**DEPARTMENT OF ECONOMICS
COLLEGE OF BUSINESS AND ECONOMICS
UNIVERSITY OF CANTERBURY
CHRISTCHURCH, NEW ZEALAND**

**OVERCONFIDENCE IN FORECASTS OF OWN
PERFORMANCE: AN EXPERIMENTAL STUDY**

by Jeremy Clark and Lana Friesen

WORKING PAPER

No. 09/2006

**Department of Economics, College of Business and
Economics, University of Canterbury, Private Bag 4800,
Christchurch, New Zealand**

WORKING PAPER No. 09/2006

OVERCONFIDENCE IN FORECASTS OF OWN PERFORMANCE: AN EXPERIMENTAL STUDY

by Jeremy Clark¹ and Lana Friesen²

November 16th, 2005

ABSTRACT. Overconfidence can have important economic consequences, but has received little direct testing within the discipline. We test for overconfidence in forecasts of own absolute or relative performance in two unfamiliar experimental tasks. Given their choice of effort at the tasks, participants have incentives to forecast accurately, and have opportunities for feedback, learning and revision. Forecast accuracy is evaluated at both the aggregate level, and at the individual level using realized outcomes. We find very limited evidence of overconfidence, with zero mean error or under-confidence more prevalent. Under-confidence is greatest in tasks with absolute rather than relative win criteria, often among subjects using greater or “smarter” effort.

Keywords: overconfidence, forecast errors, self-assessment

JEL Classifications: C91 D83 D84 J24

¹ Department of Economics, College of Business & Economics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand,
jeremy.clark@canterbury.ac.nz. Phone 011 643 364 2308 Fax: 011 643 364 2635

² Commerce Division, Lincoln University, PO Box 84, Canterbury, New Zealand,
friesenl@lincoln.ac.nz. Phone: 011 643 325 3627 Fax: 011 643 325 3847

1. Introduction

“The over-weening conceit which the greater part of men have of their own abilities, is an ancient evil remarked by the philosophers and moralists of all ages.”

(Adam Smith in *The Wealth of Nations*, 1776)

Social commentators of many generations have recognised the human tendency to be overconfident. Modern social psychology has confirmed these observations, with the pervasiveness of overconfidence one of the most robust and undisputed findings in the field. The most well known manifestation of overconfidence is the “above-average” effect, where most people report they expect an above-average likelihood of high starting salaries and job satisfaction, staying married, having gifted children, and other positive life events (Weinstein, 1980; Baker and Emery, 1993). Conversely, most people report a below-average risk of car accidents (Svenson, 1981), job loss or unemployment (Weinstein, 1980), and health problems (Weinstein, 1982, 1984, 1987, 1989, 1998; Weinstein and Klein, 1995; Miller et al., 1990). People are also systematically overconfident about the precision of their knowledge (Lichtenstein et al., 1982), and prone to self-attribution bias, over-attributing success to their own skill but blaming chance for failure (Taylor and Brown, 1988).

Few appear exempt from overconfidence, which apparently afflicts students and the general population alike (Weinstein, 1987). Attempts at verbal warnings and “de-biasing” manipulations have been largely unsuccessful (Weinstein and Klein, 1995). The only group who seem consistently realistic are the clinically depressed (Alloy and Ahrens, 1987; Pyszczynski and Holt, 1987). However, the degree of overconfidence is variable and tends to be more prevalent for events over which

respondents perceive some measure of control, or lack personal experience, or judge to be infrequent (Weinstein, 1980, 1982, 1984, 1987).

Overconfidence can have significant economic consequences because people rely on self-assessments when choosing career paths, investing in higher education, starting businesses, committing to marriage, and so on. Even Adam Smith (1776, p. 211) recognized that the “presumptuous hope of success” would result in too much effort exerted in the pursuit of high-risk and return careers.¹ More recently, overconfidence about one’s relative ability has been suggested as an explanation for the high rate of new business failures (Camerer and Lovallo, 1999), and for excessive job market search and unemployment (Dubra, 2004). In addition, if parties overestimate their chance of success at binding arbitration, expensive bargaining impasses are more likely to occur (Babcock and Loewenstein 1997; Dickinson, 2005). Conversely, underestimating one’s risk of negative life events may result in failure to undertake appropriate risk-reducing activities and expenditures (Weinstein, 1982).

In financial markets, overconfidence in one’s own knowledge and ability can lead to excessive trading and lower returns (Barber and Odeon, 2001), distortions in corporate investment decisions (Malmendier and Tate, 2005), value-reducing mergers (Roll, 1986), and various security market anomalies (Daniel et al., 1998). Some authors even suggest that overconfidence may render capital markets inefficient (Manove and Padilla, 1999).

On the other hand, the fact that overconfidence seems conspicuously absent only among the clinically depressed (Alloy and Ahrens, 1987; Pyszczynski et al., 1987) has led some psychologists to suggest that overconfidence could be beneficial, serving to “foster motivation, persistence at tasks, and ultimately more effective performance” (Taylor and Brown, 1988, p. 199). Two recent papers that model the

¹ Frank and Cook (1995) describe these as “winner-take-all” careers.

benefits of overconfidence are Benabou and Tirole (2002), where overconfidence can help overcome imperfect willpower, and Compte and Postlewaite (2004) where overconfidence increases the chance of success in risky activities.

While economists are increasingly recognising the potential consequences of overconfidence, little direct testing has been undertaken. Most of the studies reported above either assume overconfidence exists and focus on its consequences, or test for it only indirectly. Indeed, testing for the presence of overconfidence is inherently difficult. In particular, how should the accuracy of individual expectations be evaluated? A person may quite rightly expect to be above average in terms of career success or below average in terms of heart attack risk, for example. In addition, over time a person's actions may change their risk exposure, rendering any initial assessment incorrect.²

The approach taken in psychology has been to test for relative forecast bias in the aggregate, thereby avoiding the need for individual benchmarks. This approach was pioneered by Weinstein (1980) who asked individuals to rate their chance of experiencing a pleasant or unpleasant outcome relative to their peers using a seven point scale centred at zero, with -3 for “much below average” and $+3$ for “much above average”. Evaluated at a group level it is not possible for everyone to be either above- or below average. This methodology has resulted in the robust “above-average” effect described earlier.

In contrast, the economists who have investigated forecast bias have tended to elicit individuals' quantitative probability estimates of future outcomes, such as income or longevity. These predictions are then compared with either current aggregate realizations (Hamermesh, 1985; Fischhoff et al., 2000; Hurd and Rohwedder,

² Eliciting expectations is also problematic because people have difficulty conceptualising probabilities, and responses can depend on the method used. We address this issue in Section 2.

2003) or, where longitudinal data is available, with actual individual realizations (Das and van Soest, 1997, 1999; Dominitz, 1998; Smith, 2001). The findings on overconfidence using these methods have been much more equivocal. For example, Hamermesh (1985) found that while respondents were pessimistic about survival to age 60, they were optimistic about surviving to age 80. Teenagers were found by Fischhoff et al. (2000) to be overconfident about future career success, but also have a greatly exaggerated belief of early death. Using longitudinal data, Dominitz (1998) found that Americans were too optimistic about changes in future income, while Das and van Soest (1997, 1999) found that Dutch households generally underestimated future income changes.

We take a different approach in this paper and use an incentive-based experiment to directly measure people's forecasts of their own absolute or relative performance in two initially unfamiliar computerized tasks. One task involves searching over a spreadsheet for the maximum value of a function, the other decoding five-letter words. Subjects are rewarded *both* for performance on the tasks and accuracy of their predictions. Because we use individual realized outcomes for a benchmark, we are able to test for forecast bias in both relative *and* absolute settings, and compare our results in the relative case with Weinstein's hypothetical group level evaluations.

We also ensure that subjects receive feedback about forecast error, having them play two practice rounds, then two sets of ten rounds of each task. They revise their forecasts for the second set of a task after playing the first. Finally, by collecting descriptive data on subjects and passively measuring their keyboard inputs for each task, we can also test whether variation in forecast error is predicted by subjects' characteristics, whether overconfidence does lead to better outcomes via better effort, and whether subjects learn to be better forecasters.

We find that overconfidence is not as robust as the psychology literature would suggest, with zero mean error or under-confidence more prevalent. We also find no evidence that *ex ante* overconfidence confers any advantage in subsequent outcomes either directly or indirectly via effort quantity or quality. Regarding learning, individuals do update forecasts in the direction of initial outcomes, though insufficiently to eliminate subsequent forecast errors as ability improves. Finally, we find intriguing evidence that people have less difficulty predicting their success relative to others than against an absolute performance threshold.

The remainder of the paper runs as follows. Section 2 describes our experimental design. Section 3 presents our results, and Section 4 a brief discussion and conclusion.

2. Experimental Design: Mountain Climbing and Espionage

Taking our cues from the psychology literature on overconfidence, we chose a design where subjects would have to make predictions about tasks over which they initially lacked personal experience, yet perceived some measure of control. In order to reflect economic contexts where overconfidence may play a role, we also needed tasks that required real yet quantifiable effort, diverse skills, and that provided opportunities for learning. We settled on two. The first task was to maximize an unknown function by searching across contiguous cells of a computer spreadsheet. The underlying function generating the values on the spreadsheet was:

$$f(x, y) = \alpha \{b_3 - [(x - b_1)^2 + (y - b_2)^2 + (x - b_1)(y - b_2)]\} \quad (1)$$

where x and y are the coordinates of a cell, and α , b_1 , b_2 , and b_3 are parameters. This task is a simplified version of the two-variable optimisation problem used by van Dijk, Sonnemans, and Winden (2001) to examine the effect of different incentive

schemes on work effort. Within each round the unknown function (1) took on the same smooth, paraboloid shape, with a maximum value (or "height") of 100,400. The most direct route to the peak always required 200 moves. With 4 contiguous moves allowed per second, 240 moves were possible within a 60 second round. Between rounds, the location of the function's maximum was shifted randomly around the circumference of a circle centred at the spreadsheet's origin. Thus, the degree of difficulty remained the same in each round. A reproduction of the spreadsheet seen by subjects is given in Figure 1.

The second task we chose was decoding five letter words, drawn from a random sample of 1155 such words taken from the 1993 *Shorter Oxford English Dictionary*.³ The coding scheme for each round was comprised of a unique letter-for-letter mapping, and was randomly reset between rounds. With 4 letter guesses allowed per second, 240 letter guesses were possible within a 60 second round. A reproduction of the decoding task is given in Figure 2.

Both the maximization and verbal decoding tasks were fully computerized.⁴ Each participant read computerized instructions for one task, tried it for 22 rounds, read instructions for the other task, and tried it for 22 rounds. These 22 rounds were comprised of two practice rounds, a first set of ten, and a second set of ten. Thus, each participant completed two sets of one task, then two sets of the other. In half of sessions, the maximization task came first, and in half the verbal decoding task came first.

Forecasts of own performance in a task were first elicited after the two practice rounds. Subjects were asked how many of the first and second sets of ten rounds they thought they would win. After completing the first ten rounds, subjects

³ Proper names and obscure words were avoided.

⁴ The experimental instructions are included for reviewers in Supplementary Appendix A.

were reminded of their initial forecast for the *second* ten rounds, and asked to re-enter a prediction, whether revised or not. This sequence was repeated for both tasks. Subjects were given immediate feedback on whether they had won or lost each round, and were told of their earnings accumulation from prediction and performance after each set. Figure 3 illustrates the sequence of decisions subjects made for a given task.

2.1 Criteria for Winning

We looked for evidence of overconfidence in both relative and absolute settings. Twelve subjects participated in each session. In the relative treatments, the best five of the twelve participants won each round.⁵ For the maximization task, the five of twelve participants who had reached the highest spreadsheet value by the end of a round won, while for the verbal task, the five of twelve participants who had decoded the greatest number of letters by the end of the round won. Ties were automatically broken at random, so that there were always exactly five winners each round, yielding an average win rate of 4.17 in each set of ten rounds.

In the absolute treatments, subjects had to meet an absolute standard to win each round. For the maximization task, the standard was to reach a value of 100,000 on the spreadsheet within 60 seconds. A cell's value on the spreadsheet appeared only when the subject placed the cursor on it. For the verbal task, the standard was decoding two five-letter words within 60 seconds. Subjects were free to continue working at a task for the full 60 seconds of a round, even after meeting the standard. In analysing effort measures, however, we do not use non-incentive "post-win" data.

⁵ The tournament design literature suggests that effort is promoted best and collusive laziness avoided with larger group sizes (eight rather than four or two), and intermediate rather than extreme win rates (Orrison et al., 1998; Harbring and Irlenbusch, 2003).

2.2 Eliciting Forecasts

Exactly *how* beliefs should be elicited has raised disagreements within Economics and Psychology. At one end, economists such as Dominitz and Manski (1997) argue that with suitable preparation, individuals can meaningfully provide entire probability distributions to forecast variables such as income. Eliciting full distributions has the advantage of enabling researchers to test for overconfidence both in terms of mean *and* dispersion, as people may form subjective distributions that are too tightly concentrated around a (biased) mean. In contrast, psychologists such as Weinstein (1998) argue that the public has difficulty generating even point estimates of the mean of a subjective distribution. Rather, ordinal measures such as "very likely" and "somewhat likely" should be used.

Researchers have commonly taken an intermediate approach of eliciting point estimates as numbers between one and 100, or simply one and ten (Viscusi, 1990; Fischhoff et al., 2000; Morrison and Rutstrom, 2000; Jamison and Karlan, 2003). We also adopt an intermediate approach, asking subjects to report the number of rounds out of ten that they expect to win in each set. We believe that this question is framed more naturally than asking about the probability of winning a particular round, and as such is likely to yield responses that are more meaningful. However, by eliciting only a point estimate, we can test only for overconfidence as a bias in the mean of subjects' subjective distributions, rather than in the dispersion.

2.3 Rewarding Forecast Accuracy and Moral Hazard

Subjects were rewarded for *both* their performance on the tasks, and the accuracy of their predictions of that performance. For performance in a set, subjects earned two points for each round won. For accuracy in forecasting performance in the set, subjects earned points according to a quadratic scoring rule:

$$5 - \frac{(P - W)^2}{20} \quad (2)$$

where P and W refer respectively to the predicted and realized number of rounds won in the set.⁶ Accuracy was evaluated for the first set of a task, then for the second set based on the revised predictions made after the first set.

If subjects were risk neutral and making predictions about events they could not privately manipulate (a kind of moral hazard), reporting the expected value of their subjective probability distribution would maximize earnings. However in these tasks subjects have partial control over the number of wins they are forecasting through their (imperfectly observed) choice of effort during a set of rounds. Subjects could purposefully reduce the quality or quantity of their effort in a task in order to lose rounds in order to make performance line up with low initial forecasts.⁷ We address this problem in two ways.

First, to reduce the incentive that the scoring rule creates for subjects to alter their effort in the tasks, we set the marginal reward from winning an additional round (2 points) to exceed any possible marginal loss from increasing forecast error by an additional round (ranging from 0.05 points to 0.95 points).⁸ Over all four sets of a session subjects could accumulate a maximum of 20 points from accurate prediction, versus 80 points from actual performance.

Second, because this reward structure does not *eliminate* moral hazard incentives, as the internal cost to subjects of exerting effort in the tasks is not

⁶ Quadratic scoring rules have commonly been used in experiments to elicit subjects' beliefs about the choices of other subjects or their types (Sonnemans, Schram and Offerman (1998), Morrison and Rutstrom (2000), Huck and Weizsacker (2002), Nyarko and Schotter (2002)), or to elicit probabilities in the context of Bayesian inference (McKelvey and Page (1990), Grether (1992)).

⁷ A similar, though less substantial problem, has arisen in asset market experiments where subjects have some (limited) impact on the prices they are forecasting. Rewarding accuracy had no impact on actual prices in these markets (see e.g. Williams (1987)). In a different context, Offerman and Sonnemans (2001) show that incentives had no impact on effort exerted in a prediction task.

⁸ The maximum penalty from increased forecast error occurs when the error is increased from 9 to 10 rounds, costing an additional 0.95 points. The minimum occurs when the error is increased from 0 to 1 rounds, costing only 0.05 extra points.

controlled, we also check for moral hazard in practice by running four additional relative tournament sessions with non-incentive forecasts.⁹ In these sessions subjects were asked simply to try their best to predict how many rounds they would win, and accumulated points based only on performance. Since participants in these sessions could accumulate fewer points toward the final draw, an un-announced participation fee of \$4 was added to their final earnings.

To induce risk neutrality in theory, (though less clearly in practice (Selten et al., 1999)), we rewarded subjects for accurate prediction and performance with points that they accumulate for use in a single draw at the end of the session (as in Berg et al. (1986)). Each subject entered a final private draw between an NZ\$8 prize and a NZ\$30 prize, with the probability of winning \$30 given by the number of points accumulated from performance and accurate prediction over the session. Points accumulated could vary between 0 and 100.

3. Experimental Results

The fully computerized experiment was conducted at the University of Canterbury.¹⁰ In total, 239 student subjects participated in 20 different sessions of the experiment as summarized in Table 1. Students were recruited from large first and second year courses in economics, statistics, and mathematics. Each session took approximately one hour and thirty minutes, and earnings ranged from NZ\$10 to \$42, with an average of \$23.97.¹¹

Table 2 provides a summary of our main results, with more detailed descriptive statistics provided in Supplementary Appendix B. For reasons that will be

⁹ We believed the incentive to purposefully lose rounds would be strongest in relative tournaments, where after the first set had been completed subjects would have a good idea of their ability relative to other participants.

¹⁰ The data presented here was part of a larger experiment collecting additional data.

¹¹ The adult hourly minimum wage in New Zealand at the time of the experiment was NZ\$8.50.

described in Section 3.2, we pool results for sessions with and without a scoring rule. Whether from experience spillovers or fatigue, results sometimes differed depending on the order in which tasks were experienced, and so we report the results in a disaggregated form.¹²

The number of wins in Set I of a task tended to be lower in the absolute than in the relative treatments. However this was compensated by a larger number of absolute wins in Set II, so that overall there was no significant difference in the total win rate across the two winning criteria. Subjects anticipated this potential for learning in the absolute treatments, and initially predicted winning on average approximately one extra round in Set II. This effect is absent in the relative treatments where the initial predictions for the two sets are indistinguishable.

3.1 Little Overconfidence

We move now to our principal investigation of overconfidence in predicting own performance. Our design enables us to test for overconfidence using the *ex ante* group approach of Weinstein, or an individual bias approach using realized outcomes. Group bias is testable only in the relative sessions, by asking whether the predicted number of wins per set, averaged across all subjects, differed significantly from the promised average win rate of 4.17. The last two columns of Table 2 provide the results.

We find limited evidence of group overconfidence bias, but only for the maximization task, and only when it was the first task experienced. In that tournament and order, t-tests indicate that the average predicted number of wins

¹² In relative tournaments, Mann-Whitney tests detected significant order effects for predictions and forecast error in the maximization task, though not the decoding task. In absolute contests, order effects were found for revised forecast errors in the maximization task, and in some measures of effort quality and quantity in both tasks.

(4.76) over 72 participants in Set I was significantly higher than the (known) actual number of wins (4.17). This initial overconfidence persisted even in the revised Set II prediction (4.71), made after the group experienced the ten rounds of Set I. In contrast, when the maximization task was experienced second, subjects were initially under-confident, predicting a mean win rate of only 3.61, before revising to no significant difference from 4.17 in Set II forecasts. Mean predicted wins for the verbal decoding task were generally indistinguishable from 4.17, the only deviation being initial Set I under-confidence (3.61) when the task was experienced first.

Moving to realized outcome analysis, we compare the predictions of each individual to his or her outcome. Here we test for mean forecast errors that differ significantly from zero, and can do so in tasks with both relative *and* absolute criteria for success. First, we define forecast error, $ERROR_i$ as the difference between the number of rounds subject i predicted she was going to win, and the number she did win. $ERROR_i$ may be positive, zero, or negative, indicating *ex post* that the subject's forecast turned out to be overconfident, accurate, or under-confident, respectively. We look for systematic forecast errors by testing whether the sample mean of $ERROR_i$ is significantly different from zero, whether positive or negative. The mean error rates are also reported in Table 2.

Beginning with the relative tournaments, the results are consistent with those found using group bias analysis. We find robust overconfidence in maximization when it was experienced first, under-confidence giving way to zero mean bias when it was experienced second, and zero mean error for verbal decoding. The initial under-confidence previously found in the verbal tournament when experienced first is no longer significant.

Scatter plots of individual revised Set II predictions and outcomes for the relative tournaments are shown in Figure 4. Points above the 45-degree line indicate

under-confidence, while points below reflect overconfidence. The figure shows that where subjects were accurate on average (3 of 4 cases) this resulted from an even split between those who were over- and under-confident rather than from widespread individual accuracy. Overconfidence in the final case (maximization when first) resulted from a majority having positive forecast errors.

Moving to the accuracy of forecasts in absolute contests, we find a greater deviation from zero mean error, but not in the direction of overconfidence. For the maximization task when it comes first, we find no significant mean forecast error for either set. When maximization comes second, subjects begin with positive mean error in Set I and then overcorrect to negative mean error for the revised forecast for Set II. In fact, the initial forecasts for Set II that would have turned out to be accurate were revised to be excessively negative.

Even more strikingly, subjects showed persistent negative mean forecast error in the verbal decoding contest regardless of order. Revised Set II predictions reduced the error over initial Set II predictions, but not by nearly enough to eliminate it. In fact, the revised predictions for Set II were no more accurate than the initial predictions for Set I, as subjects failed to anticipate continuing improvement even after 12 rounds of experience! On average subjects ultimately expected to win between 1.2 and 1.7 fewer rounds over a set than they actually did.

Scatter plots of individual Set II predictions and outcomes in absolute contests are shown in Figure 5. As with forecasts of relative ability, individual errors were often quite large. Under-confidence in the verbal task is clearly illustrated, with almost all subjects having negative forecast errors.

A summary of these results is provided in Figure 6, and leads to the following observations. First, mean accuracy is more common than either over- or under-confidence. Second, under-confidence is more prevalent than over-confidence.

However the results tend to be task specific, with overconfidence more prevalent in the maximization task, and under-confidence in the verbal task. In addition, significant mean forecast error (in either direction) is more common in the absolute treatments than the relative, persisting in Set II revised forecasts in three of four cases for the former, versus one in four cases for the latter. Possible explanations for these patterns are discussed in Section 4.

3.2 Forecast Incentives, Moral Hazard and Accuracy

We find no evidence that using a quadratic scoring rule to modestly reward subjects for forecast accuracy altered the observable effort they made in either the maximization or verbal tournaments.¹³ We compared measures of prediction, effort quality and quantity, and outcomes for sessions with and without the scoring rule in both task orders.¹⁴ Mann-Whitney tests discerned no significant differences in any variables in any set or order, with one borderline exception. In particular, the initial predictions for Set I of the maximization task when it came first was marginally more pessimistic with the scoring rule than without ($\rho = 0.104$).

Did rewarding subjects for forecast accuracy make them better forecasters? Our results suggest not. Our modest scoring rule did not reduce either the mean or variance of forecast errors. In particular, recall that our design yields six *ERROR_i* observations per subject: one for the first set of a task, an initial and revised one for the second set of the task, and three analogous measures for the second task. As

¹³ Unfortunately, an error appeared in the instructions illustrating the working of the quadratic scoring rule. In particular, the scoring rule (2) was first (correctly) described to subjects in words. Two worked examples immediately followed, however, in which the term in parentheses on the left hand side of (2) was shown as multiplied by 2, rather than raised to the power 2. However the points generated on the right hand side in each case corresponded to squaring. None of the 240 subjects queried the error, and given the correct verbal description of the formula in the immediately preceding paragraph, we believe they would have understood that the bracket should have been squared. See Supplementary Appendix A for more detail.

¹⁴ Measures of effort quantity and quality are described more fully in the next section.

Mann-Whitney tests confirm, none of these *ERROR* measures was significantly different when the scoring rule was used than when it was omitted.¹⁵ Similarly, Levene's Test for the Equality of Variance detected no difference in the variance of any of the *ERROR* measures.

In short, we find that using a modest quadratic scoring rule to reward accurate forecasts brought neither moral hazard nor greater forecast accuracy. This is why we pooled sessions in Table 2 that differed only in their use of forecast incentives.

3.3 Auxiliary Findings

Aside from examining mean forecast errors, we also test 1) whether they vary systematically with subject characteristics, 2) whether *ex ante* overconfidence helps subjects do better in subsequent tasks, and 3) whether subjects learn from initial errors. To this end, we passively recorded participants' keystroke entries during each task, and collected demographic data following each session. From the keystroke records we derived measures of the quantity and quality of subjects' efforts at each task. Each measure was averaged over the ten rounds of a set.

Effort quantity was defined as the number of keys pressed per round in the verbal decoding task, or number of spreadsheet cells entered in the maximization task. Censoring hampers the measurement of effort quantity in tasks with absolute criteria for winning, because efforts beyond those needed to win a round are not counted. Using the censored results directly could make it appear that successful subjects in a given round made little effort, while the unsuccessful made much. Instead, when

¹⁵ These results are confirmed in subsequent regression analysis.

measuring a subject's average effort quantity over a set in absolute contests, we average only over the rounds of the set that he or she lost.¹⁶

Effort quality was defined initially using a number of plausible measures that we programmed the computer server to calculate as data was collected.¹⁷ From these alternative measures we selected those whose *ex post* bivariate correlation with winning in a given set exceeded 0.3, but whose correlation in that set with similar measures did not exceed 0.6. For the decoding task, our filter left *Letters Decoded Per Keys Pressed*, because subjects who made fewer, targeted guesses won more often than subjects who guessed randomly or who entered every letter on the keyboard systematically. For the maximization task, our filter left *Percentage Smart Moves* and *Percentage New Moves*. The first refers to an absence of persevering on the spreadsheet in directions where the function value was falling, and the second to the absence of backtracking through cells that had already been searched. As in the verbal task, purposeful rather than random search strategies were more successful on average. Descriptive statistics for these and all demographic and treatment variables are provided in Supplementary Appendix B.

To test whether errors vary with subject characteristics, we regressed initial and revised Set II forecast errors on Set II effort quantity and quality measures and demographics, under both relative and absolute winning criteria. The results are reported in Appendix I. Since our Set II effort quantity and quality measures might

¹⁶ This excludes three and five subjects from analysis of Set II of the maximization and verbal contests, respectively. Those excluded won all ten rounds of the set. As an alternative specification, we omitted effort quantity altogether from the four regressions for absolute contests. The fit of these regressions is poorer in three of the four cases, and effort quality measures that were not already significant (in initial and revised maximization) became so.

¹⁷ For the verbal task, we defined three measures of effort quality: (1) the proportion of letters successfully decoded as a percentage of all keys pressed (including number or punctuation keys), (2) the proportion of letters guessed two or more times for a given letter position, and (3) the proportion of letters guessed three or more times for a given letter position. For the maximization task we measured quality by (1) the lack of perseverance in incorrect directions, (2) the lack of backtracking through previously explored cells, and (3) the average height gained over total number of moves.

themselves be influenced by forecast error, we also ran corresponding instrumental variable (IV) regressions, with Set II effort measures instrumented on corresponding Set I measures. Hausman specification tests failed to reject the null of exogeneity in all cases, and so we report only OLS results.¹⁸

As is clear from Appendix I, very little of the variation in initial or revised subject forecast error is explained by subject characteristics such as gender, age, self-reported grade range (A, B, or C), or first language status. There is a limited tendency for males to show a higher (overconfident) forecast error in initial forecasts in some treatments, but not in revised forecasts. More curiously, there is a tendency for those who used greater effort quantity or quality in the verbal task towards negative forecast errors (under-confidence), both in initial and revised forecasts.

Next, we test whether subjects who exhibited initial overconfidence regarding the first task encountered persevered to do better on the second set of the subsequent task. As reported in Appendix II, we regress second task Set II win rates on effort quantity and quality in that task and set, then repeat the regression adding Set I and then revised Set II forecast errors from the first task. A significant coefficient on the first task forecast errors would indicate a direct effect of overconfidence on subsequent cross-task performance, while a drop in the coefficients on effort across regressions would indicate an indirect effect of overconfidence on performance via higher effort quality or quantity. Contrary to the hypotheses summarized by Taylor and Brown (1988), we find no evidence that initial overconfidence has either direct or indirect effects on subsequent performance under either criteria of winning or order of

¹⁸ Chi-square values ranged from 0.26 to 5.75 in eight tests by task, criteria for winning, and initial or revised forecasts. The corresponding p-values ranged from 1.000 to 0.889. In any event, coefficient estimates and significance levels were very similar in OLS and IV regressions.

tasks. Effort coefficients (jointly tested) never differ significantly as forecast errors are added, and first task forecast errors are never statistically significant.

Turning finally to learning, we find reassuring evidence that subjects do update predictions for Set II of a task in the direction of the realized Set I forecast error. In particular, we regress subjects' *revised* predicted number of wins for Set II of a task on their *initial* predicted number of wins for Set II (their "priors"), and the error of their forecast from Set I (new information). OLS regressions were run separately for each task under both relative and absolute winning criteria, as shown in Table 3. In all four cases the prior and new information were significant at any conventional level, and subjects on average lowered revised predictions if Set I predictions turned out to be overconfident. Subjects were also more anchored to initial Set II predictions in the relative tournaments than in the absolute contests, with a significant difference in coefficient magnitudes between columns (3) to (1) and (4) to (2). Subjects however did not always revise Set II forecasts *sufficiently* to eliminate initial mean bias, or to preserve the lack of it. Returning to Table 2, there are eight treatment cells of task, order, and criteria for winning. In three of these treatment cells, Set II mean forecast error differed significantly from zero both in initial *and* revised forecasts, and in only one of these (absolute, verbal, maximization first) did the revision reduce the mean forecast error significantly (using a t-test). In a fourth treatment cell (absolute, maximization, verbal first) mean forecast error did not differ from zero initially, but changed significantly in revision, ending significantly negative! In the four remaining treatment cells however, mean forecast error was indiscernibly close to zero initially and did not change significantly in revision.

4. Discussion and Conclusions

Overconfidence can have important economic consequences, but has received little direct testing within the discipline. The evidence that does exist is mixed, in contrast with the robust psychological evidence for the “above-average” effect. However, this data has been collected from hypothetical surveys without feedback or opportunities for learning or revision. We use laboratory experiments to address these limitations. Subjects participate in two experimental tasks, and are rewarded for both performance and accuracy of their predictions.

We fail to find confirmation of a robust overconfidence bias, either at the aggregate level as done in earlier studies, or at the individual level where predictions are compared to outcomes. We find limited evidence of overconfidence for only one of two tasks, in one of two orders, with one of two criteria for winning – maximizing an unknown function on a spreadsheet when it is the first task encountered, and when the criteria for winning is performance relative to others. We find no such bias about relative performance in the same task when it is experienced after another task, nor in a verbal decoding task in any order, nor in any forecasts of performance of either task against an absolute threshold.¹⁹

Overall our results suggest that mean accuracy is more common than either over- or under-confidence, and that when systematic errors occur under-confidence is actually more prevalent than over-confidence. Moreover, the results tend to be task specific with overconfidence more prevalent in the maximization task, and under-confidence more common in the verbal task. In addition, predictions tended to be less accurate in the absolute treatments than the relative. We found significant under-confidence persisting in revised Set II forecasts in three of the four cases for absolute

¹⁹ We do find initial overconfidence for absolute maximization performance when it is experienced after verbal contests, but it does not persist to initial or revised Set II forecasts.

contests, versus significant overconfidence in only one of four cases for relative tournaments.

What features may account for these differences? First, we originally hypothesized that predicting absolute success rates would be easier than relative ones, because the former requires knowledge of your own abilities, whereas the latter *also* required knowledge of others' abilities. On reflection, however, predictions about relative ability require different, rather than additional information. In the decoding task, for example, subjects need only estimate whether their ability ranks in the top five or bottom seven of the twelve participants in the session, and not also whether they could decode a certain number of letters.

Second, we postulate that the feedback from the two practice rounds that preceded each task provided more helpful information in the relative tournaments than in the absolute contests. In the relative tournaments, 42% of subjects (or five of twelve) were guaranteed to win each practice round, and so every subject gained two signals of their ability relative to other session participants. If relative ability remained stable across the twelve participants as the session progressed, these signals would be very informative. In contrast, in absolute contests, few subjects were able to attain the thresholds required in only two practice rounds.²⁰ This would tell most subjects only that they were starting well below the threshold required, and give them two observations from which to extrapolate their own trajectories of improvement over the following 20 rounds.

Finally, we speculate that the under-confident forecast bias for the verbal task in particular was made possible by that task's greater scope for improvement with experience. Decoding words required considerable keyboard skills, and offered more

²⁰ 3 and 21 percent won the first and second practice rounds of the maximization task, respectively. For the decoding task, the corresponding numbers were 0 and 15 percent.

channels for improvement in technique. Indeed, there was an “increasing return to improvement,” as letters decoded early in a round assisted in the decoding of subsequent words within that round. In contrast, maximization did not require intense keyboard skills, and improvements in technique were limited to insights gained over triangulation, efficient use of the arrow keys, and the need to pay constant attention. Our speculation is partially supported by the observation that average success rates changed more dramatically from first to second sets in absolute decoding contests (+1.76 rounds when first, +1.85 rounds when second) than in absolute maximization contests (+0.77 rounds when first, +1.85 rounds when second). With more potential for improvement comes a greater potential not to recognize it.

Among auxiliary results, we found that little in the variation of forecast errors could be explained by subject characteristics such as gender or first language. However subjects who exerted greater or “smarter” effort in the verbal task in particular did not adequately account for this in their forecasts, and thus had substantially more negative errors than others. This might suggest that most subjects had little idea *ex ante* idea of which decoding strategy would be the most successful. We also found no evidence in support of the hypothesis that subjects with initial overconfidence performed better in a subsequent task, whether directly, or indirectly via higher effort quantity or quality. Regarding learning, subjects generally revised second set predictions in the direction of first set outcomes, but not always sufficiently to eliminate initial mean forecast errors where they occurred.

Finally, regarding the methodology of eliciting predictions in experiments, we find that offering modest incentives for accurate forecasts using a quadratic scoring rule did not seem to trigger moral hazard problems in subsequent performance. But neither did it reduce the mean or variance of forecast errors. Increasing the relative reward from prediction over performance might improve the latter, but also trigger the

former. Until this is investigated, our results lend support to the practice of measuring peoples' expectations over outcomes that have incentives with prediction elicitation questions that do not.

Table 1: Experimental Treatments

Treatment	Winning Criteria	First Task	Scoring Rule	Sessions ^a	Number of Observations
1	Relative	Maximization	Yes	1, 2, 9, 10	48
2	Relative	Maximization	No	17, 18	24
3	Relative	Verbal	Yes	5, 6, 15, 16	48
4	Relative	Verbal	No	19, 20	24
5	Absolute	Maximization	Yes	3, 4, 11, 12	48
6	Absolute	Verbal	Yes	7, 8, 13, 14	46 ^{b,c}

^a Each session contained 12 participants.

^b There were only 11 participants in one of these sessions.

^c One of the subjects somehow circumvented the programmed constraint that spreadsheet search in the maximization contest be between contiguous cells. Fortunately, this was in an absolute contest session, and had no spillover effects to other participants. The results for this subject were excluded from the analysis.

Table 2: Summary of Mean Results

	ABSOLUTE CONTESTS		RELATIVE TOURNAMENTS	
	Maximization First N = 48	Verbal First N = 46	Maximization First N = 72	Verbal First N=72
MAXIMIZATION				
Predicted Wins/10 Set I	3.77	3.67	4.76**	3.61**
Predicted Wins/10 Set II	4.63	4.76	4.89***	3.75
Actual Wins Set I	3.19	2.74	4.17	4.17
Revised Predicted Wins Set II	4.33	3.80	4.71**	3.97
Actual Wins Set II	3.96	4.59	4.17	4.17
Individual Error (Pred. – Wins)				
Set I	0.58	0.93**	0.60*	-0.56*
Initial Set II	0.67	0.17	0.72**	-0.42
Revised Set II	0.38	-0.78**	0.54*	-0.19
VERBAL DECODING				
Predicted Wins/10 Set I	2.44	2.57	3.75	3.61**
Predicted Wins/10 Set II	2.94	3.46	3.96	3.96
Actual Wins Set I	3.42	3.07	4.15 ^a	4.17
Revised Predicted Wins Set II	3.58	3.67	4.21	3.97
Actual Wins Set II	5.27	4.83	4.15 ^a	4.15 ^a
Individual Error(Pred. – Wins)				
Set I	-0.98**	-0.50	-0.40	-0.56
Initial Set II	-2.33***	-1.37**	-0.19	-0.19
Revised Set II	-1.69**	-1.15***	-0.01	-0.18

*, **, *** denote significance at the 10%, 5% and 1% levels, respectively, in two tailed t-tests. In the case of Predicted Wins, denotes significant difference from average win rate. In the case of the Individual Error measures, indicates significant difference from zero.

^a A minor feedback error resulted in one subject in one set in each of three sessions being told and rewarded as if he or she had won one fewer rounds than was the case. This lowered the aggregate wins averaged per round to 4.08 (= 49 wins/120 rounds) for the flawed set, and 4.15 when combined with the five error-free sets.

Table 3: Learning

DEPENDENT VARIABLE: Revised Predicted Wins Set II				
	ABSOLUTE CONTESTS		RELATIVE TOURNAMENTS	
	<u>Maximization</u>	<u>Verbal</u>	<u>Maximization</u>	<u>Verbal</u>
<i>Initial Predicted Wins Set II</i>	0.782 ^{***} (0.057)	0.741 ^{***} (0.081)	0.979 ^{***} (0.051)	0.942 ^{***} (0.060)
<i>Forecast Error Set I</i>	-0.675 ^{***} (0.050)	-0.643 ^{***} (0.056)	-0.639 ^{***} (0.039)	-0.600 ^{***} (0.039)
<i>Order of Task (Max. First =1)</i>	0.397 (0.248)	-0.014 (0.287)	0.358 [*] (0.208)	0.328 (0.229)
<i>Constant</i>	0.714 ^{**} (0.309)	0.792 ^{**} (0.353)	-0.054 (0.244)	-0.090 (0.294)
<i>N</i>	94	94	144	144
<i>Adjusted R²</i>	0.734	0.644	0.775	0.714

*, **, *** denote significance at the 10%, 5% and 1% levels, respectively, in two tailed t-tests. High values of *Forecast Error* indicate predictions that are more confident. Numbers in parentheses are standard errors. Estimated using Stata 7.

Figure 1: A Subject's View of the Maximization Task

Practice Round 1											
Target: 100000						11					
	-6	-5	-4	-3	-2	-1	0	1	2	3	4
11											
10											
9											
8											
7											
6											
5											
4											
3											
2							1654				
1											
0											
-1											
-2											
-3											
-4											
-5											

This (hypothetical) subject is in the first practice round, with 11 of 60 seconds remaining. By moving the cursor from cell to cell, he has reached the value (or “height”) of 1654. This version of the maximization task has an absolute criteria for winning: reaching or exceeding the target value of 100,000.

Figure 2: A Subject's View of the Verbal Decoding Task

This (hypothetical) subject is in the first practice round, with 14 of 60 seconds remaining. He has successfully decoded one entire word, “mraqo” as “scone”, and the second and third letters of a second word, “ifkog.” Correctly decoded letters in the current word are shaded.

Figure 3: Sequence of Decisions for a Given Task

Figure 4: Final Predictions vs. Outcomes for Set II of Relative Tournaments

Figure 5: Final Predictions vs. Outcomes in Set II of Absolute Contests

Figure 6: Incidence of Mean Error Overconfidence, Accuracy, and Under-Confidence

<i>Forecast Error in</i>	Overconfidence	Accuracy	Under-confidence
<i>Set I</i>	▲■	△△■□	▲□
<i>Initial Set II</i>	▲	▲△△■■	□□
<i>Revised Set II</i>	▲	▲△△■	■□□

Key: ▲ = relative maximization tournaments (in both orders)
 △ = relative verbal decoding tournaments (in both orders)
 ■ = absolute maximization contests (in both orders)
 □ = absolute verbal decoding contests (in both orders)

Appendix I: Determinants of Initial and Revised Forecast Errors for Set II

DEPENDENT VARIABLE	Initial Set II Forecast Error				Revised Set II Forecast Error			
	Relative Tournaments		Absolute Contests		Relative Tournaments		Absolute Contests	
	Max	Verbal	Max	Verbal	Max	Verbal	Max	Verbal
<i>Demographics</i>								
<i>MALE</i>	--	0.869 [*] (0.479)	-1.265 ^{**} (0.574)	--	--	--	--	--
<i>AGE</i>	--	--	--	--	--	--	--	-0.065 ^{**} (0.028)
<i>GRADE AVE</i>	--	--	--	--	--	--	--	--
<i>ENGLISH 1ST</i>	--	--	--	--	--	--	--	--
<i>TAKEN ECON</i>	--	--	1.741 ^{***} (0.650)	--	--	--	--	0.794 [*] (0.447)
<i>TAKEN STAT</i>	--	--	--	-1.241 [*] (0.654)	--	--	-0.963 [*] (0.577)	-1.042 ^{**} (0.461)
<i>UNIV YEAR</i>	--	--	--	--	--	--	--	--
<i>Effort Variables</i>								
<i>CELLS ENTERED</i>	--	--	--	--	0.034 ^{**} (0.014)	--	-0.030 [*] (0.018)	--
<i>% SMART CELLS</i>	-0.116 ^{***} (0.028)	--	-0.225 ^{***} (0.078)	--	--	--	--	--
<i>% NEW CELLS</i>	--	--	--	--	--	--	--	--
<i>KEYS PRESSED</i>	-0.083 ^{***} (0.010)	--	-0.056 ^{***} (0.009)	--	-0.023 ^{***} (0.008)	--	-0.030 ^{***} (0.006)	--
<i>#DECODED PER KEY PRESSED</i>	-105.8 ^{***} (13.57)	--	-88.28 ^{***} (14.13)	--	-50.93 ^{***} (10.97)	--	-59.47 ^{***} (9.966)	--

Appendix I Continued:

DEPENDENT VARIABLE	Initial Set II Forecast Error				Revised Set II Forecast Error			
	Relative Tournaments		Absolute Contests		Relative Tournaments		Absolute Contests	
	Max	Verbal	Max	Verbal	Max	Verbal	Max	Verbal
<i>Other Variables</i>								
<i>ORDER (MV=1)</i>	0.957 [*] (0.543)	--	1.356 ^{**} (0.592)	-1.110 ^{**} (0.512)	--	--	1.563 ^{***} (0.499)	--
<i>SCORING RULE</i>	--	--			--	--		
<i>WON 1 PRACTICE ROUND</i>	--	--	1.568 ^{**} (0.718)	--	1.181 ^{**} (0.489)	--	--	--
<i>WON 2 PRACTICE ROUNDS</i>	--	1.723 ^{***} (0.607)	--	<i>a</i>	--	1.274 ^{***} (0.490)	--	<i>a</i>
<i>CONSTANT</i>	--	17.89 ^{***} (2.887)	55.60 ^{***} (14.99)	13.18 ^{***} (2.716)	--	6.167 ^{***} (2.333)	33.71 ^{***} (12.64)	9.457 ^{***} (1.916)
<i>N^b</i>	143	141	91	89	143	141	91	89
<i>AdjustedR²</i>	0.084	0.350	0.261	0.409	0.058	0.123	0.140	0.336

*, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. Run on Stata 7. Numbers in parentheses are standard errors. Eligible but insignificant coefficients are retained but not reported, and are denoted by "--".

^a No individuals won both practice rounds in this treatment.

^b Sample sizes vary because of a missing observation for *AGE*, and incomplete data for *CELLS ENTERED* (5) or *KEYS PRESSED* (3) using our censored measures for absolute contests, and for *#DECODED PER KEY PRESSED* (2) for the relative verbal tournaments.

Appendix IIa: Overconfidence and Better Outcomes - Relative

DEPENDENT VARIABLE: Actual Number of Wins in Set II

	Maximization when Second			Verbal Decoding when Second		
<i>Effort in Set II Of Second Task</i>						
<i>CELLS ENTERED</i>	0.004 (0.017)	0.003 (0.017)	0.002 (0.018)			
<i>% SMART CELLS</i>	0.104*** (0.036)	0.104*** (0.037)	0.103*** (0.037)			
<i>KEYS PRESSED</i>				0.069** (0.011)	0.071*** (0.011)	0.071*** (0.011)
<i>#DECODED PER KEY PRESSED</i>				99.14*** (12.65)	98.67*** (12.67)	97.42*** (12.83)
<i>Forecast Error In Set I of Prior Task</i>		-0.019 (0.102)	-0.027 (0.106)		-0.069 (0.075)	-0.068 (0.075)
<i>Revised Forecast Error In Set II of Prior Task</i>			0.048 (0.167)			-0.069 (0.095)
<i>Other Variables</i>						
<i>MALE</i>	2.054*** (0.693)	2.067*** (0.701)	2.105*** (0.718)	--	--	--
<i>WON 1 PRACTICE ROUND</i>	--	--	--	1.003* (0.567)	0.976* (0.567)	0.959* (0.571)
<i>WON 2 PRACTICE ROUNDS</i>	--	--	--	1.432** (0.637)	1.479** (0.640)	1.434** (0.645)
<i>TAKEN STAT</i>	1.091* (0.647)	1.082* (0.654)	--	--	--	--
<i>GRADE AVE</i>	--	--	--	--	--	--
<i>N</i>	72	72	72	72	72	72
Adjusted R ²	0.287	0.276	0.266	0.664	0.663	0.661

*, **, *** indicate significance at the 10%, 5% and 1% levels, respectively in two tailed tests. Run on Stata 7. Numbers in parentheses are standard errors. Some eligible but insignificant coefficients are retained but not reported, and are denoted by "--".

Appendix IIb: Overconfidence and Better Outcomes - Absolute

DEPENDENT VARIABLE: Actual Number of Wins in Set II

	Maximization when Second			Verbal Decoding when Second		
<i>Effort in Set II Of Second Task</i>						
<i>CELLS ENTERED</i>	0.037** (0.015)	0.035** (0.016)	0.034** (0.015)			
<i>% SMART CELLS</i>	0.259*** (0.077)	0.247*** (0.078)	0.249*** (0.077)			
<i>% NEW CELLS</i>	0.161 (0.161)	0.185 (0.164)	0.139 (0.163)			
<i>KEYS PRESSED</i>				0.064*** (0.007)	0.064*** (0.007)	0.067*** (0.008)
<i>#DECODED PER KEY PRESSED</i>				106.1*** (12.60)	106.4*** (13.13)	106.9*** (13.21)
<i>Forecast Error In Set I of Prior Task</i>		-0.102 (0.116)	-0.050 (0.118)		0.007 (0.082)	0.002 (0.082)
<i>Revised Forecast Error In Set II of Prior Task</i>			-0.242 (0.155)			-0.102 (0.128)
<i>Other Variables</i>						
<i>MALE</i>	1.107* (0.574)	1.115* (0.576)	1.227** (0.568)	--	--	--
<i>WON 1 PRACTICE ROUND</i>	--	--	--	--	--	--
<i>WON 2 PRACTICE ROUNDS</i>	--	--	--	<i>a</i>	<i>a</i>	<i>a</i>
<i>TAKEN STAT</i>	--	--	--	1.114** (0.448)	1.117** (0.455)	1.008** (0.477)
<i>GRADE AVE</i>	1.103** (0.446)	1.066** (0.450)	1.022** (0.441)	--	--	--
<i>N^b</i>	43	43	43	45	45	45
Adjusted R ²	0.525	0.522	0.542	0.716	0.708	0.705

*, **, *** indicate significance at the 10%, 5% and 1% levels, respectively in two tailed tests. Run on Stata 7. Numbers in parentheses are standard errors. Some eligible but insignificant coefficients are retained but not reported, and are denoted by "--".

^a No individuals won both practice rounds in this treatment.

^b Sample sizes vary because of incomplete data for gross effort using our censored measure for absolute contests.

References

- Alloy, Lauren B. and Anthony H. Ahrens (1987) "Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others", *Journal of Personality and Social Psychology*, 52(2), 366-378.
- Babcock, Linda and George Loewenstein (1997) "Explaining bargaining impasse: the role of self-serving biases", *Journal of Economic Perspectives*, 11(1), 109-26.
- Baker, Lynn A. and Robert Emery (1993) "When every relationship is above average: perceptions and expectations of divorce at the time of marriage", *Law and Human Behavior*, 17(4), 439-450.
- Barber, Brad M. and Terrance Odean (2001) "Boys will be boys: gender, overconfidence, and common stock investment", *Quarterly Journal of Economics*, 116, 261-292.
- Benabou, Roland and Jean Tirole (2002) "Self-confidence and personal motivation", *Quarterly Journal of Economics* 117, 871-915.
- Berg, Joyce E., Daley, Lane A., Dickhaut, John W. and John R. O'Brien (1986) "Controlling preferences for lotteries on units of experimental exchange", *Quarterly Journal of Economics*, 101, 281-306.
- Camerer, Colin F. and Dan Lovallo (1999) "Overconfidence and excess entry: An experimental approach", *American Economic Review*, 89(1), 306-318.
- Compte, Olivier and Postlewaite, Andrew (2004) "Confidence-enhanced performance", *American Economic Review*, 94(5), 1536-57.
- Daniel, Kent, Hirshleifer, David and Avanidhar Subrahmanyam (1998) "Investor psychology and security market under- and overreactions", *Journal of Finance*, 53(6), 1839-85.
- Das, M., and A. van Soest (1997) "Expected and realized income changes: evidence from the Dutch Socio-Economic Panel", *Journal of Economic Behavior and Organization*, 32, 137-154.
- Das, M., and A. van Soest (1999) "A panel data model for subjective information on household income growth", *Journal of Economic Behavior and Organization*, 40, 409-426.
- Dickinson, David L. (2005) "The chilling effect of optimism: The case of final-offer arbitration", *Journal of Socio-Economics*, forthcoming.
- Dominitz, J., (1998) "Earnings expectations, revisions, and realizations", *Review of Economics and Statistics*, 80(3) 374-88.

- Dominitz, Jeff and Charles F. Manski (1997) "Using expectations data to study subjective income expectations", *Journal of the American Statistical Association*, 92(439), 855-867.
- Dubra, Juan (2004) "Optimism and overconfidence in search", *Review of Economic Dynamics*, 7(1), 198-218.
- Fischhoff, Baruch; Parker, Andrew; De Bruin, Wandi; Downs, Julie; Palmgren, Claire; Dawes, Robin and Charles Manski (2000) "Teen expectations for significant life events", *Public Opinion Quarterly*, 64, 189-205.
- Frank, Robert F. and Philip J. Cook (1995) *The Winner-Take-All Society*, New York: The Free Press.
- Grether, David (1992) "Testing Bayes rule and the representativeness heuristic: some experimental evidence", *Journal of Economic Behavior and Organization* 17, 31-57.
- Hamermesh, Daniel S. (1985) "Expectations, life expectancy, and economic behavior", *Quarterly Journal of Economics*, 100(2), 389-408.
- Harbring, Christine and Bernd Irlenbusch (2003) "An experimental study on tournament design", *Labour Economics*, 10(4), 443-64.
- Huck, Steffen and Georg Weizsacker (2002) "Do players correctly estimate what others do? Evidence of conservatism in beliefs", *Journal of Economic Behavior and Organization*, 47, 71-85.
- Hurd, Michael and Susann Rohwedder (2003) "The retirement-consumption puzzle: anticipated and actual declines in spending at retirement", *NBER Working Paper 9586*.
- Jamison, Julian S. and Dean S. Karlan (2003) "When curiosity kills the profits: an experimental examination", Mimeo, January, Kellogg School of Management, Northwestern University.
- Lichtenstein, Sarah, Fischhoff, Baruch and Lawrence Phillips (1982) "Calibration of probabilities: the state of the art to 1980," in (Kahneman, Daniel; Slovic, Paul and Amos Tversky, eds) *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.
- Malmendier, Ulrike and Geoffrey Tate (2005) "CEO overconfidence and corporate investment", *Journal of Finance*, forthcoming.
- Manove, Michael and A Jorge Padilla (1999) "Banking (conservatively) with optimists", *Rand Journal of Economics*, 30(2), 324-50.
- McKelvey, Richard and Talbot Page (1990) "Public and private information: an experimental study of information pooling", *Econometrica*, 58(6), 1321-39.

- Miller, Arthur; Ashton, William; McHoskey, John and Joel Gimbel (1990) "What price attractiveness? Stereotype and risk factors in sun tanning behaviour", *Journal of Applied Social Psychology*, 20(15), 1272-1300
- Morrison, William and E. Elisabet Rutstrom (2000) "The role of beliefs in an investment game experiment", Mimeo, University of South Carolina.
- Nyarko, Yaw and Andrew Schotter (2002) "An experimental study of belief learning using elicited beliefs", *Econometrica*, 70(3), 971-1005
- Offerman, Theo and Joep Sonnemans (2001) "Is the quadratic scoring rule really incentive compatible?", working paper at www.fee.uva.nl/creed/pdf/files/qspapers5.pdf.
- Orrison, Alannah, Schotter, Andrew and Keith Weigelt (1998) "On the design of optimal organizations using tournaments: an experimental examination", mimeo, July, New York University
- Pyszczynski, Tom and Kathleen Holt (1987) "Depression, self-focused attention, and expectancies for positive and negative future life events for self and others", *Journal of Personality and Social Psychology*, 52(5), 994-1001.
- Roll, Richard (1986) "The hubris hypothesis of corporate takeovers", *Journal of Business*, 59(2), 197-216.
- Selten, Reinhard; Sadrieh, Abdolkarim and Klaus Abbink (1999) "Money does not induce risk neutral behavior, but binary lotteries do even worse", *Theory and Decision* 46, 211-49.
- Smith, Adam (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*, reprinted (Andrew Skinner, ed) Penguin Books, 1983.
- Smith, V. Kerry, Donald H. Taylor Jr. and Frank A. Sloan (2001) "Longevity expectations and death: Can people predict their own demise?" *American Economic Review*, 91(4), 1126-34.
- Sonnemans, Joep, Arthur Schram and Theo Offerman (1998) "Public good provision and public bad prevention: the effect of framing", *Journal of Economic Behavior and Organization* 34, 143-61.
- Svenson, Ola (1981) "Are we all less risky and more skillful than our fellow drivers?" *Acta Psychologica*, 47, 143-148.
- Taylor, Shelley and Jonathon Brown (1988) "Illusion and well-being: a social psychological perspective on mental health", *Psychological Bulletin* 103, 193-210.
- van Dijk, Frans, Joep Sonnemans and Frans van Winden (2001) "Incentive systems in a real effort experiment", *European Economic Review* 45, 187-214.

- Viscusi, W. Kip (1990) "Do smokers underestimate risks?" *Journal of Political Economy*, 98(6), 1253-1269.
- Weinstein, Neil D. (1980) "Unrealistic optimism about future life events", *Journal of Personality and Social Psychology*, 39(5), 806-820.
- Weinstein, Neil D. (1982) "Unrealistic optimism about susceptibility to health problems", *Journal of Behavioral Medicine*, 5(4), 441-60.
- Weinstein, Neil D. (1984) "Why it won't happen to me: perceptions of risk factors and susceptibility", *Health Psychology*, 3(5), 431-457.
- Weinstein, Neil D. (1987) "Unrealistic optimism about susceptibility to health problems: conclusions from a community-wide sample", *Journal of Behavioral Medicine*, 10(5), 481-500.
- Weinstein, Neil D. (1989) "Optimistic biases about personal risks", *Science*, 246, 1232-33.
- Weinstein, Neil D. (1998) "Accuracy of smokers' risk perceptions", *Annals of Behavioral Medicine*, 20(2), 135-140.
- Weinstein, Neil D. and William M. Klein (1995) "Resistance of personal risk perceptions to debiasing interventions", *Health Psychology*, 14(2), 132-140.
- Williams, Arlington W. (1987) "The formation of price forecasts in experimental markets", *Journal of Money, Credit, and Banking*, 19(1), 1-18.