	2.267 (0.313)***
Unspec	4.966 (1.121)***	3.189 (0.581)***	2.462 (0.429)***	2.204 (0.453)***
NonGM	6.375 (1.271)***	3.840 (0.649)***	3.777 (0.692)***	3.499 (0.435)***
Price	−2.817 (0.793)***	−6.388 (0.728)***	−2.539 (0.357)***	−3.571 (0.407)***
	Standard deviations $\sqrt{diag(\sum)}$			
ASC	4.371 (1.226)***	2.361 (0.761)***	2.509 (0.736)***	1.898 (0.419)***
CAN	4.209 (0.957)***	3.294 (0.488)***	2.054 (0.391)***	2.211 (0.341)***
Unspec	7.750 (1.510)***	4.319 (0.576)***	2.984 (0.548)***	3.277 (0.585)***
NonGM	7.243 (1.285)***	5.672 (0.497)***	3.844 (0.641)***	3.766 (0.596)***
Price	3.173 (0.702)***	6.152 (0.439)***	2.338 (0.403)***	3.038 (0.372)***
	Diagonal values in Cholesky factor **L**, **LL** = E			
ASC (A)	4.371 (1.226)***	2.361 (0.761)***	2.509 (0.736)***	1.898 (0.419)***
CAN (C)	4.198 (0.923)***	3.273 (0.505)***	1.988 (0.361)***	1.714 (0.320)***
Unspec (U)	2.828 (0.687)***	4.191 (0.604)***	2.156 (0.420)***	2.877 (0.587)***
NonGM (N)	1.748 (0.437)***	3.686 (0.621)***	2.144 (0.454)***	2.212 (0.378)***
Price (P)	1.563 (0.867)*	4.667 (0.642)***	1.803 (0.346)***	2.189 (0.325)***
	Below-diagonal values in Cholesky factor **L**			
C – A	0.299 (1.026)	0.370 (0.536)	0.516 (0.586)	1.396 (0.417)***
U – A	4.648 (1.506)***	−0.322 (0.827)	−2.012 (0.618)***	0.456 (0.641)
U – C	5.520 (1.145)***	−0.993 (0.656)	0.460 (0.420)	−1.501 (0.475)***
N – A	2.496 (1.361)*	−0.594 (0.764)	−2.934 (0.764)***	1.629 (0.612)***
N – C	6.567 (1.148)***	0.376 (0.532)	0.814 (0.646)	−1.738 (0.438)***
N – U	−0.218 (0.644)	4.254 (0.616)***	0.952 (0.746)	−1.901 (0.772)**
P – A	0.525 (0.786)	−2.389 (0.518)***	−1.101 (0.452)**	1.201 (0.511)**
P – C	−0.515 (0.652)	2.710 (0.819)***	0.159 (0.395)	−0.217 (0.446)
P – U	2.397 (0.565)***	1.554 (0.503)***	0.365 (0.473)	−1.555 (0.498)***
P – N	1.158 (0.811)	−0.776 (0.405)*	0.920 (0.432)**	−0.728 (0.490)

***Significant at the 1% level; **significant at the 5% level; *significant at the 10% level.

Table A2 Quantiles from attribute WTP distributions

Coefficient	WTP, C$									
	SPI					*RPI*				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
ASC	−6.74	−1.58	−0.54	0.43	4.69	−1.42	−0.18	0.04	0.25	1.03
CAN	−6.51	−0.13	0.99	2.29	7.34	−2.85	−0.06	0.28	0.77	3.43
Unspec	−13.00	−1.16	0.72	3.00	14.35	−2.34	−0.11	0.29	0.83	3.66
NonGM	−11.57	−0.47	1.33	3.44	13.08	−3.54	−0.23	0.32	1.00	4.14
	SP2					*RP2*				
ASC	−3.43	−0.30	0.35	0.87	3.64	−2.16	−0.22	0.12	0.60	3.27
CAN	−3.69	0.20	0.77	1.58	5.79	−3.20	−0.01	0.39	1.03	4.21
Unspec	−7.11	−0.39	0.42	1.50	6.76	−2.47	−0.01	0.60	1.19	3.43
NonGM	−10.36	−0.31	0.69	2.28	11.21	−6.65	−0.15	0.62	1.69	7.63

WTP, willingness to pay.