SIEPR Discussion Paper No. 06-42

A Test of Confidence Enhanced Performance: Evidence from US College Debaters

By
Jonathan Meer
Stanford University
And
Edward D. Van Wesep
Stanford University

August 2007

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

A Test of Confidence Enhanced Performance:
Evidence from US College Debaters

Jonathan Meer
Stanford University

Edward D. Van Wesep
Stanford University

July 2007

*We gratefully thank B. Douglas Bernheim, Evan L. Morrison, Harvey S. Rosen and Andres Santos. This research was supported in part in part by the Stanford Institute for Economic Policy Research and the Taube Foundation.

Abstract

We test the theory put forth by Compte and Postlewaite (2004) that overconfidence might persist because it is welfare improving. They argue that because confidence enhances performance, some overconfidence is optimal in spite of its negative effect on decision-making. One implication of their model is that while an agent's bias (first moment of prediction error) may not change as she gains experience in an activity, her predictive accuracy (second moment of prediction error) should improve. We test this implication by comparing predictions of success by university debaters with outcomes in debate rounds and evaluating how the first and second moments of their prediction errors change with experience. As predicted by the theory, we find that while debaters remain overconfident in spite of experience, they become more accurate in their predictions. These findings support the view that overconfidence may persist because it is welfare improving. JEL D83, D84

Jonathan Meer
Department of Economics
579 Serra Mall
Stanford University
Stanford, CA 94301
jmeer@stanford.edu

Edward D. Van Wesep
Department of Economics
579 Serra Mall
Stanford University
Stanford, CA 94301
vanwesep@stanford.edu

1. Introduction

The study of learning and incentives is central to many areas of economics. For instance, the standard justification for equilibrium analysis is not that agents make the complex calculations to derive how they should play, but rather that as they become experienced, an equilibrium naturally develops. The market provides incentives to learn how to behave optimally.

A considerable literature in the field of psychology as well as recent work in experimental economics has shown that people are systematically overconfident in their abilities and that experience is typically not sufficient to eliminate this overconfidence. Doctors' and nurses' perceptions of their medical knowledge do not correlate with demonstrated knowledge (Tracey *et al.* (1997); Marteau *et al.* (1989)), while the confidence of witnesses to a crime in identifying suspects in a line-up is only weakly correlated with accuracy (Sporer *et al.* (1995)).

Even when correlations between perceived aptitude and performance exist, they tend to be small. A 1982 meta-analysis by Mabe and West of 55 studies of self-perception and performance found an average correlation of 0.29. Moreover, these low correlations are not caused by noise: Harris and Schaubroeck (1988) find that the correlation of employee self assessments with that of peers and supervisors is roughly 0.35 but the correlation of peer and supervisor assessments is closer to 0.6.[1] Apparently, one's peers are more able to judge aptitude than one's self.

Public and private sector decision-makers must understand the nature and size of agent overconfidence in order to make optimal decisions. For example, a government determining optimal tax incentives for entrepreneurship must assess the degree to which entrepreneurs perceive opportunities, not the actual level of opportunities available. A board deciding whether to agree to a merger proposed by its CEO must understand the degree to which CEO projections are inflated by overconfidence. Malmendier and Tate

[1] See Ehrlinger and Dunning (2003) for further references on the topic

(2005, 2007) show that CEO overconfidence can explain the link between high free cash flow and excessive corporate investment as well as the negative market reactions to acquisition announcements. Massey and Thaler (2005) show that overconfidence can explain overpricing of draft picks by NFL general managers (which has large financial consequences for the teams involved).

The economic literature (mostly) implies only negative effects from overconfidence, but as overconfidence appears to be ingrained in the human psyche, it seems likely that it serves some purpose. Compte and Postlewaite (2004) argue that if confidence itself enhances performance, some overconfidence is welfare improving even if it sometimes leads to poor decision-making. Because their paper began with the observation that confidence often improves performance, a test of this theory must focus on more subtle implications of the model. This paper is an attempt to perform such a test.

We ask whether American college students participating in debate competitions are able to more accurately assess how they have performed in debates as they gain experience. After each contest, debaters were separately asked what each believed was the likelihood of having won the contest and were offered monetary incentives to provide accurate answers. A prediction can be modeled as follows: $\mathbf{p=\varphi+\varepsilon}$ where $\boldsymbol{\varphi}$ is the true probability of having won and $\boldsymbol{\varepsilon}$ is an error term. A debater can be said to make better predictions if the first and second moments of $\boldsymbol{\varepsilon}$ are closer to 0.

Debate tournaments are ideal for testing this theory: Speaking in public is one of the examples used by Compte and Postlewaite for which lack of confidence (via a trembling voice, for example) can impair performance. Furthermore, the method by which agents gain overconfidence in their paper is by attributing past success to skill and past failure to chance. Their primary example is of a lawyer:

> “When he loses, he may think that this was due to the fact that the defendant was a member of a minority group while the jury was all white: the activity was a failure, but the reasons for the failure are seen as atypi-

> cal and not likely to arise in the future. He may then disregard the event when evaluating his failure."

Debaters, much like lawyers, may well attribute losses to judge error (or a preconceived bias in favor of the opponents' side) rather than poor performance. Debate tournaments also feature participants with a range of experience and will allow us to estimate the effects of experience on prediction bias and accuracy using cross-sectional data.

We find that debaters tend to be overconfident and that this overconfidence does not decline with experience. Skill, on the other hand, does tend to reduce overconfidence, but only elite debaters are unbiased in their predictions.[2] Debaters do, however, become more accurate in their predictions (as measured by mean squared error) as they gain experience. These results are in line with Compte and Postlewaite's theory: debaters do learn about win likelihoods over time as they gain experience, but appear to do so without reducing their overconfidence. Intuitively, while confidence is necessary for success in debate, this benefit to inaccuracy is counterbalanced with benefits to accurately interpreting events in a round. The former effect pulls the first moment of $\boldsymbol{\varepsilon}$ away from zero, but the latter pulls all moments toward zero.

The paper is organized as follows: Section 2 describes the structure of debate tournaments; Section 3 outlines the data collection process and presents summary statistics; Section 4 describes the estimation strategy; Section 5 outlines our results; Section 6 describes various robustness checks and Section 7 concludes.

[2] The fact that more skilled debaters are less biased is also in line with previous research in psychology. Kruger and Dunning (1999) argue that the incompetent are simply not competent enough to make accurate assessments. As they gain competence in a task, they also gain competence in the assessment, decreasing forecast bias. Furthermore, elite debaters have high win likelihoods and therefore little room for overconfidence.

2. Debate Tournaments

Two-person teams from schools across the United States regularly attend tournaments hosted by debating unions of member universities. Host schools provide judges but do not compete themselves. A tournament has five rounds in which each team competes against another attending team. In rounds one and two, teams are randomly paired against each other and each pair of teams is randomly assigned a judge.[3] Within a particular pairing, the two teams debate a resolution proposed by the "Government" team (where the "Government" designation is randomly assigned in the initial pairings). After the debate is finished, the teams leave the room and await their pairings for the next round while the judge assesses the teams' performances.

First, the judge assigns a rank of 1-4 to each of the four participants where a rank of one means that participant was the best debater in the round. Second, the judge must assign an integer value between 1 and 30 to each debater as her "Speaker Points."[4] These are meant as an objective ranking of how each debater performed and are not relative to the other participants in the round. Finally, the judge must determine which team won. The losing pair can have neither a higher combined speaker point total nor a lower combined rank total.

Judges submit their assessments and the information is aggregated for use in assigning pairings for the following round. In the final three rounds, teams are paired against other teams with the same win/loss record. Within each win/loss group, the team with the highest combined speaker points for prior rounds faces the team with the lowest such value, the team with the second highest faces the team with the second lowest, etc. There is some discretion taken by the organizers when the number of teams in

[3] Pairings are not entirely random. Each school assigns their own teams designations of A, B, etc. where the best team from that school is assigned A, the second best B, etc. In rounds 1 and 2, no two A teams can meet and no teams from the same school can meet. The former rule prevents good teams from meeting too early in the tournament and the latter is to ensure that the rounds are interesting for all participants.

[4] In practice, nearly all speaker points awarded are between 22 and 28

a bracket is odd or when teams from the same school are matched, but these occurrences are unusual.

After completion of the five rounds described above, the top four or eight teams, depending on the tournament, are invited to join a playoff called the "out-rounds" to compete for the tournament championship.

3. Data

The data for this study were collected at the Brown University Debate Tournament in November 2001. This tournament was part of the American Parliamentary Debate Association (APDA) circuit for the 2001-2002 school year and was the 9th tournament (out of approximately 25) of the year.[5]

Immediately following the completion of each round, debaters were asked to predict the likelihood that they won said round. Specifically, they were asked upon returning from a debate to write the likelihood that they won the previous round on an index card with their unique debater ID and drop it in a box. Debaters were not informed by judges as to whether they won a particular round until the end of the tournament and, while their pairings in subsequent rounds did reveal some information as to their likely win/loss bracket, their predictions were taken prior to the announcement of pairings in each round.

In order to elicit meaningful predictions, debaters were informed that after completion of the tournament, the ten most accurate debaters who participated in all five rounds would be entered into a raffle for $1,000, where accuracy is defined to be the mean squared error where error is prediction minus outcome. The outcome is binary, taking a value of 1 if the debater won the previous round and 0 otherwise. Participants

[5] There are often multiple tournaments on a given weekend. There are about 25 weekends on which tournaments are held in any year. The Brown tournament was on the 9th such weekend in 2001.

were also offered a proof that expected mean squared error is minimized by submitting a probability equal to the subjective probability of winning.

Technically, their actual objective was not to minimize mean squared error but to maximize the probability that they were in the bottom five in mean squared error. This had the potential to cause respondents to guess "0" or "1" rather than their true probability estimates if they believed others would do the same, but the data do not show that this occurred.

The data are 923 debater-round combinations, of which 285 include estimates of win likelihoods. Each observation is defined by a debater ID and round number and includes data about the debater's prediction (if one was made), the round in particular (from the judge's ballot), and the debater herself. Table 1 presents descriptions and summary statistics.

Ballot information is Rank and Speaker Points for the debater, her partner and their opponents as well as a binary "outcome" variable which takes a value of one is the debater's team won and 0 otherwise. Personal information includes sex, and a school dummy that takes a value of 1 if the debater was a student at a school in the 2001 US News and World Reports Top 15 National Universities or Top 3 Liberal Arts Colleges and 0 otherwise.

There are also dummies accounting for experience. The Low Experience variable takes a value of 1 if the debater was a novice (i.e. in the first year of her APDA eligibility) and zero otherwise. The High Experience variable takes a value of 1 if the debater had reached the out-rounds at some point in her career and zero otherwise. As this was the ninth tournament of the year, no novice had yet reached the out-rounds in a prior tournament, so the two categories are mutually exclusive.[6] Those debaters who are neither Low nor High Experience are said to have Medium Experience.

[6] In fact, it is rare for novices to ever reach out-rounds.

At this point in the 2001-2002 school year, a low experience debater had participated in zero to forty rounds (depending on the number of prior tournaments she had attended), while a medium or high experience debater had participated in at least twenty-five.[7] Medium experience debaters, however, typically attend at least ten debate tournaments in their first year and often attend as many as twenty per year for all four years of college. High experience debaters are usually in their third or fourth year of college and have attended fifteen to twenty debates per year. Since there are five rounds per tournament, medium and high experience debaters have generally participated in 70-170 and 170-320 rounds, respectively, prior to this tournament.

Our estimation strategy requires an estimate of debater skill not correlated with outcome in a particular round. We therefore construct a skill variable which, for each debater-round observation, is that debater's average speaker points in the other rounds.

Inherent in this framework is the assumption that debaters are experienced at making the predictions asked of them. Most debaters spend a significant amount of time between rounds discussing how the round went with friends and a fundamental component of these discussions is a prediction of whether the debate was won.

Not all debaters chose to participate in the study. Some did not participate for any of the five rounds, while others stopped participating during the tournament. Table 3 shows summary statistics for the 484 observations corresponding to debaters who *never* participated, versus the 439 observations corresponding to debaters who participated in at least one round. There are some differences in the two populations, but they are unrelated to performance: differences in probability of winning and speaker points are minimal.

[7] Debaters are allowed to retain novice status if they debated at fewer than five tournaments in their first year.

4. Estimation Strategy

We can model the debater's prediction as $\mathbf{p = \varphi+\varepsilon}$ where $\mathbf{p}$ is the debater's prediction, $\boldsymbol{\varphi}$ is the true probability of having won and $\boldsymbol{\varepsilon}$ is an error. A debater is better at predicting outcomes when the first and second moments of $\boldsymbol{\varepsilon}$ are closer to 0. Because we do not observe $\boldsymbol{\varphi}$, we cannot directly estimate the influence of experience and other variables on predictive ability. We do observe the outcome $\mathbf{I}$, which is a noisy indicator of $\boldsymbol{\varphi}$, but substituting $\mathbf{I}$ for $\boldsymbol{\varphi}$ will not result in consistent estimates of accuracy. With observable data $\mathbf{X}$,

$$\mathbf{E((p\text{-}I)^2 \mid X) = E[(\varphi+\varepsilon)^2+\varphi-2\varphi(\varphi+\varepsilon) \mid X]}$$

$$\mathbf{= E(\varepsilon^2 \mid X)+ \varphi(X)\cdot(1-\varphi(X))}$$

Because $\boldsymbol{\varphi}$ is unobservable, the proposed method of substituting $\mathbf{I}$ for $\boldsymbol{\varphi}$ will not suffice. Instead we use the outcome data to generate estimates of $\boldsymbol{\varphi}$ and use these estimates instead of $\mathbf{I}$ as a substitute for $\boldsymbol{\varphi}$.

We first estimate $\boldsymbol{\varphi}$ using a probit regression of $\mathbf{I}$ on observables $\mathbf{X}$: $\mathbf{Pr(I=1|X) = \Phi(X'\beta)}$, where Φ is the normal CDF. This model yields an estimate of the debater's prediction error, which we then regress on dummies for experience as well as other observables, yielding the effect of experience on bias. We also estimate the effects of experience on accuracy by regressing the square of the debater's error, $\mathbf{p\text{-}\Phi(X'\beta)}$, adjusted using a process described below, on experience dummies and other variables.

Let our error in estimating the true win probability be $\mathbf{u = \Phi(X'\beta)\text{-}\varphi}$. Without adjustments, our measure of accuracy is not consistent:

$$\begin{aligned}\mathbf{E((p\text{-}\Phi(X'\beta))^2 \mid \theta)} &\mathbf{= E((\varepsilon\text{-}u)^2 \mid X)}\\ &\mathbf{= E(\varepsilon^2+u^2\text{-}2\varepsilon u \mid X)}\\ &\mathbf{= E(\varepsilon^2 \mid X)+E(u^2 \mid X)\text{-}2E(\varepsilon u \mid X)}\end{aligned}$$

Our prediction error $\mathbf{u}$ results from two factors: first, we do not observe a number of factors that are specific to each round, such a debater's comfort with a given topic or whether the judge had some preconceived bias. Second, the estimates of $\boldsymbol{\beta}$ that yield $\boldsymbol{\Phi(X'\beta)}$ are unlikely to be exactly correct. We can decompose our error as $\mathbf{u = v + z}$ where $\mathbf{v}$ arises from the first source of error and $\mathbf{z}$ from the second.

The debater error $\boldsymbol{\varepsilon}$ arises from overestimation of skill, misinterpreting judge or competitor actions during the debate, and so on. $\boldsymbol{\varepsilon}$ is therefore likely to be uncorrelated with $\mathbf{v}$. Because z results from estimation of a probit of outcome on X, $\boldsymbol{\varepsilon}$ is also uncorrelated with z. Therefore,

$$\mathbf{E(\varepsilon^2 \mid X)+E(u^2 \mid X)-2E(\varepsilon u \mid X) = E(\varepsilon^2 \mid X)+E(u^2 \mid X)}$$

The effect on debater error of any misestimation depends on $\boldsymbol{\Phi'(X'\beta)}$ and is therefore not independent of $\mathbf{X}$. We adjust the dependence of $\mathbf{E(u^2 \mid X)}$ on X by estimating a covariance matrix for our estimates of $\boldsymbol{\beta}$ and using Monte Carlo methods to generate the variance in our win probability estimates resulting from this type of error for each observation. These methods yield $\mathbf{E(z^2 \mid X)}$. Then:

$$\begin{aligned}\mathbf{E((p-\Phi(X'\beta))^2 \mid X)} &\mathbf{= E(\varepsilon^2 \mid \theta) + E(u^2 \mid X)}\\ &\mathbf{= E(\varepsilon^2 \mid X) + E(v^2 \mid X) + E(z^2 \mid X) + 2E(vz \mid X)}\\ &\mathbf{= E(\varepsilon^2 \mid X) + E(v^2) + E(z^2 \mid X)}\end{aligned}$$

We define our estimate of accuracy to be $\mathbf{(p-\Phi(X'\beta))^2- E(z^2 \mid X)}$ and regress this on $\mathbf{X}$. Our estimate of the constant term is inflated by $\mathbf{E(v^2)}$, which is unobservable, but otherwise our coefficient estimates are consistent.

One might argue that a more appropriate measure of predictive accuracy would be the variance: $\mathbf{E(\varepsilon^2 \mid X) - E(\varepsilon \mid X)^2}$. For reasons outlined in appendix B, these data do not allow us to estimate variance, nor is it clear that either measure of accuracy is better than the other.

There is one final difficulty in setting up our model: By a significant margin, the most important factor in allowing us to predict the probability of a win is the skill of the debater. True skill is unobserved, so we must infer it from our observations of debater speaker points, which are intended to measure the skill of a debater relative to the debating universe, not relative to others in the round. Unfortunately, positive surprises (from the debater's perspective) in speaker points will be positively correlated with surprises in outcome, since the team with higher combined speaker points wins. We therefore construct a measure of skill that does not include a round in question but which is constructed from observations at the tournament. We proxy for skill for a particular debater-round observation as the normalized average of debater speaker points over the rounds not including that observation. As long as expected speaker points are constant across rounds, this measure of skill avoids correlation with the error term while consistently measuring true skill.

5. Results

As discussed above, we begin by estimating each debater's probability of winning in a given round as a function of skill, experience level, sex, school, and an interaction of skill with a dummy variable for whether the debate took place in the first or second round. Throughout the analysis we use three specifications to estimate this probability, each relying on different subsets of the data. Specification I uses all 923 debater-round observations to predict outcome from observables. Specification II uses only the 439 observations associated with debaters who participated at some point and specification III uses only the 285 observations which include debater predictions. All three use a probit model, though results from a linear probability model and logit model (available on request) had, given a specification, nearly identical qualitative results. We do not believe that there is systematic nonparticipation or attrition, so we prefer specification I, which uses all of our available data.[8] The effect of experience on bias and accuracy does not differ qualitatively across the different specifications.

The marginal effects (evaluated at the mean) of these models are presented in Table 4, with the standard errors adjusted for clustering by debater. Because rounds one and two were paired randomly (with the requirement that "A" teams from each school cannot face each other) skill in those rounds has a large and statistically significant impact on win likelihood. In later rounds, when debaters are paired against others in the same win-loss bracket, skill is less important, though still significant in specifications I and II. In each specification, experience is related monotonically to winning.

To assess whether debaters learn to be less biased or more accurate, we create a variable, Error, which is the difference between the debater's own predicted probability and the predicted probability of winning from our probit model. We then regress Error on male, high experience, medium experience, and normalized skill, the normalized value of average speaker points as defined in Section 4. These results are presented in Table 5. In specification I, the coefficients on both experience dummies are small and insignificant. The point estimates of the impact of experience on bias vary according to the specification, but lack statistical significance in all three. On the other hand, skill is a large and statistically significant factor. A debater with skill one standard deviation above the mean has slightly more than half of the average prediction bias of a debater with mean skill.

We estimate the effects of experience on accuracy by regressing Error^2 on the same right hand side variables as before and present our results in Table 6. In this case, we find that experience does play a large and statistically significant role in accuracy, while the importance of skill is lessened considerably. The expected squared deviation

[8] Table 3 presents summary statistics for participants and non-participants.

for a highly experienced debater with mean skill is approximately one-half of that of a novice with the same skill.

The effect of skill is both small and insignificant; an increase of one standard deviation in skill reduces expected squared deviation by only 0.0164 (s.e. = 0.0149). The effect of gender, on the other hand, is large and significant: men tend to be less accurate. Specification III of our bias regression hints that men might be more biased, though the gender coefficient is insignificant in the other specifications, but in this model, all three specifications show that men are less accurate.

These results highlight that experience affects accuracy but not bias: people learn to be less extreme in their predictions but do not learn to be less biased. This supports the theory that overconfidence may be welfare enhancing: debaters' beliefs of their win likelihoods do converge over time, but to a biased level. It is clearly valuable for debaters to have a good understanding of how well they performed in a round: the better this understanding, the more they can change bad behaviors and improve good ones for the next tournament. On the other hand, confidence is also clearly important for winning. The two effects counteract each other when applied to bias: More bias implies more confidence but also less understanding of how to succeed in a debate. There is no such conflict in the second moment, and debaters do appear to learn to be more accurate.

6. Robustness and Sensitivity to Assumptions

There are several potential concerns we must address in order to be more sure of our results. The results below are available in Appendix A.

The monetary incentives offered for participation had two potential flaws. As mentioned above, the actual goal is not to minimize mean squared error, but rather to maximize the probability of being among the lowest five participants in mean squared error. If a large number of debaters participated in all five rounds, then the optimal strategy would be to guess zero or one depending on whether the subjective win estimate is less than or greater than 1/2. Two results suggest that this problem did not arise. First, only 15 debaters made predictions in all five rounds, so the probability of being in the bottom five with an error above 0 was high. Second, all 25 zero predictions are associated with losses. If debaters were biasing predictions to zero and one, then some participants with low probabilities of winning should have predicted zero yet won.

The second potential flaw of our reward system is that it provides more intense incentives to debaters who have win likelihoods close to zero or one. Their probability of achieving low mean squared errors is higher than that of more middling debaters. As noted in Table 3, winning percentage, average skill and the standard deviation of skill are similar across groups, so the difference in incentive intensity does not seem to have affected the decision to participate. We cannot know whether it affected the effort participants made in making predictions, but it seems unlikely that it would have affected effort but not the initial participation decision.

A further concern may be that those who dropped out during the tournament did so because of a poor prediction in a previous round. Examining the round prior to dropping out shows that the absolute value of difference between the predicted probability of winning and the debater's reported probability is 0.199, compared to 0.149 for all other rounds. The difference is insignificant, p = 0.291. Moreover, those who left the sample during the tournament did not differ significantly in predictive ability from those who never left the sample, with those who dropped at any point having a difference between predicted and reported probability of 0.160, and those who never dropped out having a value of 0.163. The difference is insignificant, p = 0.936.

Another potential problem is that skill is imperfectly measured. If judges have systematic biases in the way they assign speaker points that are not correlated with how they assign a winner, our results will be inaccurate. A simple example would be if judges assign higher speaker points to a team when their opponents perform better and lower points when their opponents perform worse. Table A-1 shows descriptive statistics for speaker points by round. We expect that rounds one and two have the largest skill differences, as debaters in those rounds are randomly paired (with the provision that no "A" teams may face each other). While round one appears to have slightly lower average speaker points and lower variance relative to the later rounds, round two does not have this character. Therefore, we are confident that skill measurement is not biased.

In order to evaluate the robustness of our model to changes in our measure of experience, we estimate a new model where experience is binary: Novice or Experienced, where the latter variable combines our Medium and High experience measures. We also allow the effect of skill to vary by experience level. The results, shown in Table A-2, indicate that bias is reduced by skill, as before, but not by experience. Estimating the effect of experience and skill on accuracy in a similar way, we find that experience is highly significant, while skill is not. Gender, too, is an important factor, with men being less accurate than women. These results, shown in Table A-3, also confirm our conclusions above.

7. Conclusions

The finding that experience does not affect bias, even after hundreds of observations by an individual, is in line with previous research on overconfidence bias, as are the findings that being more skilled reduces this overconfidence. The novel result in this study is that while debaters appear unable (or unwilling, consciously or sub-consciously) to learn through experience to reduce their overconfidence, they are still able to learn at a more abstract level: the accuracy of their predictions improves even while the bias is unchanging. These results are robust to a variety of changes in the specification of our win probability estimates and definitions of experience. This combination of results provides support for Compte and Postlewaite's claim that overconfidence bias may persist because it is welfare improving.

An obvious direction for future research would be to follow a panel of subjects over an extended period of time. In this manner, more accurate determinations of both the amount and rate of learning can be ascertained.

Table 1: Summary Statistics

Variable	Description	Mean	Standard Deviation
Outcome	1 if the debater won that round, 0 otherwise	0.488	0.501
Debater prediction	Reported prediction of winning that round	0.650	0.351
Economist prediction I	Estimated probability of winning as computed in Table 4, Column I	0.488	0.216
Economist prediction II	Estimated probability of winning as computed in Table 4, Column II	0.489	0.234
Economist prediction III	Estimated probability of winning as computed in Table 4, Column III	0.487	0.226
Error I	Difference between debater's reported probability of winning and economist's predicted probability as computed in Table 4, Column I	0.161	0.342
Error II	Difference between debater's reported probability of winning and economist's predicted probability as computed in Table 4, Column II	0.160	0.348
Error III	Difference between debater's reported probability of winning and economist's predicted probability as computed in Table 4, Column III	0.162	0.348
Adjusted Error^2 I	Square of error, adjusted as described in Section 4 using predictions calculated from Table 4, Column I	0.143	0.153
Adjusted Error^2 II	Square of error, adjusted as described in Section 4 using predictions calculated from Table 4, Column II	0.146	0.164
Adjusted Error^2 III	Square of error, adjusted as described in Section 4 using predictions calculated from Table 4, Column III	0.147	0.162
Skill	Mean of Speaker Points in all other rounds	24.92	1.04
Low experience	1 if the debater is of low experience, 0 otherwise	0.554	0.498
Medium experience	1 if the debater is of medium experience, 0 otherwise	0.291	0.455
High experience	1 if the debater is of high experience, 0 otherwise	0.154	0.362
Male	1 if the debater is a male, 0 otherwise	0.653	0.477
Top school	1 if the debater attends a school ranked by U.S. News and World Reports in 2001 as being in the top 15 national universities or top 3 liberal arts colleges, 0 otherwise	0.491	0.501

Based on 285 observations for 95 debaters.

Table 2: Speaker Points and Winning Percentage by Experience Level

Experience Level	Number of Observations	Speaker Points	Winning Percentage
Low	158	24.42 (1.072)	0.354 (0.480)
Medium	83	25.09 (0.919)	0.566 (0.499)
High	44	26.27 (1.065)	0.818 (0.390)

Based on 285 observations for 95 debaters. Standard deviations are reported in parentheses.

Table 3: Participation

	Never participated	Participated in at least one round
Number of observations	484	439
Outcome	0.494 (0.500)	0.508 (0.501)
Skill	24.91 (1.24)	24.95 (1.24)
Low experience	0.455 (0.498)	0.501 (0.501)
Medium experience	0.411 (0.493)**	0.317 (0.466)**
High experience	0.134 (0.341)**	0.182 (0.386)**
Male	0.599 (0.491)**	0.674 (0.469)**
Top school	0.581 (0.494)**	0.501 (0.501)**

Summary statistics for 484 outcomes, representing 108 debaters, who never participated in the study, versus 439 outcomes, representing 95 debaters, who participated in at least one round. Standard deviations are reported in parentheses. ** indicates differences that are statistically significant at the 5% level.

Table 4: Estimated Probability of Winning

	I	II	III
High experience	0.1519 (0.0665)**	0.2290 (0.0909)**	0.2281 (0.1187)*
Medium experience	0.0635 (0.0394)	0.1516 (0.0587)**	0.1073 (0.0742)
Normalized skill	0.1124 (0.0296)***	0.0848 (0.0435)*	0.0699 (0.0555)
Round*normalized skill	0.1695 (0.0420)***	0.2187 (0.0649)***	0.2334 (0.0759)***
Male	0.0092 (0.0380)	-0.0160 (0.0547)	-0.0931 (0.0669)
Top school	0.0190 (0.0378)	0.0335 (0.0533)	0.0544 (0.0664)

Coefficients are marginal effects, evaluated at the mean, of a probit regression of Outcome on observables under three treatments. Treatment I uses all 923 debater-round observations; treatment II uses only the 439 observations associated with debaters who participated in the study; treatment III uses only the 285 observations in which debaters submitted predictions. Standard errors are reported in parentheses and adjusted for clustering within debaters. *, ** and *** signify significance at the 10%, 5% and 1% level respectively.

Table 5: Bias

	I	II	III
High experience	-0.0081 (0.0874)	-0.0816 (0.0887)	-0.0755 (0.0883)
Medium experience	0.0189 (0.0488)	-0.0676 (0.0491)	-0.0248 (0.0491)
Normalized skill	-0.0810 (0.0267)***	-0.0690 (0.0270)**	-0.0644 (0.0267)**
Male	0.0087 (0.0398)	0.0317 (0.0398)	0.0991 (0.0395)**
Top school	-0.0468 (0.0420)	-0.0591 (0.0420)	-0.0777 (0.0420)*
Constant	0.1736 (0.0404)***	0.2006 (0.0407)***	0.1545 (0.0405)***

Coefficients are reported from OLS regression of Error on observables under three treatments. Treatment I uses all 923 debater-round observations; treatment II uses only the 439 observations associated with debaters who participated in the study; treatment III uses only the 285 observations in which debaters submitted prediction. Standard errors are reported in parentheses and adjusted for clustering within debaters. ** and *** signify significance at the 10%, 5% and 1% level respectively.

Table 6: Accuracy

	I	II	III
High experience	-0.0694 (0.0368)*	-0.0883 (0.0378)**	-0.0955 (0.0371)**
Medium experience	-0.0487 (0.0217)**	-0.0752 (0.0212)***	-0.0624 (0.0221)***
Normalized skill	-0.0164 (0.0149)	-0.0156 (0.0150)	-0.0133 (0.0150)
Male	0.0541 (0.0209)**	0.0634 (0.0214)***	0.0877 (0.0205)***
Top school	-0.00233 (0.0219)	-0.0116 (0.0228)	-0.0233 (0.0223)
Constant	0.132 (0.0193)***	0.143 (0.0199)***	0.129 (0.0194)***

Coefficients are reported from OLS regression of adjusted squared error on observables under three treatments. Treatment I uses all 923 debater-round observations; treatment II uses only the 439 observations associated with debaters who participated in the study; treatment III uses only the 285 observations in which debaters submitted prediction. Standard errors are reported in parentheses and adjusted for clustering within debaters. ** and *** signify significance at the 10%, 5% and 1% level respectively.

References

1) Compte, O. and Postlewaite, A. (2004). “Confidence-Enhanced Performance,” American Economic Review, 94, 1536-1557

2) Ehrlinger, J. and Dunning, D. (2003). “How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance,” Journal of Personality and Social Psychology, 84, 5-17

3) Harris, M. M. and Schaubroeck, J. (1988). “A Meta-Analysis of Self-Supervisor, Self-Peer and Peer-Supervisor Ratings,” Personnel Psychology, 41, 43-62

4) Kruger, J. M. and Dunning, D. (1999). “Unskilled and Unaware of it: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments,” Journal of Personality and Social Psychology, 77, 1121-1134

5) Mabe, P. A. III and West, S. G. (1982). “Validity of Self-Evaluation of Ability: A Review and Mets-Analysis,” Journal of Applied Psychology, 67, 280-296

6) Malmendier, U. and Tate, G. (2005). “CEO Overconfidence and Corporate Investment,” Journal of Finance, 60, 2661-2700

7) Malmendier, U. and Tate, G. (2007). “Who Makes Acquisitions? CEO Overconfidence and the Market’s Reaction,” NBER working paper 10813

8) Marteau, T. M., Johnston, M., Wynne, G. and Evans, T. R. (1989). “Cognitive Factors in the Explanation of the Mismatch Between Confidence and Competence in Performing Basic Life Support,” Psychology and Health, 3, 173-182

9) Massey, C. and Thaler, R. (2005). “Overconfidence vs. Market Efficiency in the National Football League,” NBER working paper 11270

10) Niederle, M. and Vesterlund, L. (2005). “Do Women Shy away from Competition? Do Men Compete too Much?,” NBER working paper 11474

11) Santos-Pinto, L. and Sobel, J. (2005). “A Model of Positive Self-Image in Subjective Assessments,” American Economic Review, 95, 1386-1402

12) Sporer, S. L., Penrod, S., Reed, D. and Cutler, B. (1995). “Choosing, Confidence and Accuracy: A Meta-Analysis of the Confidence—Accuracy Relation in Eyewitness Identification Studies,” Psychological Bulletin, 118, 315-327

13) Tracey, J. M., Arroll, B., Richmond, D. E. and Barham, P. M. (1997). “The Validity of General Practitioners Self-Assessment of Knowledge,” British Medical Journal, 315, 1426-1428

Appendix A

Table A-1: Speaker Points by Round

	Number of Observations	Mean	Standard Deviation	Min	Max
Round 1	192	24.73	1.08	21	28
Round 2	195	24.93	1.23	22	28
Round 3	177	25.03	1.36	20	28
Round 4	186	24.94	1.27	22	28
Round 5	173	25.03	1.25	21	28

Mean, standard deviation, minimum and maximum speaker points assigned by round, from 923 debater-round observations.

Table A-2: Bias With Binary Experience

	I	II	III	IV	V	VI
Experience	0.0167 (0.0495)	-0.0684 (0.0502)	-0.0320 (0.0499)	0.0128 (0.0463)	-0.0727 (0.0465)	-0.0359 (0.0467)
Normalized skill	-0.0868 (0.0222)***	-0.0731 (0.0225)***	-0.0744 (0.0222)***	-0.1202 (0.0324)***	-0.1092 (0.0328)***	-0.1064 (0.0325)***
Normalized skill*experienced	-	-	-	0.0693 (0.0433)	0.0750 (0.0436)*	0.0665 (0.0433)
Male	0.0063 (0.0391)	0.0300 (0.0393)	0.0962 (0.0390)**	-0.0047 (0.0413)	0.0180 (0.0413)	0.0856 (0.0410)**
Top school	-0.0457 (0.0423)	-0.0593 (0.0425)	-0.0731 (0.0424)*	-0.0509 (0.0428)	-0.0649 (0.0429)	-0.0781 (0.0428)*
Constant	0.1710 (0.0405)***	0.1985 (0.0408)***	0.1488 (0.0406)***	0.1630 (0.0396)***	0.1898 (0.0399)***	0.1411 (0.0397)***

Coefficients are reported from OLS regression of Error on observables under six treatments. Treatments I&IV use all 923 debater-round observations; treatments II&V use only the 439 observations associated with debaters who participated in the study; treatments III&VI use only the 285 observations in which debaters submitted prediction. Treatments IV-VI allow the effect of skill to vary by experience. Standard errors are reported in parentheses and adjusted for clustering within debaters. *, ** and *** signify significance at the 10%, 5% and 1% level respectively.

Table A-3: Accuracy With Binary Experience

	I	II	III	IV	V	VI
Experienced	-0.0513 (0.0219)**	-0.0769 (0.0216)***	-0.0665 (0.0223)***	-0.0495 (0.0216)**	-0.0755 (0.0213)***	-0.0646 (0.0222)***
Normalized skill	-0.0202 (0.0123)	-0.0181 (0.0121)	-0.0194 (0.0123)	-0.00512 (0.0188)	-0.00687 (0.0191)	-0.00346 (0.0188)
Normalized skill*experienced	-	-	-	-0.0315 (0.0229)	-0.0233 (0.0226)	-0.0332 (0.0226)
Male	0.0524 (0.0204)**	0.0623 (0.0210)***	0.0851 (0.0215)***	0.0574 (0.0215)***	0.0660 (0.0221)***	0.0903 (0.0210)***
Top school	-0.0006 (0.0213)	-0.0106 (0.0221)	-0.0204 (0.0215)	0.00180 (0.0213)	-0.00887 (0.0221)	-0.0179 (0.0214)
Constant	0.130 (0.0190)***	0.142 (0.0199)***	0.126 (0.0189)***	0.134 (0.0198)***	0.145 (0.0202)***	0.130 (0.0196)***

Coefficients are reported from OLS regression of adjusted squared error on observables under six treatments. Treatments I&IV use all 923 debater-round observations; treatments II&V use only the 439 observations associated with debaters who participated in the study; treatments III&VI use only the 285 observations in which debaters submitted prediction. Treatments IV-VI allow the effect of skill to vary by experience. Standard errors are reported in parentheses and adjusted for clustering within debaters. *, ** and *** signify significance at the 10%, 5% and 1% level respectively.

Table A-4: Prediction Variance

	I	II	III
High experience	-0.0532 (0.0462)	-0.0549 (0.0482)	-0.0538 (0.0486)
Medium experience	-0.0366 (0.0220)*	-0.0382 (0.0229)*	-0.0374 (0.0230)
Normalized skill	0.0020 (0.0139)	0.0027 (0.0144)	0.0024 (0.0145)
Male	0.0416 (0.0182)**	0.0421 (0.0187)**	0.0428 (0.0187)**
Top school	0.0193 (0.0185)	0.0192 (0.0190)	0.0191 (0.0192)
Constant	0.0907 (0.0174)***	0.0924 (0.0180)***	0.0919 (0.0180)***

Coefficients are reported from OLS regression of the squared error minus expected error on observables under three treatments. Treatment I uses all 923 debater-round observations; treatment II uses only the 439 observations associated with debaters who participated in the study; treatment III uses only the 285 observations in which debaters submitted prediction. Standard errors are reported in parentheses and adjusted for clustering within debaters. *, ** and *** signify significance at the 10%, 5% and 1% level respectively.

Appendix B

Our measures of bias and accuracy are the calculated expectations of debater error and debater error squared. While the former is clearly the best measure of bias, the latter is a combination of prediction bias and prediction variance. That is, we would prefer to estimate the impact of experience on prediction variance, not on the expectation of prediction error squared. Ideally, we could use the regression reported in Table 5 to get an estimate of expected bias for each debater, $\mathbf{B = X'\gamma}$. We then can get an estimate of prediction variance as $\mathbf{(p - \Phi(X'\beta) - B)^2}$ which is the debater's prediction error, normalized to have mean zero, squared.

In words, normalized prediction error is "debater prediction minus our prediction minus expected debater error." Expected debater error, however, is expected debater prediction minus our prediction. Therefore the first difference becomes "debater prediction minus expected debater prediction." Any connection with actual win probabilities is removed via the normalization because our estimates of debater error and win probability come from the same data. Therefore, this method will illustrate what types of debater have high variance predictions relative to expected prediction but not relative to true win likelihood. We find that, relative to the model's prediction of what the debater's prediction should be, more experienced debaters do appear to converge to a lower prediction variance. While the point estimates are not significant, they are monotonic in experience. These results are available in Table A-4.