JEL classification: C91 D83 D84 J24

Rational Expectations of Own Performance: An Experimental Study

Jeremy Clark*

Lana Friesen†

October 23rd, 2003

Abstract:

According to rational choice theory, people will choose their careers, level of work effort or investment based in part on their expectations of success. But are people's expectations of their likelihood of success accurate? Evidence accumulated by psychologists suggest that on average people underestimate their risks and overestimate their abilities relative to others. We test for such optimistic bias in experiments where people must predict their relative or absolute success in incentive-based verbal and maximization contests. We ask subjects to provide initial and revised point estimates of their success rates, either hypothetically or with a scoring rule that rewards forecast accuracy. We then passively measure the quantity and quality of subjects' efforts and subsequent outcomes. Bias in forecasts is evaluated at the aggregate level as done by psychologists, but also at the individual level using realized outcomes. We find limited evidence of excess optimism only in relative maximization contests when encountered first, and excess pessimism or accuracy elsewhere. Experience across contests and updating does not always increase the accuracy of self-assessed forecasts, and even when accuracy improves biases are rarely eliminated entirely. Methodologically, we find no evidence that a modest quadratic scoring rule introduces moral hazard for own-outcome forecasts, but neither does it increase forecast accuracy or lower variance.

---

* Department of 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

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

Keywords: rational expectations; optimistic bias; self-assessment

1. Introduction and Background

Are people good judges of their own abilities? Are they unbiased on average, as rational choice theory assumes, or do they spin webs of optimistic self-delusion, as psychologists suspect? The answer to this question has profound implications, because people may rely on self-assessments when choosing career paths, investment in higher education or in the stock market, starting businesses, committing to marriage, and so on. Systematic over-confidence could lead people to spend too much effort trying to break into highly competitive or difficult schools or jobs or markets (Frank and Cook, 1995). It could suggest a role for outside intervention to "de-bias" people's expectations, or else tax or limit their choices. In contrast, a lack of systematic bias would suggest that the decisions people make are individually optimal, given available *ex ante* information.

Attempts to measure the accuracy of people's expectations of their own abilities have been thwarted by several factors. Conceptually, researchers need a true or rational benchmark against which expectations can be compared. Fischoff et al. (2000) compare teenagers' current probability point estimates of various life events against current aggregate realizations. As the authors recognize, this comparison is imperfect in that future realizations may differ from current ones, and because individuals may correctly perceive factors that increase or decrease their personal likelihoods relative to published reference population frequencies. Individual future realized outcomes would seem an ideal benchmark, except that a) few longitudinal studies yet exist that precisely compare quantitative forecasts and subsequent outcomes, and b) individuals may alter their subsequent behaviour in ways that alter the actual likelihood of particular outcomes.

Within experimental psychology Weinstein (1980) pioneered a literature that cleverly side-stepped the need for benchmarks by testing for relative forecast bias in the aggregate. He asked individuals to rate their chance of experiencing a pleasant or unpleasant event *relative to their peers*, or others of a similar occupation or age or situation. Weinstein elicited ordinal forecasts on a seven point scale centred at zero, with -3 for "much below average (risk)" and +3 for "much above average (risk)". Problems about averages vs. medians aside[1], this literature has yielded a robust anomaly that most people report they anticipate an above-average risk of high starting salaries and job satisfaction, staying married, having gifted children, and so on (Weinstein (1980), Baker and Emery (1993)). Conversely, most people report they have a below average risk of car accidents (Svenson 1981), job loss or unemployment (Weinstein (1980), or health problems (Weinstein 1982, 1984, 1987, 1989, 1995, 1998, Miller et al. (1990)).

This "optimistic group bias" has been found to be most prevalent for events over which respondents perceive some measure of control, or lack personal experience, or judge to be infrequent (Weinstein (1984), (1987)). It has been studied primarily among student subjects, but also across the general population (Weinstein (1987)), and appears resistant to verbal warnings and "de-biasing" manipulations (Weinstein (1995)). Over-optimism has been attributed to the "representativeness heuristic", or tendency of individuals to judge the likelihood of an event by the degree to which it resembles a stereotype, regardless of base rate frequency (Tversky and Kahneman (1982)). It seems conspicuously absent only among the clinically depressed (Alloy and Ahrens (1987), Pyszczynski et al. (1987)).

[1] It is thoroughly rational for most people to rate themselves below average in a distribution that is negatively skewed, and above average in a distribution that is positively skewed.

While economists are beginning to incorporate psychological findings into economic analysis (Rabin 1998), optimism bias has received surprisingly little investigation. Existing studies either focus on the consequences of overconfidence assuming it exists, or test for it using only non-incentive survey instruments.[2] For instance, in an experiment on market entry, Camerer and Lovallo (1999) suggested that overconfidence about one's *relative* ability might explain excess entry and lower profits, and extrapolated that it might also explain the high rate of new business failures. Barber and Odean (2001) compare the common stock investments of men and women, and find that men trade more excessively than women and as a consequence obtained lower profits on average. The authors attribute the difference to men being more overconfident than women about the relative precision of their knowledge.[3]

In survey investigations, Dominitz (1998) uses three consecutive rounds of the United States Survey of Economic Expectations to elicit entire probability distributions of future earnings and compare these with realized outcomes. Subject to sample attrition problems, Dominitz find that overall Americans are too optimistic about changes in future income. At the same time, expectations are revised more sharply in response to income drops than income rises. Das and van Soest (1997), (1999) follow up data on expected and realized income changes over six consecutive years of the Dutch Socio-Economic Panel. These authors also contend with sample attrition problems, and unlike Dominitz were limited to ordered categorical data. In contrast to Dominitz, Das and van Soest find that heads of households generally

[2] Overconfidence is also beginning to find its way into theoretical models. For example, Benabou and Tirole (2002) develop a model where overconfidence about one's own ability can be self-serving for individuals with imperfect willpower. The potentially "self-serving" nature of optimism biases has long been recognised in the psychology literature (Taylor and Brown (1988)).

[3] Lichtenstein, Fischhoff and Phillips (1982) provide a review of this calibration literature.

underestimated future income changes. This was particularly true for those who had experienced a fall in income, and the effect was robust over all 6 years of the panel.

One limitation of surveys, whether by economists or psychologists, is that people are not provided with incentives to think carefully about their predictions (other than the general incentive to be cooperative and appear intelligent to the interviewer). Another limitation is that annual longitudinal surveys do not provide people with timely feedback on the accuracy of their predictions, or the opportunity to revise their predictions about a fixed event as new information becomes available.

We take a different approach in this paper, using experiments to directly measure people's expectations of their own performance, either in absolute contests or relative tournaments. By using experiments, we hope to provide participants with incentives for accurate prediction, and with opportunities for feedback and revision. We will also be able to test for optimism bias using individual realized outcomes as the benchmark for evaluation, as well as group averages.[4]

To provide subjects with incentives for careful prediction, we use a quadratic scoring rule to reward them for greater forecast accuracy over their performance in contests or tournaments (see Huck and Weizsacker (2002)). Technically, quadratic scoring rules are only incentive compatible for agents making predictions concerning events over which they have no control (or moral hazard), and who are risk neutral. Nonetheless, with suitable attention paid to the relative rewards for performance and prediction, moral hazard can be eliminated in theory. Further, by passively measuring the quantity and quality of effort with and without a scoring rule, moral hazard can be tested for in practice. Risk neutrality can also be induced in theory, by using payoffs in lottery points for both prediction and performance (Berg et al. (1986)).

[4] The data set also allows us to check whether biases are self-serving, by comparing individual biases with performance.

To provide subjects with feedback, we have them enter two sets of two different competitions, revising their forecasts for the second set after playing the first. In more structured game theoretic settings, Camerer and Kunreuther (1989) find that excess optimism about market prices (and therefore earnings) can be reduced substantially by such feedback.

Finally, in the event that people are biased in their expectations of success in competitive settings, it is useful to know if this is due to errors in forecasting their own (absolute) ability, or their ability relative to others. We thus compare the accuracy of people's predictions in contests with absolute or relative performance criteria.

We find that optimistic group bias is not as robust as the psychology literature would suggest. We also find that modest scoring rules can be used without introducing moral hazard, but that they do not reduce mean forecast errors over non-incentive forecasts. 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 will run as follows. In Section 2 we outline our experimental design. In Section 3 we present our results, and in Section 4 we discuss our findings and conclude.

2. Experimental Design: Mountain Climbing and Espionage

We wanted a design in which subjects had to predict their own performance in a task requiring real yet quantifiable effort. We also wished to examine predictions over tasks that required diverse skills and opportunities for learning. We settled on two tasks. The first was maximizing the unknown function:

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

where x and y are coordinates on a spreadsheet, and α, b_1, b_2, and b_3 are parameters. Subjects attempt to maximize (1) by moving contiguously from cell to cell on a spreadsheet. This task is a simplified version of the two-variable optimization task used by van Dijk, Sonnemans, and Winden (2001) to examine the effect of different incentive schemes on work effort. In each round the unknown function took on the same smooth, paraboloid, shape, and obtained a maximum value of 100,400. The most direct route to the peak required 200 moves, with 240 moves possible per round. 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.

The second task was decoding five letter words, drawn from a random sample of 1155 such words taken from the 1993 *Shorter Oxford English Dictionary*.[5] The coding scheme was comprised of a unique letter-for-letter mapping, and was held constant for all words within a round, but randomly reset between rounds.

These maximization and verbal decoding tasks were fully computerized, and took place in 60 second rounds, with ten rounds per set. Each participant tried 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.

In any given session, 12 subjects would read computerized instructions for the first task, try two practice rounds, then try a first and second sets of 10 rounds. After the practice rounds but before the first set, subjects were asked how many of the first and second 10 rounds they thought they would win. After completing the first 10 rounds, subjects were reminded of their initial forecast for the *second* 10 rounds, and

[5] Proper names and obscure words were avoided.

asked to re-enter a prediction, whether revised or not. Upon completing the second 10 rounds, subjects were presented with instructions for the second task, and then proceeded through an analogous order of practice rounds, prediction for both sets, first set, revised prediction, and second set. Subjects were given immediate feedback on whether they had won or lost each round, and were informed of their earnings accumulation from prediction and performance after each set.

2.1 Criteria For Winning

Of 16 sessions that rewarded forecast accuracy, half (3, 4, 7, 8, 11, 12, 13 and 14) required subjects to meet an absolute standard to win each round. For maximization, the standard was to reach a value of 100,000 on the unknown function within 60 seconds, with a constraint of 4 contiguous cells' movement per second. A cell's value appeared only when the cursor was on it. For verbal decoding, the standard was decoding 2 five letter words within 60 seconds, with a constraint of 4 letter guesses per second. Subjects were free to continue maximising or decoding letters for the full 60 seconds, even after meeting the standard, though we do not use any "post-win" data that this generates.

The other eight sessions (1, 2, 5, 6, 9, 10, 15, 16) required subjects to meet a relative standard to win each round. For maximization, the 5 of 12 participants who had reached the highest function value by the end of a round won, while for decoding, the 5 of 12 participants who had decoded the greatest number of letters by the end of the round won.[6] Ties were automatically broken at random, so that there were always exactly 5 winners, and an aggregate probability of winning each round of .4167 (= 5/12).

[6] The tournament design literature suggests that effort is promoted best and collusive laziness avoided with larger group sizes (8 rather than 4 or 2), and intermediate rather than extreme win rates (Orrison et al. (1998) , Harbring and Irlenbusch (2001)).

2.2 Eliciting Predictions

Exactly *how* beliefs should be elicited has raised disagreements within and between disciplines. At one extreme, psychologists such as Weinstein (1998) argue that the public has difficulty generating risk estimates in terms of percentages, so that ordinal measures ("very likely", "somewhat likely" etc.) should be used. At the other extreme, 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. More commonly, an intermediate approach of eliciting point estimates of probability as numbers between 1 and 100, or simply 1 and 10 is used ((Weiner (1976), Viscusi (1990), Fischoff et al. (2000), Morrison and Rutstrom (2000), Jamison and Karlan (2003)). We adopt this intermediate approach and ask subjects to report the number of rounds (out of 10) that they expect to win.[7]

In 16 sessions subjects were rewarded for the *ex post* accuracy of their *ex ante* set predictions according to the quadratic scoring rule:

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

P and *W* refer to the predicted and realized number of rounds won, respectively.[8] The predicted number of rounds won for the first set of a task was compared to the first set outcome, while the revised prediction for the second set was compared to the second set outcome.

This rule gives risk neutral subjects (without moral hazard) proper incentives for reporting their true expected value for the number of rounds they would win. Note

[7] This estimate can be converted into a point estimate of the probability of winning each round by dividing by ten.

[8] The scoring rule takes the same form as that used by Huck and Weizsacker (2002) to elicit beliefs about the choices of other players. Quadratic scoring rules have also been used to elicit probabilities in the context of Bayesian inference (e.g. McKelvey and Page (1990), Grether (1992)) as well as beliefs about other players (e.g. Sonnemans, Schram and Offerman (1998), Morrison and Rutstrom (2000), Nyarko and Schotter (2000)).

from (2) that earnings from forecast accuracy would be maximized if the number of rounds won turned out to match the initial prediction perfectly. Forecast earnings would be minimized at 0 per set if the forecast error was at the maximum possible level of 10 rounds. Unfortunately, the rule also suffers from weak payoff dominance, in that small forecast errors of one or two rounds cost little (Davis and Holt (1993)). In addition, because subjects have partial control over the outcomes, they could face moral hazard incentives, such as deliberately making low forecasts of success and then "throwing" rounds in order to receive payment for accurate forecasts. We address this below. To induce risk neutrality in theory, (though less clearly in practice (Selten, Sadrieh and Abbink (1999))), we reward subjects for accurate prediction and performance using points that they accumulate for use in a single draw at the end of the session (as in Berg et al. (1986)). Each subject enters a final private draw between an $8 prize and a $30 prize, with the probability of winning $30 given by the number of points accumulated from performance and accurate prediction over the session.

2.3 Addressing Moral Hazard in Design

Since people's performance in a contest is constrained positively by their ability, we focus on the problem of people underperforming in order to make their *ex post* outcome align with low *ex ante* forecasts. This could occur under two conditions. First, a subject could make an honest but overly pessimistic prediction of winning x of 10 rounds, and then throw any rounds *after* winning that number. Second, a subject could purposefully make a low prediction, then purposefully under-perform to meet it.

To foreclose both incentives, we set the marginal reward from winning an additional round, 2 points, to exceed any possible marginal cost from increasing forecast error by an additional round (see Table 1). These marginal costs range from

Table 1: Points Payoffs from Performance and Prediction Per Set

# Rounds	Payoff From Performance Total	Marginal	Penalty From Forecast Error Total	Marginal
0	0	--	.00	--
1	2	**2**	.05	**.05**
2	4	**2**	.20	**.15**
3	6	**2**	.45	**.25**
4	8	**2**	.80	**.35**
5	10	**2**	1.25	**.45**
6	12	**2**	1.80	**.55**
7	14	**2**	2.45	**.65**
8	16	**2**	3.20	**.75**
9	18	**2**	4.05	**.85**
10	20	**2**	5.00	**.95**

.05 points to .95 points. Thus in all cases subject will do well to make accurate predictions, but better if they actually win as many rounds as possible. Over all four sets of a session, subjects could accumulate a maximum of 20 points from accurate prediction, and 80 points from winning every round.

Theoretical incentives aside, we also check whether rewarding forecast accuracy creates moral hazard in practice by running additional relative tournament sessions (17 to 20) with non-incentive forecasts.[9] These sessions were replications of tournament sessions 1, 2, 5 and 6, except that 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.

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

3. Experimental Results

239 student subjects participated over 20 sessions of the experiment between March and May of 2003 at the University of Canterbury.[10] Students were recruited from large first and second year courses in economics, statistics, and mathematics.[11] Each session took approximately one hour and thirty minutes, and earnings ranged from NZ$10 to $42, with an average of $23.97.[12]

Table 2 provides summary statistics of our results. Perhaps not surprisingly, contest or tournament outcomes sometimes differed depending on the order in which they were experienced. In relative tournaments, Mann-Whitney tests detected significant order effects for predictions and forecast error in the maximization task, though none were found in 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. Order effects could derive positively from skills or experience gained in the first task spilling over to the second, or negatively from fatigue. While no order effects were detected for relative verbal tournaments, we report them in disaggregated form in keeping with the other cases where separation is warranted.

3.1 Ex Post Tests for Moral Hazard and Scoring Rule Effects

Consistent with the incentives in our design, we find no evidence that using a modest quadratic scoring rule to reward subjects for forecast accuracy in relative tournaments created moral hazard. When maximization came first, we compared all measures of prediction, effort quality and quantity, and outcomes for Sessions 1, 2, 9

[10] One of the 239 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 are dropped from analysis.

[11] Demographic characteristics were collected from students following each session, but have not yet been analyzed. The data presented here was part of a larger experiment collecting additional data.

[12] The 2003 adult hourly minimum wage in New Zealand was $8.50.

Table 2: Summary of Mean Results

	ABSOLUTE CONTESTS		RELATIVE TOURNAMENTS	
	Maximisation First N = 48	Verbal First N = 46	Maximisation First N = 72	Verbal First N=72
MAXIMISATION				
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 Pred. Wins Set II	4.33	3.80	4.71**	3.97
Actual Wins Set II	3.96	4.59	4.17	4.17
Individual Bias (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 Pred. 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 Bias (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 Bias measures, indicates significant difference from zero bias.

[a] A minor feedback errors resulted in one subject in one set in each of sessions 2, 10 and 20 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 (= 49wins/120 rounds) for the flawed set, and 4.15 when combined with the five error-free sets.

and 10 (with the scoring rule) against Sessions 17 and 18 (without). For variables measured round by round, such as effort level and quality, we average an individuals' observations over the ten rounds of a set as our unit of observation.[13] Mann-Whitney tests discerned no significant differences in any variables, with one borderline exception. In particular, the distribution of initial predictions for Set I of the maximization tournament was marginally lower with the scoring rule than without (Mann Whitney r = .104, N_I = 48, N_{II} = 24). While this difference is consistent with moral hazard effects, no such differences were detected for initial or revised Set II predictions under maximization, nor under any variables for the subsequent verbal decoding tournament. More importantly, there was no evidence of differences in effort quality or quantity when the scoring rule was used for any set or task. Difference in effort would be a necessary condition for evidence of moral hazard in the form of "thrown" rounds.

When the verbal decoding task came first, we compared measures of prediction, effort and outcomes for Sessions 5, 6, 15 and 16 (with the scoring rule) against Sessions 19 and 20 (without). Mann Whitney tests detected no significant difference in any measure for either type of tournament (again, N_I = 48, N_{II} = 24).

While it is a relief to know that rewarding subjects for forecast accuracy did not appear to influence their subsequent performance via moral hazard, it would be more interesting yet to know if it made them better forecasters. Our results here suggest that our modest scoring rule did not reduce either forecast bias or variance. In particular, we define forecast bias, $BIAS_i$ as the difference between the number of rounds subject i said she was going to win, and the number she did win. Our design

[13] In the verbal decoding contests we collected data on keys pressed, letters pressed, and frequency of repetitive incorrect guesses. In the maximization contests we collected data on the value reached, number of moves, perseverance in incorrect directions, and amount of backtracking through previously explored cells.

yields six *BIAS* 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 other task. To our surprise, none of the *BIAS* measures were discernibly different when the scoring rule was used. Similarly, Levene's Test for the Equality of Variance detected no difference in the variance of any of the *BIAS* measures with the scoring rule than without.[14]

In short, we find that using a quadratic scoring rule with low payoffs relative to those from performance had no discernible effect on moral hazard or forecast accuracy over non-incentive forecasts. We shall thus pool results for corresponding sessions that did and did not reward forecast accuracy.

3.3 Where Have All the Optimists Gone?

We move now to our principal investigation of forecast accuracy. Our design enables us to test for rational expectations using the group bias approach of Weinstein, or an individual bias approach using realized outcomes. Group bias is testable in the relative tournaments (Sessions 1,2,5,6,9,10,15,16,17,18,19,20) by asking whether mean predictions for each set differ significantly from the aggregate win rate of .417.[15] The last two columns of Table 2 provide the results.

We do find evidence of group optimism bias, but only for the maximization tournament, and only when it was the first tournament experienced. In that tournament and order, t-tests indicate that the mean forecast over 72 participants was significantly higher than .417 in the first set forecasts (.476), as well as in the initial (.489) and revised (.471) second set forecasts. While the average forecast is creeping in the right direction, the bias remains strong. In contrast, when the maximization

[14] In contrast, Grether (1992) found that financial incentives reduced the number of "absurd" responses.

[15] For reasons explained in footnote *a* of Table 2, the aggregate win rate for 3 of 8 cases was in fact .4153 rather than .4167.

tournament is experienced *after* verbal decoding, we find initial group pessimism bias for Set I only (.361), giving way to group-rational expectations for the initial and revised forecasts for Set II. For the verbal decoding tournament when experienced first, we find initial group pessimism bias for Set I only (.361) and group-rational expectations thereafter. When verbal decoding is experienced second, we find group-rational expectations throughout. Thus, we find group optimism bias in 3 of 12 set forecast cases, group pessimism bias in 2 of 12 cases, and group-rationality in 7 of 12 cases.

Moving to individual level analysis, we compare the predictions and outcomes for each individual in both sets of both tournaments. Here it is possible to consider data from both relative and absolute sessions. The discrepancy between each individual's prediction and outcome in a set, $BIAS_i$, should be zero when averaged across the sample in case of rational expectations, positive in case of optimistic bias, and negative in case of pessimistic bias. Beginning with relative tournaments, t-tests find that mean *BIAS* is positive (optimistic) only for the maximization tournament when it is experienced first. Consistent with the group bias tests, this excess optimism persists through to the initial and revised Set II forecasts: it is robust to experience and the opportunity for revision. The order of magnitude of the bias is around 5%: subjects expect, on average, to win ½ a round more than they actually do over the 10 rounds. When the maximization tournament is second, the mean *BIAS* is negative only for Set I and not significantly different from 0 for either forecast of Set II.

In slight contrast to group bias tests, rational expectations is not rejected for the verbal tournament regardless of tournament order, set, or opportunity for revision.

Thus, of 12 possible set forecast measures, we find optimistic bias for 3, pessimistic bias for 1, and rational expectations for 8.

Individual scatterplots of the four revised Set II predictions and outcomes for the relative tournaments are provided in Figure 1. Points above the 45-degree line reflect pessimistic (or negatively biased) predictions, while points below reflect optimistic (or positively biased) predictions. Rational expectations are not rejected for three of the four cases, and excess optimism is found for the fourth (maximization as a first tournament). Note that even in the three cases of rational expectations, the plots show considerable dispersion in the bias numbers. Zero bias on average appears to result from an even split between overly-optimistic and pessimistic subjects, rather than widespread individual accuracy. Indeed the predictions of some subjects were off by as many as 6 (out of 10), even in the second set of the second tournament! When optimism persists this is because most subjects are optimistic (42% were optimistic, 26% unbiased, and the remaining 32% pessimistic), not simply because of a few outliers. Nevertheless, in 3 of 4 cases individual biases in Set II revised forecasts averaged out to be insignificantly different from zero.

Before the advocates of rational expectations celebrate, let us move to the forecast bias tests in absolute contests. The extensive literature on group optimism bias does not offer guidance about what to expect here. When the maximization contest comes first, we find no excess optimism, but only rational expectations for Set I and both forecasts of Set II. When maximization comes second (after a bruising verbal contest), subjects begin with optimistic bias for Set I, and overcorrect to pessimistic bias for the revised forecast for Set II. That is, initial forecasts for the *second* set that would have been rational were instead revised with excessively pessimistic ones after disappointing results in the first set.

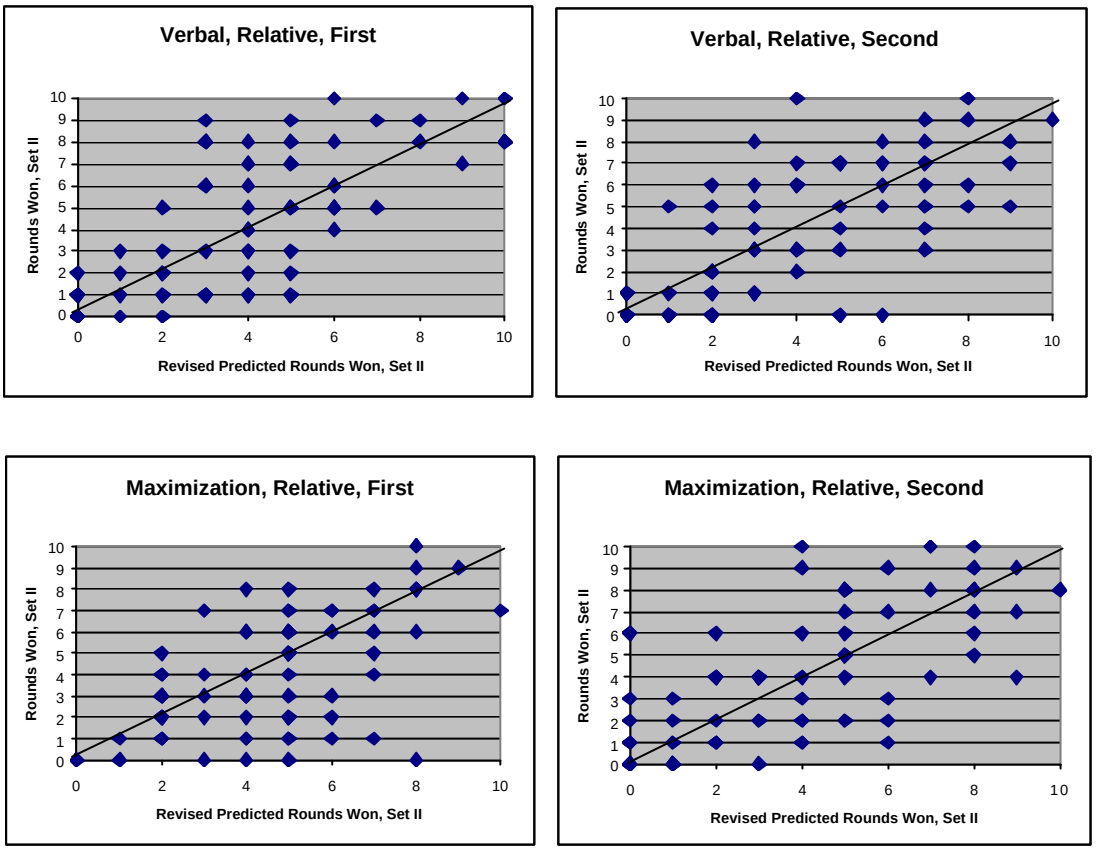

Figure 1: Final Predictions vs. Outcomes for Set II of Relative Tournaments

Even more strikingly, subjects showed pessimistic bias in almost all cases in the verbal decoding contest, regardless of the order in which it appeared. Revised Set II predictions reduced the bias over initial Set II predictions, but not by nearly enough to eliminate the bias. In fact, the predictions were even worse after one set of experience was gained! The size of this bias ranges from 12-17% depending on task order: on average subjects expected to win between 1.2 and 1.7 rounds less than they actually did over 10 rounds.

Thus, of 12 possible set forecasts in the absolute contests, we found optimistic bias for 1, pessimistic bias for 6, and rational expectations for 5.

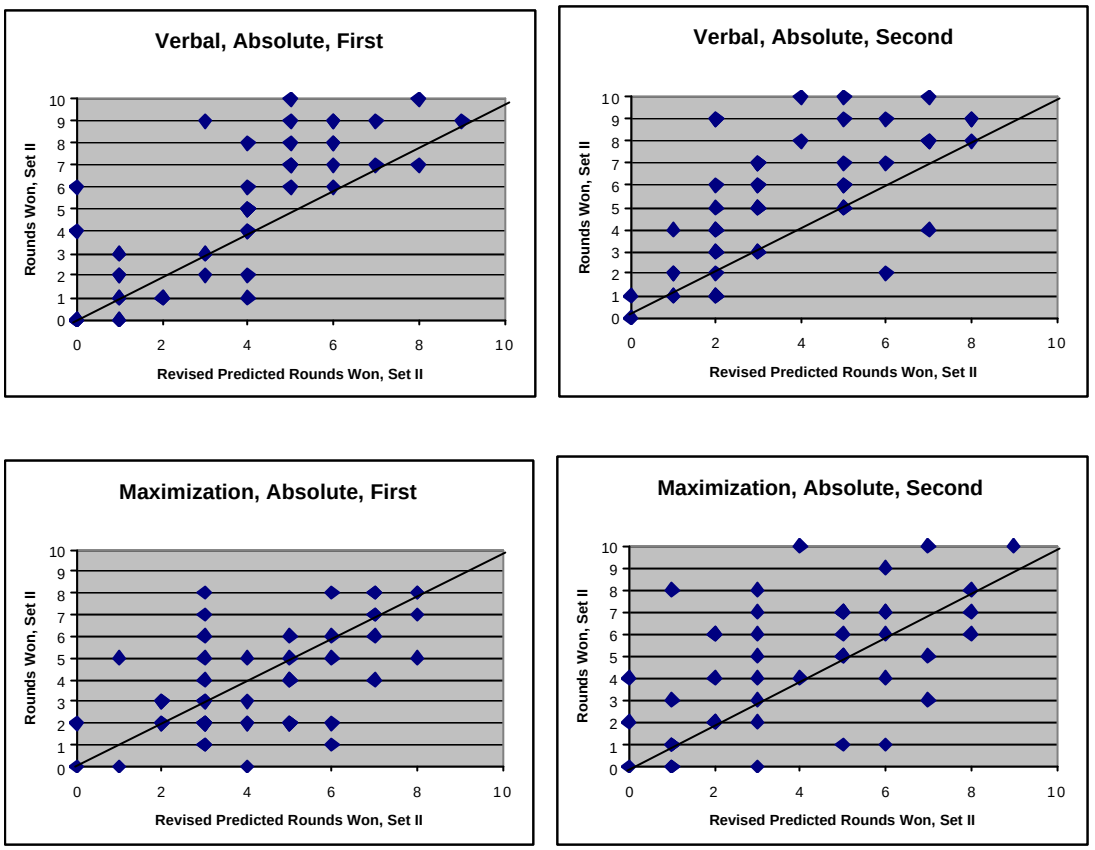

Figure 2: Final Prediction vs. Outcome in Set II of Absolute Contests

Individual scatterplots of revised Set II predictions and outcomes for the absolute contests are provided in Figure 2. As the scatterplots show, the majority of subjects were overly pessimistic in the verbal tournaments (57% and 71% of subjects when verbal came first and second respectively), and in the maximization when it came second (48%). As in the relative tournaments, individual biases were often quite large.

While subjects forecasts had greater bias in absolute contests than in relative tournaments, we do not find that they had greater variance. Levene's Test for the Equality of Variance in bias indicates no significant difference.

4. Discussion and Conclusions

A robust finding from psychological studies of risk perception is that people tend to rate themselves below-average in their risk of experiencing unpleasant events and above-average in their risk of experiencing pleasant ones (Weinstein (1980)). These findings have been derived from hypothetical surveys, with qualitative prediction elicitation, no feedback, and no opportunities for learning or revision. We design experiments that provide modest incentives for accurate, quantitative prediction, as well as opportunities for feedback, learning and forecast revision.

We fail to find confirmation of a robust optimism bias, either at the aggregate level as done in earlier studies, or at the individual level where predictions are compared to realized outcomes. We find strong evidence of optimistic bias 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 competition encountered, and when the criteria for winning is relative. We find no such optimism bias when maximization is experienced after another tournament, nor in verbal decoding tournaments in any order, nor in maximization or decoding contests in any order when the criteria for winning is absolute.[16]

Our results add support to the subsequent caveats in psychology that group optimism bias is less likely to be a problem when respondents perceive little control over round outcomes, or have personal experience of round outcomes, or judge the frequency of losing or winning to be reasonably high (Weinstein (1984), (1987)).

Positively, what we do find is support for rational expectations in relative tournaments, and some evidence of pessimistic bias in absolute contests, particularly verbal decoding. Focusing on revised forecasts for Set II (where feedback and

[16] We do find initial optimistic bias for maximization contests when they are experienced after verbal contests, but it does not persist to initial or revised forecasts for Set II.

experience is greatest) we do not reject rational expectations for 3 of 4 cases in relative tournaments, with excess optimism in the final case. In contrast, we do reject rational expectations in favour of excess pessimism for 3 of 4 cases in absolute contests, with rational expectations in the last. What accounts for these differences?

First, regarding prediction in relative vs. absolute contests, we originally hypothesized that predicting absolute success rates would be easier than predicting relative ones, because the former requires knowledge of your own abilities, whereas the latter *also* requires knowledge of others' abilities. On reflection, however, predictions about meeting relative thresholds require different, rather than additional information. In the decoding task, for example, subjects need only estimate whether their ability is in the top 5 or bottom 7 of the 12 participants in the session, and not whether they can accomplish specific tasks. Useful feedback on initial relative standing was provided in the two hypothetical practice rounds that preceded each relative tournament.

In contrast, it might be more cognitively demanding for subjects to predict whether they could decode 10 or more letters in 60 seconds, and the extent to which they would improve at doing so over 20 rounds. For the information provided in the two practice rounds preceding absolute contests would likely have revealed to subjects only that they were not yet near attaining the threshold.

Second, regarding strong pessimism in the absolute verbal contest in particular, we speculate that the scope for improvement there was greater than that in the maximization contest. Maximization did not require intense keyboard skills. Improvements in maximization ability were limited to insights gained over triangulation, efficient use of the arrow keys, and the need to pay constant attention. In contrast, verbal decoding required greater initial keyboard skills, and offered more

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 letters within that round. Our speculation is partially supported by the observation that average success rates changed more dramatically from first to second sets in decoding contests (+1.76 rounds when first, +1.85 rounds when second) than in maximization contests (+.85 rounds when first, +1.85 rounds when second). With more room for improvement over 20 rounds came a greater potential not to recognize that such improvement was possible.

Finally, regarding the methodology of eliciting predictions in experiments, we find that offering modest incentives for accurate forecasts using a quadratic scoring rule had little positive or negative effect. In particular, it seems possible to provide subjects with incentives to accurately predict their own performance without triggering moral hazard problems in subsequent performance. On the other hand, the modest incentives of the quadratic scoring rule, weakly payoff dominant at the best of times, did nothing to reduce the mean or variance of forecast bias. It is possible that increasing the relative payoff from prediction over performance would improve forecast accuracy, but it is also possible that doing so would trigger moral hazard in performance. Until this is investigated, our results lend support to the practice of measuring expectations over outcomes that have incentives with prediction elicitation questions that don't.

References

Alloy, Lauren B. and Ahrens, Anthony H. (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.

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 Howard Kunreuther (1989) "Decision processes for low probability events: policy implications" *Journal of Policy Analysis and Management*, 8, 4, 565-592.

Camerer, Colin F. and Dan Lovallo (1999) "Overconfidence and excess entry: An experimental approach" *American Economic Review*, 89(1), 306-318.

Das, M., van Soest, A., 1997. Expected and realized income changes: evidence from the Dutch Socio-Economic Panel. *Journal of Economic Behavior and Organization*, 32, 137-154.

_____ 1999. A panel data model for subjective information on household income growth. *Journal of Economic Behavior and Organization*, 40, 409-426.

Davis, Douglas and Chalres Holt (1993) *Experimental Economics* (Princeton University Press: New Jersey)

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.

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* (The Free Press: New York)

Grether, David (1992) "Testing Bayes rule and the representativeness heuristic: some experimental evidence", *Journal of Economic Behavior and Organization* 17, 31-57.

Harbring, Christine and Bernd Irlenbusch (2001) "An experimental study on tournament design" Mimeo, May, University of Bonn

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.

Jamison, Julian S. and Dean S. Karlan (2003) "When curiosity kills the profits: an experimental examination" Mimeo, January, Kellog School of Management, Northwestern University.

Lichtenstein, Sarah, Baruch Fischhoff and Lawrence Phillips (1982) "Calibration of probabilities: the state of the art to 1980," in *Judgement Under Uncertainty: Heuristics and Biases*, Daniel Kahneman, Paul Slovic and Amos Tversky editors, (Cambridge: Cambridge University Press).

Miller, Arthur; Ashton, William; McHoskey, John and Joel Gimbel (1990) "What price attractiveness? Stereotype and risk factors in suntanning 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, 2000, University of South Carolina.

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.

Rabin, Matthew (1998) "Psychology and economics" *Journal of Economic Literature*, 36, 11-46.

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

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.

Tayloer, Shelley and Jonathon Brown (1988) "Illusion and well-being: a social psychological perspective on mental health", *Psychological Bulletin* 103, 193-210.

Tversky, Amos and Daniel Kahneman (1982) "Judgement under uncertainty: heuristics and biases," in *Judgement Under Uncertainty: Heuristics and Biases 3*, Daniel Kahneman, Paul Slovic and Amos Tversky editors, (Cambridge: Cambridge University Press).

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.

Weiner, Bernard (1979) "A theory of motivation for some classroom experiences" *Journal of Educational Psychology*, 71, 1, 3-25.

Weinstein, Neil D. (1980) "Unrealistic optimism about future life events," *Journal of Personality and Social Psychology*, 1980, 39, 5, 806-820.

_____ (1982) "Unrealistic optimism about susceptibility to health problems," *Journal of Behavioral Medicine*, 5, 4, 1982.

_____ (1984) "Why it won't happen to me: perceptions of risk factors and susceptibility," *Health Psychology*, 3, 5, 431-457.

_____ (1987) "Unrealistic optimism about susceptibility to health problems: conclusions from a community-wide sample," *Journal of Behavioral Medicine*, 10, 5, 481-500.

_____ (1989) "Optimistic biases about personal risks," *Science*, 2xx, 1232-33.

_____ (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.