Stated preference methods such as contingent valuation and conjoint analysis have become standard tools for economic and public policy analysis. In cases where policy-makers are interested in estimating the value of non-market goods or those with passive use values, stated preference methods are sometimes the only tools available. The validity of stated preference methods has been debated for several decades in the contingent valuation literature. While not all researchers agree, the general consensus is that most of the arguments against stated preference methods can be avoided by careful design and implementation (see Carson, Flores and Meade). However, even the most ardent supporters of stated preference methods would attest to its disadvantages. Perhaps the greatest drawback to stated preference techniques is its hypothetical nature. People can easily say they will pay a certain amount for a good, but often find giving up actual money to be more difficult.

The tendency to over-commit to a theoretical payment for a good is referred to as “hypothetical bias” and has been found in close to 90% of studies comparing hypothetical to non-hypothetical values (Harrison and Rutstrom). Because hypothetical bias can lead to an overestimation of a good’s value, the NOAA panel on contingent valuation recommended that values estimated from hypothetical questions simply be divided by two. The level of hypothetical bias can be profound, and has been measured as high 300% of a good’s true value (List and Gallat). For years economists have sought to
explain hypothetical bias and to discover methods of removing it from stated preference surveys.

Increasingly, researchers are using calibration techniques to remove hypothetical bias from stated values. Various methods for calibrating hypothetical values to real values have been proposed by Champ and Bishop; Fox et al.; Hofler and List; and List, Margolis, and Shogren, just to name a few. However, virtually all calibration studies suffer from a failure to use theory to substantiate why individuals might overstate their values in a hypothetical setting. That is, published studies tend to report successful attempts to develop calibration methods, despite any theoretical or behavioral evidence that the methods should be successful. Since calibration tests that succeed in removing hypothetical bias are more publishable than tests that fail, readers are left wondering whether the proposed calibration methods would succeed in repeated experiments. Practitioners and policy makers would be more confident in calibration techniques if theoretical or behavioral reasons were provided as to why a particular calibration technique should perform well.

This study demonstrates that hypothetical bias is partially due to the fact that, in hypothetical situations, people are uncertain how they would act if the situation were real and required an actual payment. We refer to this phenomenon as self-uncertainty. We show that self-uncertainty is an underlying driver of hypothetical bias and is a potential reason why two, seemingly disparate, calibration techniques are successful at removing hypothetical bias. First, we confirm the results of Johannesson et al. and show that self-uncertainty is significantly related to hypothetical bias and the certainty-calibration
technique proposed by Champ and Bishop. Second, for the first time, we show that self-uncertainty is also related to the *frontier calibration* method recently proposed by Hofler and List.

In this paper, we remove the veil of mystery surrounding the frontier calibration method, providing evidence that the frontier calibration method, similar to the certainty calibration, operates by indirectly by measuring self-uncertainty through hypothetical bids. We first show that self-uncertainty, as measured on a self-reported scale increasing in certainty from 1 to 10, is positively correlated with hypothetical bias. Our statistical analysis confirms a significant negative correlation between answers to the certainty question and hypothetical bias, lending credence to the certainty calibration approach. We then show that the non-negative random error in the frontier calibration approach is related to responses to the self-uncertainty scale. By demonstrating a link between self-uncertainty and hypothetical bias, this study links two calibration techniques by illustrating their theoretical foundations. Second, answers to the certainty question subsequently enter the stochastic frontier model as an explanatory variable, producing a hybrid-calibration. Although this hybrid-calibration improves forecasts of true values relative to the frontier-calibration, the forecasts from all models are biased.¹

This second result of biased predictions of true values contradicts most findings that calibration provides unbiased predictions of true values. The reason for this contradiction is that our predictions are out-of-sample predictions not subject to data-mining, whereas many of the previous studies employ in-sample predictions or other data-mining tools where good prediction performance is a tautology. Overall, this study
finds that calibration methods can play an important role in stated preference valuation, but one should not expect them to provide unbiased estimates of true values.

This paper is organized as follows. The next section describes the certainty- and frontier- calibrations in more detail. The third section describes our experiment. The next sections show that self-uncertainty is a major cause of hypothetical bias, and that the frontier calibration works by indirectly measuring self-uncertainty. In the results, we determine whether the certainty- and frontier-calibration methods can be profitably combined using auction data and tests the ability of calibrated bids to predict true bids. The last section provides a summary and concluding comments.

Two Calibrations
This section describes two distinct methods of calibrating hypothetical values to predict non-hypothetical or true values. One is the certainty-calibration used for dichotomous choice questions and the second is the frontier-calibration used in hypothetical auctions. The certainty-calibration has taken three different forms in the literature. Champ and Bishop used the method to calibrate stated values for wind energy with actual payments. The authors asked respondents if they would like to purchase a particular amount of wind energy at a particular price. For half of the respondents this was a hypothetical question while for the other half the offer was real. In the hypothetical setting, if the respondent indicated “yes” to the hypothetical purchase opportunity, the following Certainty Question was posed: 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?

Champ and Bishop then used answers to the certainty question to calibrate the “yes/no” responses to the hypothetical question. Not surprisingly, the percentage of “yes” responses to the hypothetical question was larger than percent of “yes” responses to the real offers, indicating a hypothetical bias. However, by assuming that only those who checked 8, 9 or 10 on the certainty question would actually pay the amount asked (i.e. after changing the “yes” responses to “no” if the answer to the certainty question was less than the threshold of 8), the distribution of stated values was indistinguishable from the distribution of actual values.

However, Champ and Bishop’s recoding scheme ensured that calibrated values would be lower than stated values, and logically some threshold for the certainty question scale had to exist that would make hypothetical and true values statistically indistinguishable. Even if self-uncertainty and hypothetical bias were unrelated, there would still be some threshold for which calibrated bids values would equal true values. This result would have been more plausible had the threshold of eight been chosen \textit{a priori}, and then used to calibrate and predict true values.

Johannesson et al. provide another example of where the certainty-calibration appears to provide unbiased estimates of true values, but the reliability of the certainty-calibration remains suspect because they used in-sample as opposed to out-of-sample predictions. Subjects were asked if they would purchase a particular good at a particular price, and then if they answered “yes”, they were presented with a certainty question like
that in Champ and Bishop. A follow-up question then allowed the subjects to actually purchase the good at that price. Not surprisingly, the percentage of “yes” responses to the hypothetical purchase opportunity was larger than the real purchase opportunity, indicating a hypothetical bias. A probit regression was then used to predict the probability of a “yes-yes” response (yes to both the hypothetical and the real purchase opportunity) as opposed to a “yes-no” response based on a subjects’ answer to the certainty question.

The certainty-calibration by Johannesson et al. was then performed as follows. If a subject answered “yes” to the hypothetical question, but the predicted probability of a “yes-yes” response from the probit model was less than 50%, the “yes” answer to the hypothetical question was changed to “no.” After recoding the data, the percent of “yes” responses in the hypothetical and real samples were statistically indistinguishable. Thus, they conclude based on within-sample calibration that this certainty-calibration provides unbiased predictions of real values.

The parameter associated with the certainty question in the probit regression by Johannesson et al. was significantly positive, indicating self-uncertainty indeed caused hypothetical bias. However, it is not surprising that the probit regression provided an unbiased prediction of the true percentage of “yes-yes” responses, because it was an in-sample prediction. In the probit estimation, the parameters were chosen to provide [asymptotically] unbiased predictions. While this does not suggest that the certainty-calibration is not useful, their results would be more compelling if the probit regression has been used to predict “yes-yes” responses from a different pool of subjects, i.e., out-
of-sample predictions would provide a much better test of calibration performance than in-sample predictions.²

There are some studies that test calibration performance using out-of-sample predictions. Blumenschein et al. used a certainty question where, if a subject stated “yes” she would hypothetically purchase the good, she could only select “definitely sure” and “probably sure” instead of the one-to-ten scale. After recoding “yes” and “probably sure” answers to “no”, answers to the hypothetical and real dichotomous choice question were statistically indistinguishable. The salient feature of this study is that, unlike the previous two studies, the good calibration performance was not a tautology. It was not designed so that it must predict well. This occurred because answers to the real purchase opportunity did not enter the calibration design, and therefore the results were out-of-sample predictions whose performance, a priori, had just as much chance to fail as it did to succeed. Only when we turn to out-of-sample predictions does the certainty-calibration appear imperfect. The Johannesson, Liljas and Johannson experiment used the same methods as Blumenschein et al., but found the certainty-calibration under-predicted the number of true “yes-yes” responses.

The frontier-calibration, designed by Hofler and List for use in hypothetical Vickrey Auctions, provides an alternative to using a certainty question for calculating hypothetical bias in bids for goods. Hofler and List define the process generating actual bids in a Vickrey Auction for a private good for person i as

\[ Y_i^A = X_i\beta + v_i \]
where $Y_i^A$ is the actual (non-hypothetical) bid or true value, $v_i$ is normally distributed with a zero mean, $X_i$ is a vector of demographics and $\beta$ is a conformable parameter vector. A hypothetical bias exists when people overstate their true bid in hypothetical questions. Assume this bias can be depicted by the non-negative random error $\mu_i = Y_i^H - Y_i^A$, where $Y_i^H$ is a hypothetical bid. The process driving hypothetical bids can then be modeled as 

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

In stated preference studies researchers can only observe the error term $\varepsilon_i = v_i + \mu_i$. However, by assuming particular distributions for $v_i$ and $\mu_i$, one can obtain an estimate of $\mu_i$ based on the estimate of $\varepsilon_i$. It is common to assume that $v_i \sim N(0, \sigma^2)$ and that $\mu_i$ is either half-normal or exponential (Beckers and Hammond; Hofler and List; Reifschneider and Stevenson; Kumbhakar and Lovell; Jondrow et al.). In either case, the expected value of $\mu_i$ is increasing in $\varepsilon_i$ (see Jondrow et al.). Therefore, a larger bid residual, $\varepsilon_i$, implies a larger predicted hypothetical bias. Thus, the frontier calibration works by assuming hypothetical bias is positively correlated with bid residuals. The steps to obtaining the calibrated bids are as follows:

**Frontier Calibration Method**

Step A) Estimate a hypothetical bid function using a stochastic frontier function $Y_i^H = X_i\beta + v_i + \mu_i$ where $v_i \sim N(0, \sigma^2)$ and $\mu_i$ is a non-negative random variable.

Step B) For each individual, calculate the expected value of $\mu_i$ conditional on the observed residuals $\varepsilon_i = v_i + \mu_i$, denoted $E(\mu_i|\varepsilon_i)$. 
Step C) Obtain a frontier-calibrated bid \( \hat{Y}_i \) by calculating

\[
\hat{Y}_i = \left( \frac{X_i \hat{\beta}}{X_i \hat{\beta} + E(\mu_i | \epsilon_i)} \right) Y_i^H \text{ where } \hat{\beta} \text{ is the maximum likelihood estimate of } \beta. \]

The above model is constructed such that hypothetical bias \( \mu_i \) must be a component of the observed error \( \epsilon_i \). However, it is just as easy to construct a model where this is not the case. If we specify hypothetical bids to follow a stochastic frontier, then hypothetical and real bids could be stated as given below:

(3) \( Y_i^H = X_i \beta + v_i = X_i \beta + \epsilon_i \)

(4) \( Y_i^A = X_i \beta + v_i - \mu_i. \)

In this case, \( \epsilon_i = v_i \) and conditional bid residuals are uncorrelated with hypothetical bias, implying the correlation between \( \epsilon_i \) and \( \mu_i \) is an empirical question. If they are indeed correlated, then the residuals from a stochastic frontier function provide indirect estimates of hypothetical bias, and may improve predictions of true bids.

The certainty-calibration operates by assuming answers to the certainty question are correlated with hypothetical bias, and the frontier-calibration assumes bid residuals are correlated with hypothetical bias. If both assumptions are correct, it stands to reason that the two calibrations can be profitably combined to form a hybrid-calibration by making the distribution of the hypothetical bias (\( \mu_i \) in the frontier calibration) dependent upon the certainty question. The next section describes an experiment used to collect hypothetical and real bids from a Vickrey Auction. The third section then demonstrates that the assumptions underlying both the certainty- and the frontier-calibration are correct, and the fourth section develops and tests a hybrid calibration.
Experiment Description

The data used in this study are from an auction designed to mimic the Hofler and List field experiment. The subjects were students in an undergraduate agricultural economics class. Students were shown a portable lawn chair with the university colors and logo and were asked to submit sealed hypothetical bids. Students were asked to bid as if the auction were real. The hypothetical auction was described as a Vickrey Auction and students were given examples of the Vickrey Auction process prior to the experiment. Students were encouraged to ask questions about the auction to ensure that the rules were fully understood, and were verbally instructed to participate as if the bidding process were real. After writing their hypothetical bids, they were asked to complete a certainty question asking them how certain they are on a scale of one to ten they would submit a bid equal to or greater than their hypothetical bid if the auction were real. The certainty question was similar to the one used by Champ and Bishop.

Comparing hypothetical to non-hypothetical bids allows a direct measurement of hypothetical bias, so an actual auction followed the first experiment. After the hypothetical bids were collected, the students were asked to submit real bids for a Vickrey Auction. The students were told that the auction winner would have to pay the second highest bid amount by cash or check within one week. Students were told to sign a form indicating that they understood this second auction was real. The two auctions were held at the beginning of class, so there was no reason for the students to hurry through the experiment. Demographic characteristics such as age, sex and frequency of outdoor activities were collected, but none of these were significantly correlated with
hypothetical bids. Table 1 provides the descriptive statistics of the experiment. Hypothetical bias was pervasive in this experiment. The average hypothetical bid was 90% higher than the average real bid.

**Self-Uncertainty, Hypothetical Bias and Bid Residuals**

Hypothetical bias may arise from confusion, a desire to answer the question quickly, a strategic maneuver to free-ride or a desire to manipulate the outcome. Bias from confusion and anxious participants can be avoided by a well-designed survey or experiment. This article focuses on the hypothetical bias for stated preference studies dealing with private goods, so free-riding is irrelevant, though certainly a major factor in other settings such as the valuation of public goods. In this section, we test the hypothesis that hypothetical bias is partially the result of self-uncertainty. In hypothetical situations, individuals are simply unsure how they would behave in real situations. This leads them to say they will pay more for a good than they really will, and causes them to submit higher hypothetical bids than individuals with similar demographic characteristics.

First, the hypothesis that hypothetical bias stems partially from self-uncertainty is tested. Self-uncertainty was measured by the students’ answer to the certainty question, where a lower value implies greater self-uncertainty. Hypothetical bias was calculated by subtracting actual bids from hypothetical bids. Hypothetical bias was then regressed against answers to the certainty scale question. The results shown in table 2 demonstrate a positive relationship between self-uncertainty and hypothetical bias. A lower value on
the certainty question is associated with greater hypothetical bias. Although this finding
does not imply that self-uncertainty is the only cause of hypothetical bias, it is a
significant component. This corroborates the Johannesson et al. finding that self-
uncertainty does provide information on the hypothetical bias, and that certainty-
calibration is not simply an arbitrary method of reducing hypothetical values.

Next, we tested the underlying assumption of the frontier-calibration that the bid
residuals were positively correlated with hypothetical bias. Demographic variables were
found to have no effect on hypothetical bids, so hypothetical bids were modeled as $Y_i^H =
\beta_0 + \eta_i$ using ordinary least squares where $\eta_i$ is an error term. The failure of participants’
demographic characteristics to prove significant and useful in improving bid predictions
has been noted in other studies (Umberger and Feuz; Johannesson et al.). In cases where
demographics significantly affect bid amounts, the predicted bid should also be a
function of demographic data. While $\eta_i$ is not a non-negative error term like those
employed in frontier models, a higher value still implies a greater bid residual. Thus, a
larger value of $\eta_i$ implies a larger value of $\epsilon_i$ in a frontier model. Table 2 shows the result
of a regression of hypothetical bias on bid residuals, and the correlation is statistically
significant and positive. The high coefficient of determination compared to the certainty
question regression suggests that conditional bid residuals explain more hypothetical bias
than answers to certainty questions.

Table 2 also shows a regression of hypothetical bid residuals against answers to
the certainty question, revealing a positive and significant correlation between self-
uncertainty and bid residuals. From this, one can conclude that self-uncertainty causes
people to not only overestimate how much they are willing to pay, but to also submit higher hypothetical bids than those of similar demographics.\textsuperscript{4} Self-uncertainty results in greater bid residuals and greater hypothetical bias, thus providing theoretical justification for the frontier-calibration.

The auction results suggest that the frontier calibration and the certainty calibration are intimately related. This naturally leads one to wonder whether they can be profitably combined. The certainty-calibration was designed for dichotomous choice questions, but can easily be modified for use in auctions. It was previously shown that the frontier-calibration works because the error term $\mu_i$ in (2) is an indirect measure of self-uncertainty. Since the certainty question is a direct measure of self-uncertainty, one can make the distribution of $\mu_i$ conditional on answers to the certainty question. The next section tests whether this modification improves inferences obtained from hypothetical bids.

**Combining Calibrations**

This section utilizes the auction data described previously to determine whether embedding answers to the certainty question within the frontier calibration approach provides better forecasts of observed bids. This is referred to as a *hybrid-calibration*. Hypothetical bids are modeled as:

$$Y_i^H = X_i\beta + v_i + \mu_i = X_i\beta + \varepsilon_i$$
where $X_i \beta + v_i$ is the stochastic true bid and $\mu_i$ is the hypothetical bias. The distribution of $v_i$ is assumed $N(0, \sigma^2)$ while $\mu_i$ is allowed to follow two possible distributions. The first is an exponential distribution whose distribution function is given by

$$f(u_i) = \alpha \exp(-\alpha_i \mu_i).$$

(6) The expected value of $\mu_i$ is $\alpha_i^{-1}$. Because $\mu_i$ is the indirect measure of hypothetical bias in the frontier-calibration approach, this measure might be enhanced by including information about individuals’ level of uncertainty. Individuals’ answers to the certainty question ($C_i$) are incorporated by specifying $\alpha_i$ to follow

$$\alpha_i = \alpha_0 + \alpha_1(C_i/10)$$

(7) where $C_i$ is divided by ten to facilitate convergence. To test whether the certainty question improves predictions of actual bids, the model is also estimated where $\alpha_1$ is constrained to equal zero. This constraint produces the frontier-calibration model proposed by Hofler and List. With some modification, the log-likelihood function for $\varepsilon_i = v_i + u_i$ is given by Aigner, Lovell and Schmidt as

$$LLF = \sum_i \left[ \ln(\alpha_i) + 0.5\alpha_i^2 \sigma^2 - \alpha_i (\varepsilon_i) + \ln \left( \Phi \left( \frac{\varepsilon_i}{\sigma} - \sigma \alpha_i \right) \right) \right]$$

(8) where $\Phi$ is the standard normal cumulative distribution function. Predicted hypothetical bias for each individual is calculated by determining the expected value of $\mu_i$ given the residual $\varepsilon_i$. This expectation is (Aigner, Lovell and Schmidt).
(9) \[ E(\mu_i \mid \varepsilon_i) = \sigma \left( \frac{\phi(A_i)}{1 - \Phi(A_i)} - A_i \right) \]

where \( A_i = -\varepsilon_i/\sigma + \sigma \alpha_i \). As the estimated residual, \( \hat{\varepsilon}_i = Y_i^H - X_i \hat{\beta} \) increases, the level of the predicted hypothetical bias increases as well. In the alternative specification, the error term \( \mu_i \) is allowed to follow a half-normal distribution where \( \mu_i \sim N(0, \alpha_i^2) \). The log-likelihood function for this model is (Kumbhakar and Lovell)

\[
\text{LLF} = \sum_i \left[ -\ln(\tilde{\sigma}^2) + \ln \left( 1 - \Phi \left( \frac{-\left( Y_i^H - X_i \hat{\beta} \right)}{\tilde{\sigma}} \right) \right) - 0.5 \tilde{\sigma}^{-2} \left( Y_i^H - X_i \hat{\beta} \right)^2 \right]
\]

(10) \[ \tilde{\sigma} = \sqrt{\sigma^2 + \alpha_i^2} \]
\[ \lambda = \alpha_i / \sigma \]

and the estimated hypothetical bias is

\[
E(\mu_i \mid \varepsilon_i) = \sigma^* \left[ \frac{\phi \left( \frac{\varepsilon_i \lambda_i}{\tilde{\sigma}} \right)}{1 - \Phi \left( \frac{-\varepsilon_i \lambda_i}{\tilde{\sigma}} \right)} + \varepsilon_i \lambda_i / \sigma \right]
\]

(11) \[ \tilde{\sigma} = \sqrt{\sigma^2 + \alpha_i^2} \]
\[ \sigma^* = \alpha_i \tilde{\sigma}^{-1} \]
\[ \lambda_i = \alpha_i / \sigma \]

The interpretation of \( \alpha_i \) differs in the normal / half-normal model than the normal / exponential model. An increase in \( \alpha_i \) decreases the predicted hypothetical bias in the exponential case and increases it in the half-normal case.
Estimates for the normal / exponential and the normal / half-normal models with
and without the certainty question are shown in table 3. As expected, $\alpha_1$ is negative for
the exponential model and positive for the half-normal model, implying that individuals
who express greater certainty are projected to exhibit less hypothetical bias. This also
demonstrates the hypothesized casual relationship between self-uncertainty and bid
residuals. The test statistics suggest that $\alpha_1$ is significant in both models at the 10% level.
Likelihood ratio tests reject the null hypothesis that $\alpha_1 = 0$ in both cases at the 1% level.\(^7\)

An interesting question is whether incorporating answers to the certainty question
in the frontier model of hypothetical bids improves forecasts of actual bids. This test was
conducted using a bootstrap method that works as follows. Let $W_i^{\text{NOCERTAIN}}$ be the
squared forecast error, which is the squared difference between the calibrated bid and the
actual bid for the $i^{\text{th}}$ person when $\alpha_1$ (the indirect hypothetical bias measure) is
constrained to equal zero. Similarly, let $W_i^{\text{CERTAIN}}$ be the squared forecast error when the
value of $\alpha_1$ is unrestrained. Finally, let $W_i = W_i^{\text{NOCERTAIN}} - W_i^{\text{CERTAIN}}$. The contribution
of the certainty question to frontier calibration is measured by testing the null hypothesis
that $E(W_i) = 0$ versus the alternative hypothesis the $E(W_i) > 0$. Values of $W_i$ are expected
to be non-normally distributed, so a nonparametric bootstrap is conducted. A total of
1,000 bootstraps are conducted where, within each bootstrap, individual values of $W_i$ are
randomly chosen with replacement to yield 48 simulated $W_i$ values, where the average
$W_i$ at each bootstrap is denoted $\overline{W}_i$. The percent of positive $\overline{W}_i$’s can serve as a p-value
for the statistical test. In 100% of bootstraps $\overline{W}_i$ was positive for the normal / exponential
and the normal / half-normal case, indicating that embedding the certainty calibration
within the frontier calibration reduces the forecast error between predictions of actual bids from hypothetical bids and actual bids. Therefore, we conclude that the inclusion of certainty question in the hybrid-calibration significantly improves forecasts of actual bids.

**Comparing Calibrations**

Next, we analyze the ability of the hybrid-calibration to improve forecasts of true values relative to uncalibrated bids and other calibration methods. Forecast accuracy was measured by taking the squared difference between calibrated bids and true bids, averaging this squared difference across all individuals and taking its square root. This is referred to as the root-mean-squared error (RMSE). As table 3 shows, the RMSE is lower using the normal / exponential model than the normal / half-normal model (despite the fact that the normal / half-normal model had a higher log-likelihood function value), and the RMSE improves under both models when the certainty question is used.

The RMSE when using uncalibrated bids to forecast true bids was $22.84, indicating that all four models in table 3 improve inferences from stated values. As cited before, the NOAA panel once recommended that stated values be reduced by one-half. We refer to this approach as the NOAA-calibration. How do the calibrated bids from the models in table 3 compare to a system where hypothetical bids are divided by two? Surprisingly, the NOAA-calibration resulted in the lowest RMSE of 10.93. The average NOAA-calibrated bid was $13.41, which was very close to the average true bid of $14.19. In another study, Fox et al. use hypothetical and real auction data to calculate
calibration factors of 0.87 for males and 0.63 for females, which when applied to our data yields a RMSE of 16.39. This suggests that the frontier- and hybrid-calibration produces predictions of real bids which are within the range of existing calibration methods.

Next, we evaluate the ability of calibration to provide unbiased estimates of mean true values. To test the null hypothesis that the average calibrated bid is identical to the average true bid, a nonparametric bootstrap as described previously was performed. Nonparametric bootstraps were used because real bids are not normally distributed, as shown in figure 1. Let $D_i$ equal the calibrated bid minus the real bid for subject $i$. A total of 1,000 bootstraps were performed where at each bootstrap, original values of $D_i$ were sampled with replacement. Values of $D_i$ were sampled, as opposed sampling calibrated and real bids separately, because calibrated and real bids are correlated. At each bootstrap the average value of $D_i$, denoted $\overline{D}_i$, was calculated.

The percent of times $\overline{D}_i$ is positive can be interpreted as the p-value in a statistical test. If the percent of positive $\overline{D}_i$’s is greater than 95% (less than 5%), we can reject the null hypothesis that $E(D_i) = 0$ in favor of the hypothesis that it is positive (negative) at the 5% level. For the frontier- and hybrid-calibrations, the average $D_i$ was negative in 100% bootstraps. For the Fox-calibration the average $D_i$ was positive in 100% bootstraps. Only the NOAA-calibration was unbiased, as $D_i$ was positive 28% of the time. This test concludes that, for these data, the frontier- and hybrid-calibrations under-predict true values, the Fox-calibration over-predicts true values, and the NOAA-calibration is unbiased.
Another important distinction between the calibrations is made when we look at the distribution of predicted and real bids. Both the NOAA- and the Fox-calibration depict the variability of real bids much better than the frontier- or hybrid-calibrated bids. See figure 1 containing histograms for bids from the frontier-, hybrid-, NOAA- and Fox-calibrated bids and the real bids.

Hybrid- and frontier-calibrated bids fail at accurately depicting the variability of true values. The distribution from the NOAA- and Fox-calibrated bids provide a much better picture of the variance of true values. In cases where the researcher is interested in the distribution of values, and not just the average value, the frontier- and hybrid-calibration method leaves much to be desired in these data. Overall, the NOAA-calibration performed better at depicting both the mean and variance of true values. This of course does not imply that the NOAA-calibration is always the preferred technique, as its success may be isolated to the data in this study.

**Concluding Remarks**

This study provides empirical evidence that individuals with greater self-uncertainty (those who are less confident in predicting how they would behave in non-hypothetical situations) display greater hypothetical bias and larger bid residuals (they tend to submit larger hypothetical bids relative to those of similar demographics). Using this as a working assumption in a behavioral model, this provides theoretical justification for the certainty- and the frontier-calibration. Both calibrations were combined to produce a
hybrid-calibration that proved useful for predicting true values based on hypothetical auctions.

Although results suggest that calibration does improve predictions of true values, contrary to other studies we find that most calibrated values are biased estimates of true values. However, this is not unexpected. Unless we are given economic or statistical reasons to believe that calibrated values are unbiased predictors of true values, studies demonstrating specific instances where they appear unbiased should be interpreted with caution, especially considering that manuscripts illustrating a failed calibration attempt have little chance of being published.

Calibration--the act of modifying hypothetical values to predict real values--is a daunting task. Essentially, calibration entails using one random variable to estimate another random variable, where the only link between the two variables is a correlation. While this study aptly demonstrates that this correlation is conditional on self-uncertainty, this finding also indicates that the literature on calibration remains far from complete. The fact that the more sophisticated frontier calibrations were inferior to the arbitrary NOAA-calibration reconfirms that the science of calibration is still young.

Perhaps calibrated values will always be biased estimators of true values. Nonetheless, increasing attention to calibration of hypothetical values to actual values for all goods, particularly non-market goods, will shed light on magnitude and distribution of hypothetical bias, increasing the accuracy of stated preference techniques. The purpose of this paper, as with all calibration studies, is to render the issue of hypothetical bias more benign in the policy debate over the merits of stated preference techniques.
Footnotes

1. The definition of “true” values throughout this paper is values revealed through actual payments of money in an incentive-compatible valuation mechanism. In our data, the true values are real Vickrey Auction bids.

2. Let \( f(C_i, \beta) \) be a function predicting the probability of a “yes-yes” response based on the \( i^{th} \) subject’s answer to the certainty-question, denoted \( C_i \). A value of \( C_i = 1 \) (\( C_i = 10 \)) indicates low (high) certainty. \( \beta \) is a parameter vector. Let \( I_i = 1 \) if the \( i^{th} \) subject answers “yes-yes” and zero otherwise. The Johannesson study evaluated forecasts of \( I_i \)’s using a parameter vector \( \beta \) that was estimated from the true value of the \( I_i \)’s. That is, the same group of individuals was used to estimate \( \beta \) and to judge calibration performance. For this reason, the predictions of the Johannesson study are in-sample predictions, and the value of \( \beta \) was chosen to maximize prediction accuracy. If \( \beta \) was estimated from one group of students, and then used to predict \( I_i \)’s from another group of students, the predictions would be out-of-sample.

3. Hofler and List also consider an alternative calibration where hypothetical bids are calibrated using the unconditional expectation of \( \mu_i \). This entails replacing Step 3 of the frontier calibration with “Obtain a frontier-calibrated bid by multiplying \( Y_i^H \) by

\[
\hat{Y}_i = \left( \frac{X_i \hat{\beta}}{X_i \hat{\beta} + \hat{E}(\mu_i)} \right) Y_i^H \text{where } \hat{\beta} \text{ is the maximum likelihood estimate of } \beta.
\]

This method performed equally well at removing hypothetical bias.
4. In our case, the predicted hypothetical bid was not contingent on demographics because they did not help explain bids. For this reason, we group all the students who participated into the same demographic. When demographics are important, the bid residual would be expressed as $\eta_i = Y_i^H - X_i \beta$ where $X_i$ is a data matrix and $\beta$ is a conformable parameter vector.

5. Aigner, Knox Lovell and Schmidt derive the log-likelihood function for the case when a non-negative error is subtracted from the frontier. We wish to thank William Greene of the Stern School of business for providing the log-likelihood function when the error is added to the frontier.

6. Assistance from William Greene is again acknowledged as in Footnote 5. Simulations were conducted to ensure (8) and (9) are correct, since they could not be derived.

7. The likelihood ratio test statistic for the null hypothesis that $\alpha_1 = 0$ for half-normal model is $2(201.75 - 199.49) = 4.52$ for the exponential model and $2(192.60-190.08) = 5.04$. The p-values for these test statistics are 0.03 for both.

8. A calibration factor of 0.87 implies that hypothetical bids should be multiplied by 0.87 to obtain predictions of true bids. See page 463 of Fox et al. for more detail on how these calibration factors were derived.
References


Hofler, R. and J. A. List. “Valuation On The Frontier: Calibrating Actual and


Umberger, W. J. and D. M. Feuz. “The Usefulness of Experimental Auctions in
<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean (Standard Deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothetical Bid</td>
<td>$26.82 (20.65)</td>
</tr>
<tr>
<td>Actual Bid</td>
<td>$14.19 (9.30)</td>
</tr>
<tr>
<td>Answer to Certainty Question</td>
<td>8.08 (2.15)</td>
</tr>
<tr>
<td>Number of Participants</td>
<td>48</td>
</tr>
</tbody>
</table>
Table 2. Relationship Between Certainty Question, Hypothetical Bid Residuals and Hypothetical Bias

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variable = Hypothetical Bias = (Hypothetical Bid – Real Bid)</th>
<th>Dependent Variable = Hypothetical Bid Residual&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimate &lt;br&gt;(T-Statistic In Parenthesis)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>43.3671 &lt;br&gt;(18.35)*****</td>
<td>12.6333 &lt;br&gt;(29.80)***</td>
</tr>
<tr>
<td>Value of Hypothetical</td>
<td>0.8321 &lt;br&gt;(40.10)***</td>
<td></td>
</tr>
<tr>
<td>Bid Residual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer To Certainty</td>
<td>-3.8021 &lt;br&gt;(-13.45)***</td>
<td>-3.3983 &lt;br&gt;(-2.62)*****</td>
</tr>
<tr>
<td>Question</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient of</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>Determination</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** and ** Indicates significance at the 1% and 5% level, respectively.

<sup>a</sup> Hypothetical bid residuals are calculated as the residuals from a linear regression of hypothetical bids on an intercept.
Table 3. Stochastic Frontier Estimation

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Normal / Half-Normal Model</th>
<th>Normal / Exponential Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Certainty Question</td>
<td>With Certainty Question</td>
</tr>
<tr>
<td><strong>Parameter Estimate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(T-Statistic)</td>
</tr>
<tr>
<td>Intercept</td>
<td>6.9522 (3.93)***</td>
<td>7.7079 (3.33)***</td>
</tr>
<tr>
<td>σ</td>
<td>4.9729 (2.88)***</td>
<td>5.3663 (2.90)***</td>
</tr>
<tr>
<td>α₀</td>
<td>0.0503 (5.66)***</td>
<td>0.0104 (0.7644)</td>
</tr>
<tr>
<td>α₁</td>
<td>0.0574 (2.4730)**</td>
<td></td>
</tr>
<tr>
<td>Log-Likelihood Function</td>
<td>-201.75</td>
<td>-199.49</td>
</tr>
<tr>
<td>Mean Calibrated Bid</td>
<td>6.61</td>
<td>7.30</td>
</tr>
<tr>
<td>Root-Mean Squared Error From Using Calibrated Bid To Forecast True Bids</td>
<td>11.51</td>
<td>10.99</td>
</tr>
</tbody>
</table>

Notes: Due to the non-linearity of this model, parameter estimates are sensitive to starting values. The values shown here are the estimates that provide the highest likelihood function value after trying many different starting values. The superscripts *, **, and *** denote significance at the 10%, 5% and 1% level, respectively.
Figure 1. Distribution of Calibrated and Real Bids