Can Calibration Reconcile Stated and Observed Preferences?

Author: Bailey Norwood
426 Agricultural Hall
Stillwater, OK 74078
(405) 744-9820
baileyn@okstate.edu

Final Submission to the Journal of Agricultural and Applied Economics

September 16, 2004

The author is assistant professor, Department of Agricultural Economics, Oklahoma State University. The author wishes to thank two anonymous reviewers for their constructive comments, as well as the students who participated in the study.
Can Calibration Reconcile Stated and Observed Preferences?

While stated preference has become a standard non-market valuation tool, hypothetical bias remains its most noted weakness. When people are asked how much they value a good, they tend to state an amount larger than they are truly willing to pay. In studies comparing hypothetical values from state preference methods to “true” values from experiments involving real money, hypothetical values are almost always larger. List and Gallet report a range of calibration factors (hypothetical values divided by true values) from thirty studies, many of which are greater than three, indicating great care must be taken when predicting true values from stated preference methods.

Hypothetical bias is a phenomena for both private and public goods. The only difference is the factors causing hypothetical bias. Both suffer from the uncertainty people possess over how they would react in real situations. But with public goods, people may also employ a strategy of stating a high value for the public good, hoping it will be provided through donations, after which they can free-ride. This is often called a strategic-bias, but since it creates a wedge between stated values and true willingness-to-pay, this study treats it as just another form of hypothetical bias.

Several methods have been proposed to reduce hypothetical bias. This study focuses on methods that require only data contained in the stated preference experiment. One method is to multiply hypothetical values by a factor less than one, commonly called a calibration factor. The difficult task is identifying the specific calibration factor. For instance, the well-known NOAA panel once recommended that all hypothetical values be reduced by one-half, implying a calibration factor of one-half for all individuals. Recent research has sought less arbitrary means
of determining calibration factors; methods where the calibration factor has been scrutinized in a laboratory. Also, since not all individuals will display the same degree of hypothetical bias, researchers have attempted to formalize mechanisms where the calibration factor differs across individuals.

Two such mechanisms are the certainty-calibration and the frontier-calibration. Both assign individuals different calibration factors based on their response to stated preference surveys, and both have successfully survived controlled experiments measuring their ability to predict true values. However, there are several reasons why the certainty- and frontier-calibration methods are not yet considered standard tools. The first is that they simply have not passed enough laboratory tests. The certainty calibration has been tested just twice and the frontier calibration only once. Second, even if both methods are useful in predicting true values, no study has yet identified which method is better by comparing the two in the same experiment. Third, while the certainty calibration has been tested in a dichotomous choice setting and the frontier calibration has been tested using auction data, no study has tested whether they can be extended to conjoint analyses, a standard non-market valuation tool.

This study contributes to the literature by addressing each three concerns. Data taken from a classroom experiment is used to predict observed behavior from stated preference surveys. Students were granted attendance points that contributed towards their final grade. They were then given the opportunity to donate points towards a public good that would enhance their grade further. This public good was designed to mimic a voluntary marketing checkoff. While the game rewarded cooperation by the entire class, it especially rewarded individuals when they were among the few to free-ride. Before given the opportunity to donate to the public good, conjoint surveys were administered to elicit student preferences on how the public good
should be designed. Indeed, students exhibited a hypothetical bias—they were not as enthusiastic to donate towards the public good as they stated in surveys. In other words, the utility of the public good estimated from stated preference surveys was higher than the utility revealed by the students’ choices.

To correct for this hypothetical bias, preferences for the public good were calibrated according to the certainty- and the frontier-calibration, producing two calibrated utility functions. The ability of each calibrated utility function to predict actual donations was then tested. In terms of prediction accuracy, the two calibrations are comparable, but they did not improve upon the uncalibrated utility function. However, results of this study and previous studies suggest that calibrated and uncalibrated utility can provide an upper and lower bound to true values.

The next two sections provide a brief description of the two calibrations, and is followed by a description of the experiment used to test calibration performance. The fifth section describes the estimation procedure and the sixth section compares the forecasting ability of calibrated and uncalibrated models. The last section provides concluding comments.

**The Certainty-Calibration**

The certainty-calibration is increasingly used in non-market valuation studies, including Blumenschein et al (1998; 2001), Champ and Bishop, Poe et al. and Vossler et al. This paper utilizes the certainty-calibration as described by Champ and Bishop, Poe et al. and Vossler et al. Mail surveys in the Champ and Bishop study were issued asking respondents if they would be willing to donate a particular amount of money to a public good, assuming hypothetically they
were given the opportunity. Following the hypothetical question, if the person indicated “yes” she would donate, she was presented with a certainty question stated “On a scale of 1 to 10, where 1 means very uncertain and 10 means very certain, how certain are you that you would purchase the wind power offered in Question 1 if you had the opportunity to actually purchase it?”

Another sample of respondents received a similar survey, but were actually given an opportunity to donate real money towards the public good. Not surprisingly, the percentage of “yes” responses was higher to the hypothetical donation than the real donation opportunity. The authors then demonstrate that, for individuals who answer “yes” to the hypothetical donation but reveal a certainty less than eight, if one recodes their answers to “no” the hypothetical bias disappears.

Please note that the certainty-calibration in this paper is always assumed to employ a threshold of eight. This threshold must be set in advance, as there will always be some threshold that equates hypothetical and observed values. For the certainty-calibration to be scientifically valid, there must be one threshold that works well across all data. Empirical tests of the eight-threshold are encouraging but not complete. The eight-threshold was first calculated to be optimal in Champ and Bishop. The Vossler et al. study also found the eight-threshold unbiased, but the Poe et al. study found it underestimated true values. However, Poe et al. concluded a seven-threshold to be optimal, which close to eight. This suggests that while the eight-threshold is not perfect it may still be the most reliable. The present study provides an additional test of the certainty-calibration using the eight-threshold.
It should be noted that other versions of the certainty-calibration exist which are not analyzed in this study. Johanneson et al. use a certainty scale of zero to ten (as opposed to one to ten) and estimated an optimal threshold using a probit model. Blumenschein et al. (1998) use categories of certainty such as “definitely sure” and “probably sure” instead of a numerical scale. Time only allowed testing one version of the certainty-calibration, so only the most widely cited version is used.

The Frontier Calibration

An alternative calibration technique was recently offered by Hofler and List. This calibration was designed for auction data. The authors first elicited bids in a hypothetical Vickrey auction where no money or goods were exchanged. Let the hypothetical bid for individual $i$ be denoted $Y_i^H$. After individuals submitted hypothetical bids, they were then invited to submit real bids, where the winner would have to pay real money in exchange for the good. Label these real bids as $Y_i^A$. Hofler and List then estimated the following stochastic frontier function

$$Y_i^H = X_i \beta + v_i + \mu_i$$

where $v_i$ is normally distributed with a zero mean and $\mu_i$ is a non-negative random error. The authors then assumed that $Y_i^A = X_i \beta + v_i$, making $\mu_i$ identical to the hypothetical bias. The assumption that $\mu_i$ is the hypothetical bias proved useful. After estimating real bids as

$$\hat{Y}_i^A = Y_i^H \left[ \frac{X_i \beta}{X_i \beta + \mu_i} \right],$$

which is akin to assuming $\mu_i = 0$, the predicted bids $\hat{Y}_i^A$ from the hypothetical auction and the bids from the real auction were statistically indistinguishable.
The underlying assumption behind the frontier calibration is that in real auctions, people with similar demographics described by $X_i$ will submit similar bids. These “similar” bids are described statistically as $X_i\beta + v_i$. In hypothetical auctions, some of these people will submit higher bids than those in their demographic, but this is solely due to the hypothetical nature of the auction, and when the auction is made real their bids will fall back to the process $X_i\beta + v_i$. Another view of the frontier calibration interprets $X_i\beta + v_i$ as the lower bound on bids. Thus, the frontier calibration also works by using the lowest hypothetical bids as the prediction of true bids for each demographic.

Mimicking a Voluntary Beef Checkoff

The potential for both the certainty- and the frontier-calibration method to remove hypothetical bias has been demonstrated in previous studies. However, more experiments are needed before either should be used widely. Also, the certainty-calibration was used in a dichotomous choice setting while the frontier-calibration was employed in a Vickrey Auction. Conjoint analysis is another popular stated preference approach, so whether the calibrations work well in conjoint analysis warrants attention.

This study uses data from a classroom exercise designed to illustrate the difficulties of providing public goods through voluntary donations. The beef checkoff was used as a case study. The beef checkoff is a program authorized by the Beef Promotion and Research Act of 1986, where U.S. cattle producers are assessed a one dollar fee for each head of cattle that is sold. An equivalent fee is assessed on beef and cattle imports as well. These funds are used to increase beef demand through generic beef advertising, food safety research, consumer education and foreign marketing. Although the ability of the checkoff to increase beef demand is difficult
to quantify (see Coulibaly and Brorsen), the checkoff is supported by a majority of cattle producers (High Plains Journal; Winn et al.).

Despite its popularity, the current checkoff has been challenged numerous times in court. Until recently these challenges were unsuccessful. In 2002, the Livestock Marketing Association argued that using mandatory fees to fund advertisements violates the first amendment, and the courts agreed. An appeal has been made to the Supreme Court, and until this appeal is ruled upon, the future of the beef checkoff remains uncertain.

The U.S. Supreme Court has already ruled the mushroom checkoff unconstitutional on the grounds of free speech, so a similar ruling for the beef checkoff seems likely. In preparation for a ruling against the checkoff, twenty-five states have passed legislation to implement a voluntary beef checkoff, where checkoff fees are collected just as they are now, but producers may receive a full refund if requested. The federation of the state beef councils, which was in place before the mandatory checkoff was created, could then pool all the participating states’ revenues to fund large-scale beef promotion activities (Reese).

A voluntary beef checkoff would be subject to free-riding. Producers may request a refund of their checkoff fees, yet still reap the benefit of higher beef prices due to checkoff activities. This makes checkoff activities a public good for cattle producers. To illustrate the difficulties of providing public goods through donations, an experiment mimicking a voluntary checkoff was conducted in the classroom. Participation points, which enhanced students’ final grade, were used in place of money. These points could then be donated to a public good. Whether these donations benefited the donator depended on the level of free-riding.

The class was a junior-level undergraduate agricultural marketing and price analysis class with 59 students. Students’ grades were partially based on participation. The participation
component was designed to reward attendance without punishing absences. Students were told that attendance would be taken on ten randomly determined days. Students present on those days would receive 100 participation points. They would then be given the opportunity to donate a fixed number of points to a public good, where the donated points would be doubled and evenly distributed among all students, even those who did not donate. The donation of participation points is akin to a donation of money to a voluntary beef checkoff.

This donation could only equal a lump sum. For example, the student could donate 80 points or no points. This mimics the per-cattle-sold fee that would be collected under a voluntary checkoff, which producers could subsequently request to be refunded. Some researchers have suggested a provision points mechanism may help alleviate free-riding in public goods provision (Poe et al.), some mentioning voluntary checkoffs specifically (Messer, Kaiser and Schulze). A provision point mechanism is a designated minimum support level, where if the minimum support for the public good is not met, all contributors to the public good will receive a full refund of their contributions. This limits the amount of free-riding, as free-riding is no longer a dominant strategy (Bagnoli and Lipman). A provision point mechanism was implemented by establishing a minimum participation rate, defined as the percent of students donating. If the minimum participation rate was not met, then all donations to the public good were returned in full.

Students indicated whether they wished to donate by filling out a form administered during class. For example, the form used at the first donation opportunity is shown in Appendix A. Students were told they would receive a participation grade of a 100, but the contribution of this 100 towards their final grade depended on how many participation points they received over the ten trials. If the minimum participation rate was never met, but the student always attended
class, the participation grade of a 100 would count 4% of their final grade. Donating to the checkoff could increase this percentage. If everyone always attended class and donated their 100 points to the checkoff, the participation grade of a 100 would count 8% of their final grade. This percentage could be increased even higher to 10% if they were the only student who did not donate 100 points at each opportunity. Students were then told that their final grade would be based on whether their numerical grade fell within predefined intervals. That is, they were graded on an absolute scale, rather than relative to each other.

To ensure each student understood the game, a spreadsheet was constructed where students could input difference scenarios (e.g. donation amount, minimum participation rates, actual donation rates) and observe the outcome. More than forty-five minutes were devoted towards explaining the experiment and answering questions. The experiment generated great interest and curiosity on the part of the students, and all were made to sign a form indicating they understood the experiment. Students were told not to discuss the experiment with other students at any time, and failure to abide by this request would be considered cheating.

This experiment provides a unique opportunity to conduct a hypothetical survey before actual donations are made. Hypothetical donations can then be calibrated using the certainty- and frontier-calibration and compared to actual donations. A conjoint analysis survey was administered to students before the experiment began. Each student was asked to complete four conjoint questions, an example of which is shown in Appendix B. Each conjoint question contained three options, the first two of which are donations opportunities. The two donation opportunities differ by the number of points one can donate and the minimum participation rate. The third option is to abstain from making a donation. For example, a subject who prefers to donate a small number of points at a high minimum participation rate is likely to select Donation
Opportunity A in Appendix B. A person who does not wish to donate would choose Neither Donation Scenario. If a student chooses one of the donation opportunities, she is asked to complete a certainty question very similar to Champ and Bishop and Vossler et al. The two donation opportunity attributes, number of points and minimum participation rate, varied randomly across and within surveys. Both were chosen from uniform distributions between 5 and 100. This random design ensures that the two attributes are orthogonal, leading to an efficient experimental design.3

The survey yielded 197 choices, which were used to estimate a random utility model for the public good. Utility for donations to the public good is stated to be a function of the donation amount, the minimum participation rate and gender. The utility from not donating was set to zero. Thus, if utility for a particular donation amount and minimum participation rate is greater than zero, the student makes the donation. Otherwise, the donation is not made. In the next section we show that the utility estimated from the choice experiment survey overestimates the number of students that would donate. The ability of the certainty- and frontier-calibration to correct for this overestimation is then evaluated.

Calibrating Utility for the Public Good

Based off student answers to the conjoint survey, the following random utility function was estimated for the public good. The utility from choosing Option A on the choice experiment survey was assumed to follow

\[ U_A = X_A \beta + \varepsilon_A = \beta_0 + \beta_1(DON_A) + \beta_2(DON_A)^2 + \beta_3(MPR_A) + \beta_4(MPR_A)^2 + \beta_5(DON_A*MPR_A) + \beta_6(MALE) + \varepsilon_A. \]
where $DON$ is the fixed amount of money that can be donated, $MPR$ is the minimum participation rate, $MALE$ is a dummy variable for males and $\varepsilon_A$ is a zero-mean random error. The value of $DON$ and $MPR$ was scaled to the $(0,1)$ interval to facilitate convergence in non-linear estimation. The subscript “A” refers to Option A on the survey. The utility from Option B is the same as (3) except with subscript “B” instead of “A”, while utility from Option C (no donation) simply equals $\varepsilon_C$.

The utility function was first estimated using the standard conditional logit model, whose estimates are shown in the second column of Table 1. This is referred to as the uncalibrated conditional logit model. The minimum participation rate had a positive and increasing effect on donations, indicating that it does help control free-riding in public goods provision. The sign on $DON$ may be seem odd, because the price coefficient is usually thought to be negative. However, the donation is not the price of a public good but an investment, so the sign on donation could be positive or negative. Finally, the estimates suggest that males expressed a lower desire to donate than females.

Although many individual coefficients are insignificant, likelihood ratio tests show that all variables are jointly significant. Moreover, when utility is estimated without quadratic or interaction terms all coefficients except the intercept are significant. This indicates the students are indeed responsive to the public good attributes, although the response gets hidden when quadratic and interaction terms are used. In many cases this would lead one to adopt a simple linear model. However, the purpose of this study is to compare calibrations, not functional form. For objectivity, we choose to specify a single, flexible function form 	extit{a priori}, and to hold this functional form constant across calibration.
Next, the certainty-calibration is used to re-estimate utility for the public good, producing a certainty-calibrated conditional logit model. For all individuals in the conjoint survey who responded “yes” they would donate but indicated a certainty level less than eight, their answers were recoded to “no”. The recoded data were then used to estimate a conditional logit model, producing the estimates shown in the third column of Table 1. Not surprisingly, this leads to different preferences for donating to the public good. The two most salient differences are the lower intercept, indicating a lower probability of donating, and a change in the gender coefficient sign. Most females responded they would donate but indicated a low certainty score. Therefore, while females had a higher probability of donating according to the uncalibrated utility function, males had a higher probability of donating according to the certainty-calibrated utility function.

Finally, utility was estimated using the frontier calibration. Let $U_A^H$ be “hypothetical utility”, or the utility that describes individuals’ choices in hypothetical situations, and let $U_A^A$ be utility that best describes individuals choices in real situations where a real good is obtained at a real cost. Similar to Hofler and List, define the relationship between $U_A^H$ and $U_A^A$ as

\begin{equation}
U_A^H = U_A^A + v_A + \mu_A = X_A \beta + v_A + \mu_A
\end{equation}

The difference between (4) and the Hofler and List model is that we assume $v_i$ to follow the extreme-value distribution instead of a normal distribution. The two distributions are similar so this should not affect the results, though it greatly simplifies estimation. The term $\mu_A$ is assumed to equal the hypothetical bias. Recall that the utility of not donating is set to equal zero, so a positive $v_A$ and $\mu_A$ increases the propensity to donate towards Option A. The $\mu_A$ term is not included in the utility from not donating (Option C). The distribution of $\mu_i$ is assumed to follow the non-negative, exponential distribution $\mu_i \sim \alpha e^{-\mu_i}$ for $i = A, B$. With the hypothetical bias,
term $\mu_i$, (4) becomes a mixed logit model where the probability of choosing Option A on the choice experiment survey is

$$
Pr_A = \int_0^\infty \int_0^\infty \left( \frac{\exp(X_A \beta + \mu_A)}{1 + \exp(X_A \beta + \mu_A) + \exp(X_B \beta + \mu_B)} \right) \alpha^2 e^{-\alpha(\mu_A + \mu_B)} d\mu_A d\mu_B.
$$

Expressions for the probability of choosing Options B and C follow similarly. This model is referred to as the frontier-calibrated mixed logit model. The parameter estimates were obtained using simulated maximum likelihood with Halton Sequences as described by Train. The estimates, shown in the last column of Table 1, shows that the estimate of $\alpha$ is 0.0683, which implies the mean of $\mu_i$ is a large $1/0.0683 \approx 15$. Likelihood ratio tests suggest this parameter is insignificant. Nevertheless, it produces a very different profile of student preferences for the public good, as demonstrated in the next section.

**A Forecasting Contest**

Approximately one week after the conjoint survey was administered, students were given their first opportunity to donate towards the public good. The 55 students present that day were given the form in Appendix A describing the opportunity to donate 80 points (out of the 100 points they were given) to the public good. If at least 70% of students did not donate, they were told that their donated points would be refunded. Since students must either donate 80 points or no points, the total value of the public good can be estimated by multiplying 80 points times the actual donation rate. Since more points increase the probability of a higher grade, they can be used as a measure of value, similar to dollars in most valuation studies. Total contributions to the public good depend on the actual donation rate, thus, each model was judged based on its ability to forecast the donation rate.
The actual donation rate was 76%, barely exceeding the minimum participation rate. Of the 55 students present, 36 were males. Based on these demographics and using the uncalibrated model one would predict 92% of males and 98% of females to donate the 80 points (See Table 2), yielding a total donation rate of 94%. This probability is calculated by the formula

$\text{Probability of Donating} = \frac{\exp(X_{\text{OUT}} \beta)}{1 + \exp(X_{\text{OUT}} \beta)}$

where $X_{\text{OUT}}$ contains data for $X$ in (3) where $DON = 0.8$ and $MPR = 0.7$. There are two possible reasons for this hypothetical bias. With private goods, individuals who express a high level of uncertainty using the 1-10 certainty scale also tend to display greater hypothetical bias (Johannesson et al.). This experiment employed a public good, and students may also signal a high participation rate hoping the instructor would reveal this high rate. From a student’s point-of-view, this high rate may induce greater donations by others, enabling her to free-ride with greater reward. This is a strategic-bias. The objective of this study is to evaluate the ability of the certainty- and frontier-calibration to account for uncertainty and strategic-bias, improving estimates of actual donation rates.

Using the certainty-calibrated model, we would predict 40% of students to donate, well below the actual rate of 76%. The donation rate using the frontier-calibrated model is estimated by assuming that $\mu_i$ equals the hypothetical bias. To predict actual donations the term $\mu_i$ is removed from utility, allowing the probability of donating to be calculated using (6). The predicted participation rate is 39%, which is very similar to the certainty-calibration. The two calibrations differ greatly in the donation rates for males and females, but when gender types are combined the rates are almost identical. Both calibration methods produce similar results, and both underestimate the value students’ willingness-to-donate. Moreover, the difference in the
donation rates for the calibrated and uncalibrated models is statistically different from the true
donation rate of 76%. No model provided unbiased estimates of the donation rate.

Seven more donation opportunities were provided, but after the first donation opportunity
students received feedback on the outcome. In particular, the students saw how the non-
donators benefited at the expense of donators. This likely caused students’ preferences for the
public good to change, such that the results from the conjoint survey are no longer pertinent to
the public good experiment. For this reason, we only evaluate the ability of conjoint surveys to
predict behavior at the first donation opportunity.

Which method performed the best at forecasting actual donations? Note that the
forecasts in Table 2 are out-of-sample forecasts. We then may want to use forecast evaluation
techniques to rank each three models. In a recent paper, Norwood, Lusk and Brorsen illustrated
a forecast evaluation method for discrete variable with good small and large sample properties.
This method is referred to as the out-of-sample-log-likelihood function (OSLLF) criterion and is
calculated as follows

\[
(7) \quad \text{OSLLF} = \sum_{i=1}^{55} D_i \ln \left( \Pr(D_i = 1) \right) + \sum_{i=1}^{55} \left[ 1 - D_i \right] \ln \left[ 1 - \Pr(D_i = 1) \right]
\]

where \( i \) denotes the \( i \)th student, \( D_i = 1 \) if the student donated and zero otherwise and \( \Pr(D_i = 1) \)
is the probability of the \( i \)th student donating as given by each model. A higher OSLLF value
indicates more accurate forecasts. The average OSLLF value for each model across each student
is given in Table 2. The uncalibrated model produced the best forecasts, while the certainty- and
frontier-calibrated model performed similarly.

Overall, no single model predicted well. The calibrated models underestimated and
uncalibrated model overestimated donations. Several remedies are available. First, one could
employ a composite model where the projected donation rates from the uncalibrated model and
one of the calibrated models are averaged. This average is 67%, which is closer to the true
donation rate of 76%. Alternatively, if a threshold of six is used in the certainty-calibration
(instead of eight) the projected donation rate is 77%, only one percentage point away from the
true rate.

However, all of these remedies are the result of data-mining. In cases where calibration
provides biased predictions of observed behavior, there is always some ad hoc modification that
will make it unbiased. What the results do show is that neither calibration will always be
unbiased. Prior to this study the frontier-calibration has been tested only once, and was found to
be an unbiased predictor of true values. This study finds the frontier-calibration to be
downward-biased. Two previous studies have tested the certainty-calibration in out-of-sample
predictions (Vossler et al.; Poe et al.).⁶ The certainty-calibration using the eight-threshold was
unbiased in Vossler et al., but was downward-biased in this study and Poe et al. study. At this
point in time, the most we can conclude is that the two calibrations are likely to provide
predictions of actual values that are either unbiased or downward-biased. This suggests that
researchers can use calibrated and uncalibrated values as an upper and lower bounds for true
values.

**Concluding Comments**

The most common critique of stated preference methods is hypothetical bias. Regardless of
whether the good is a private or public good, individuals tend to say they value a good more than
they will actually pay for it. This has led researchers to search for methods of reducing values
derived from hypothetical questions, such that predictions of true values are improved. Two
such methods have illustrated this potential: the certainty-calibration and the frontier-calibration.
Both calibrations are appealing in that they require no additional information other than what is contained in the stated preference experiment. This is the first study to analyze either calibration in a conjoint analysis, and the first to compare them using the same data. The two calibrations provided almost identical predictions of actual values, although they differed on whether men or women donate more. Given the similar results, and the fact that the certainty-calibration is easier to implement, this study prefers the certainty-calibration over the frontier-calibration. However, while the certainty-calibration can only be used in surveys with a specific type of certainty question, the frontier-calibration can be used with any study.

This study found the frontier-calibration to underpredict true values, which is contrary to Hofler and List who found it unbiased. Similarly, we found the certainty-calibration downward-biased. Combining these results with previous studies, one should not expect either calibration to provide unbiased estimates of true values. However, this should not discourage researchers from using calibration. The research conducted thus far yields two general conclusions. First, uncalibrated models will overestimate true values. Second, calibrated models are either unbiased or will underestimate true values. Researchers can therefore use calibrated and uncalibrated models to provide upper and lower bounds to capture true values.
Footnotes

1. To illustrate the provision points mechanism, consider the following experiment conducted in a price analysis class in 2002. Students were given five points extra on an exam. Students were then allowed to donate five points or no points to a public good where the number of donated points were increased by 50% and distributed evenly among donators and non-donators alike. Although the dominant strategy is to not donate, approximately 50% of the students donated. However, the experiment was repeated with a provision point mechanism. Students were told that if 100% of students did not contribute all five points, all contributions to the public good would be refunded in full. In this case, it was everyone’s dominant strategy to contribute, and everyone did donate.

2. The students’ final exam made up 20-30% of the students grade. If a student’s participation grade counted 0%, 5% or 10% of their final grade, the final exam counted 30%, 25% or 20%, respectively.

3. The donations points and minimum participation rate changed in increments of 5 points or 5%, so there are 20 different values they could take. A D-efficiency score was used for experimental design. Survey attributes were repeatedly sampled at random until the D-efficiency measure $\|(X'X)^{-1}\|^1/P$ was maximized, where X is the matrix of explanatory variables and P is the number of parameters. For further details, see Kuhfeld, Tobias and Garratt.
4. This estimated model is \( U = 0.8713 + 1.4026(DON) + 2.8979(MPR) - 1.7039(MALE) \). All parameters except the intercept are significant at the 1% level. The intercept is not significant at the 10% level.

5. A total of eight donation opportunities were provided, whereas students were told they would have ten opportunities. This was to prevent the students’ actions to be influenced by a perception that it was the last time the game would be played.

6. The threshold value of eight was calculated from the Champ and Bishop, so it cannot be counted as an out-of-sample test. By construction, any test of the frontier-calibration is an out-of-sample test. This is because information on real values never enters into the frontier-calibration.
References


<table>
<thead>
<tr>
<th>Variable</th>
<th>Uncalibrated Conditional Logit Model</th>
<th>Certainty-Calibrated Conditional Logit Model</th>
<th>Frontier-Calibrated Mixed Logit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interception</td>
<td>0.4323 (0.4280)</td>
<td>-2.5769 (-2.8605)</td>
<td>-2.8538 (-1.5099)</td>
</tr>
<tr>
<td>Donation Amount</td>
<td>4.0917 (1.9411)</td>
<td>3.9462 (1.6271)</td>
<td>4.0020 (1.9626)</td>
</tr>
<tr>
<td>Donation Amount Squared</td>
<td>-1.7850 (-1.0586)</td>
<td>-2.5286 (-1.3149)</td>
<td>-1.5043 (-0.9139)</td>
</tr>
<tr>
<td>Minimum Participation Rate</td>
<td>1.8774 (0.9201)</td>
<td>-2.1636 (-1.0101)</td>
<td>2.0090 (1.0650)</td>
</tr>
<tr>
<td>Minimum Participation Rate Squared</td>
<td>1.8027 (1.0367)</td>
<td>5.1263 (2.9424)</td>
<td>1.9378 (1.1745)</td>
</tr>
<tr>
<td>Donation Amount * Minimum Participation Rate</td>
<td>-1.6335 (-0.9349)</td>
<td>-1.1695 (-0.6847)</td>
<td>-1.7351 (-1.1228)</td>
</tr>
<tr>
<td>Male Dummy Variable</td>
<td>-1.7065 (-2.1590)</td>
<td>0.5290 (1.6488)</td>
<td>-2.3840 (-1.5424)</td>
</tr>
<tr>
<td>Exponential Distribution Parameter ($\alpha$)</td>
<td>--------------</td>
<td>--------------</td>
<td>0.0683 (0.7207)</td>
</tr>
<tr>
<td>Log-Likelihood Function Value</td>
<td>-186.60</td>
<td>-152.57</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The donation amount $DON$ was divided by 100 and the minimum participation rate was scaled to the interval $(0,1)$. The number of observations was 197.

$^a$ The expected value of the exponential random variable equals $1/\alpha$. 

23
Table 2. Results of Donation Forecasting Contest

<table>
<thead>
<tr>
<th>Model</th>
<th>Percent of Students Donating to Public Good</th>
<th>Average Log-Likelihood Function of Actual Checkoff Donations&lt;sup&gt;a&lt;/sup&gt; (55 observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction from Uncalibrated Conditional Logit Model</td>
<td>Males: 92%</td>
<td>-0.7207</td>
</tr>
<tr>
<td></td>
<td>Females: 98%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All: 94%</td>
<td></td>
</tr>
<tr>
<td>Prediction from Certainty-Calibrated Conditional Logit Model</td>
<td>Males: 43%</td>
<td>-0.8416</td>
</tr>
<tr>
<td></td>
<td>Females: 33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All: 40%</td>
<td></td>
</tr>
<tr>
<td>Prediction from Frontier-Calibrated Mixed Logit Model</td>
<td>Males: 24%</td>
<td>-0.8934</td>
</tr>
<tr>
<td></td>
<td>Females: 68%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All: 39%</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Actual Donations

- Males: 72%
- Females: 84%
- All: 76%

A larger OSLLF value indicates more accurate forecasts.
Appendix A
Example Donation Question

Thank you for attending class. By coming to class today you are awarded 100 participation points.

Please read the following question carefully and check yes or no. Do not discuss this with any classmates.

Name _______________________________

Would you like to donate 80 of your participation points to the PUBLIC FUND? Each of your classmates present today will be asked if they would like to donate the same amount. All donations to the PUBLIC FUND will be increased by 100%, and then distributed equally among all those present, even to those who did not donate 80 points. Those absent from class will not receive any points.

If a minimum participation rate of 70% is not met, the PUBLIC FUND will not be used and all donations to the PUBLIC FUND will be refunded in full to the donor. The minimum participation rate is defined as the percent of people who donate to the PUBLIC GOOD.

At this time, would you like to donate 80 participation points? Please check one.

☐ YES ☐ NO

Note: I will not reveal your answer to anyone. Your answer will be kept strictly confidential.
Appendix B
Example Conjoint Survey Question

1.A) In the table below, please select the donation scenario you prefer, if these were the only options available. If you would not donate under either scenario, simply select “Neither Donation Scenario.”

<table>
<thead>
<tr>
<th>Donation Scenario Attribute</th>
<th>Donation Scenario A</th>
<th>Donation Scenario B</th>
<th>Neither Donation Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Points</td>
<td>30 Points</td>
<td>70 Points</td>
<td>Neither donation scenario A nor B is preferred. If these were the only two options available, I would not make a donation.</td>
</tr>
<tr>
<td>Minimum Participation Rate</td>
<td>85%</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>I would choose (please check ONLY ONE OPTION)</td>
<td></td>
<td>I would prefer to donate to Donation Scenario B</td>
<td>I would not donate in scenarios A or B</td>
</tr>
</tbody>
</table>

If you chose to donate to Donation Scenario A or B, please answer the following question. 1.B) On a scale of 1 to 10, where 1 means "very uncertain" and 10 means "very certain", how certain are you that you would voluntarily donate the points for the donation scenario you chose in Question 1.A if given the opportunity? (CIRCLE ONE NUMBER)

1  2  3  4  5  6  7  8  9  10
very uncertain  very certain