Essays in Behavioural Labour Economics

Tony So

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Economics, the University of Auckland, 2016.

Abstract

We use decision making experiments to study the impact of different pay schemes and feedback on performance and learning in a cognitively challenging task. In each of multiple rounds, subjects are presented with two cue values, Cue A and Cue B, and asked to predict the value of a third variable X, which is derived from a function of the two cue values. We measure performance with the forecast error of the prediction, the absolute difference between the actual and predicted values of X. Our pay schemes include: (1) piece rates, where subjects are paid on the basis of only their own performance; (2) a two-person winner-takes-all-tournament, where subjects are paired and the one with the highest performance earns a positive payoff while the other earns nothing; and 3) a fixed salary, where subjects are paid a flat lump-sum amount regardless of their performance. These pay schemes make up three of our experimental treatments: the Piece Rate, Tournament and Salary treatments. In our fourth treatment – the Piece Rate Win Lose treatment – we investigate the role of relative performance feedback by paying people piece rates based on their own performance, while informing them about whether they performed better or worse than a random partner. We find that the Piece Rate Win Lose and Salary treatments perform better than the Piece Rate and Tournament treatments, with no difference in performance between the former and latter two treatments. However, we only observe significant learning in the Tournament treatment.

JEL Codes: D83, J24, J33

Acknowledgements

This thesis has benefitted from the feedback provided by many people. Special thanks to my main supervisor, Ananish Chaudhuri, and co-supervisors, Jennifer Steele and Ryan Greenaway-McGrevy, for feedback, proof-reading, and motivation. Erwann Sbai and Ryan Greenaway-McGrevy have provided advice with econometrics. Paul Brown and Linda Cameron have also provided helpful comments. I also thank participants at Economic Science Association conferences in Auckland (2014), Honolulu (2014) and Sydney (2015); the Australia, New Zealand Workshop on Experimental Economics workshop in Hobart (2015); a seminar at Motu Economic and Public Policy Research in Wellington (2015) for valuable feedback. I am grateful also to John Trevithick for programming the experimental software; and to Sherry Li, Meg Paichayontvijit and Tian Leong for assistance conducting the experiments. Any errors found in this thesis are my own.

Contents

1. Introduction

A fundamental objective in economics is to maximise the productivity of workers, as this is necessary for a productive economy. While labour productivity is multi-dimensional and is affected by a range of factors, we focus on worker-level productivity and see how it is affected by two factors: the choice of pay scheme, and the feedback that is provided to the workers – in particular whether workers get feedback only on their own performance or whether they get feedback on others' performance as well. To the firm, these two factors are easily manipulated and are relatively cost effective ways of influencing worker effort, if indeed these manipulations can be shown to have an influence on worker productivity. This thesis utilises economic decision-making experiments to study how worker productivity is affected by various pay schemes, as well as how it is affected by the provision of feedback about the performance of others.

We also study how these two factors influence learning over time. Learning is an important element of productivity as most tasks require some degree of learning over time in order to do well. As learning occurs, workers pick up skills, heuristics and experience which improves their productivity. To our knowledge, there has been no prior research into how pay schemes and relative feedback affect learning.

In this thesis, we focus on three different pay schemes: piece rates, fixed salaries and tournaments. These pay schemes differ in terms of the incentives that they provide. Piece rates pay people according to their performance, so that they are motivated to perform. Piece rates require performance to be accurately observed, measurable and attributable to individual workers. In practice, this is not always the case.

Fixed salaries, on the other hand, are often employed when piece rates are not appropriate. Salaries pay workers based on the input of time, and does not depend on performance. Since salaries do not change with increased levels of effort or performance, workers do not have any incentive to exert effort over-and-above the minimum which is required, particularly when effort is costly.

Rank-order tournaments are an alternative method to pay workers. Tournaments pay workers depending on how they perform compared to others. If a worker performs better than his peers, he receives a larger 'prize', while other workers receive smaller ones or none at all. An employee's bonus is an example of a tournament-based scheme, where the relative performance of a particular worker will influence the manager's decision whether to pay a bonus or not and who to pay it to.

Theoretical research has shown that tournaments elicit a similar level of effort as do piece rates (Lazear & Rosen, 1981). Tournaments incentivise performance primarily through competition between workers. In order to earn more money, a worker will need to exert greater effort to improve their performance. If this increased performance leads to a higher rank, the worker's relative position improves and is rewarded with a larger monetary payment. If increased performance is insufficient to improve one's rank, there will be no subsequent increase in earnings. This could be because the performance of others also simultaneously improve. A worker will not necessarily earn more even if his absolute performance improves – this contrasts with piece rates where greater performance guarantees higher earnings. As a consequence, workers may not want to increase performance over and above what is required to attain a particular rank, since excess performance does not improve payoffs.

While tournaments incentivise performance through rank-dependent payoffs, there may exist other factors that also motivate performance. The act of competing may itself motivate workers to perform, even when monetary payoffs are not tied to the outcome of winning or losing. For example in a casual game of tennis, even when there are no tangible prizes or rewards for the winner, each player may still want to outplay his rival. This could be because players are stimulated by competition, or because they are motivated by the status and esteem that is associated with winning. In the case of tournaments, workers may be motivated to work harder in order to improve their chances of winning, even if they are not tangibly rewarded for it. We will refer to this notion of competition as 'competition for rank', where it is not linked to any monetary or tangible rewards. We study the effect of rank competition by providing workers feedback which informs them of how well they are performing relative to their peers, with such feedback having no effect whatsoever on monetary payoffs.

We focus on three primary research questions. The first asks which of the incentives – piece rates, fixed salaries and tournaments – induce the best performance from workers. Piece rates and tournaments depend on performance, while salaries do not. In addition to the different extrinsic incentives provided by each pay scheme, these different schemes might also influence workers' intrinsic motivation in different ways. A strand of literature suggests that performance pay schemes reduce peoples' intrinsic motivation to work, suggesting that salaries might not perform as poorly as would be otherwise anticipated (see Deci, Koestner, & Ryan, 1999). The overall effect brought about by the choice of pay scheme is therefore ambiguous. The literature regarding the efficacy of performance pay schemes remains contentious.

Our second research question aims to isolate the motivating effects of rank competition from competition for payoffs, both inherent in tournaments. Rank competition refers to the competition that is independent of monetary rewards, and is brought about solely by feedback on relative performance. If players are able to gauge their performance vis à vis others, they may improve their performance in order to improve their relative standing amongst their peers. In other words, rank competition may itself motivate performance under tournaments. On the other hand, the rank-dependent payoffs inherent in tournaments could also play a part in motiving performance. People may work harder in order to attain a higher rank, for which they are rewarded with a larger monetary payoff. Since both rank competition and payoff competition plausibly motivate performance, our second research question decomposes these effects and asks how each of these affects tournament performance.

Our final research question relates to learning. How do different pay schemes, as well as feedback on relative performance, affect people's learning? As learning occurs, people pick up skills, heuristics and experience which assists them to perform. Learning therefore plays a critical role in one's performance as it promotes better long term performance.

Our three research questions are re-iterated below:

R1. *How effective are piece rates, tournaments and fixed salaries in motivating worker performance?*

R2. *How do rank feedback and rank-dependent payoffs influence performance under tournaments?*

R3. *How is learning affected by different pay schemes and the provision of relative performance feedback?*

We examine these three research questions with a series of laboratory experiments. Lab experiments are appropriate since real-world data on worker performance and pay is difficult to obtain given the commercial sensitivity of micro-productivity data. Furthermore, lab experiments give us latitude to control for various factors, which may influence our results.

Our experiments utilise a real-effort stock forecasting task to simulate the tasks that workers face in their jobs. In this task, players in each of twenty rounds observe two numerical cue values, denoted Cue A and Cue B, and are asked to predict the underlying stock price X. This underlying stock price X is based on an underlying function of the two cue values. This task is cognitively challenging, reminiscent of the everyday tasks that workers face in their jobs. The difficult nature of this task, as well as the fact that players will need to uncover the underlying relationship in order to perform well, means that it is suitable to study learning. We measure performance in this task with the absolute forecast errors of the prediction, the absolute difference between the actual and predicted value of the stock price. The absolute forecast error indicates the accuracy of the forecast.

We pay our worker participants in three different ways: piece rates, tournaments and salaries. These three pay schemes constitute different treatments of our experiment – the Piece Rate, Tournament and Salary treatments. A further treatment manipulates feedback on relative performance. In the Piece Rate Win Lose treatment, workers are paid according to piece rates, but are provided additional information about whether they performed better or worse than a randomly matched, anonymous partner. Here relative feedback has no influence on monetary payoffs, allowing us to analyse the effect of rank competition.

Our findings show that the S treatment performers better than the PR and T treatments, suggesting that performance pay schemes are not as effective as fixed salaries in motivating performance. In terms of the tournament decomposition, we find that relative performance feedback is effective in motivating performance, with the PRWL treatment performing better than the PR treatment. We find that payoff competition plays a smaller role than rank competition in motivating performance under tournaments.

In terms of learning, however, we only find significant evidence of learning in the T treatment. Since learning was not observed in the PRWL treatment, we attribute such learning to competition over the rank-dependent payoffs inherent in tournaments. Learning is observed for both high and low performers in the T treatment.

The thesis proceeds in the following manner. Chapter 2 surveys the various strands of literature that are relevant to our study. Chapter 3 outlines the experimental design and procedures. Chapter 4 presents our overall results at the aggregated level. Chapter 5 focuses on the dynamics of learning. Chapters 6 and 7 disaggregates results by players of different ability and gender respectively, allowing us to uncover finer results which could be masked by aggregation. Chapter 8 concludes.

2. Literature Review

This chapter reviews the literature that relates to our research questions. In particular, we focus on studies that look at the motivating effects associated with pay schemes and relative performance feedback. Due to the large number of papers that is covered, we organise the review according to our three research questions.

We begin by outlining the framework for which we will adopt when interpreting empirical results. Next, we define and classify different pay schemes and discuss their theoretical properties. In line with our first research question, the following section of the review covers empirical studies that look at the motivating effects of different pay schemes.

Our second research question decomposes the effects that relative performance feedback and payoffs have under tournaments. We review the relevant literature, most of which focuses on the effect relative feedback has on performance. We also discuss the small number of papers that decompose tournaments and are able to shed light on the effect of rank-dependent payoffs.

Our third research question focuses on learning. To our knowledge there are few papers that focuses on learning in the context of pay schemes and feedback. We review these papers, as well as those that are related to the notion of learning in general.

The literature review concludes with a discussion of the novelty of our study. Our research methodology is laid out in the following chapter.

2.1. Conceptual Framework

2.1.1. Principal-Agent Model

We begin this literature review by explaining the conceptual framework which we will adopt when discussing the various effects that could come into play. Once this framework has been established, we will define and classify different types of pay schemes.

It is standard in the personnel economics literature to model workers' productivity in a principal-agent framework. The firm (principal) pays workers according to a particular pay scheme in order to produce output q, for which the firm sells at price P, the price which they can sell this output at.[1] The worker (agent) produces this output q by exerting effort[2] e, according to some production function:

$$q = f(e) + \varepsilon$$

We assume that higher effort entails higher production ($f'(e) > 0$) and that there are diminishing returns to effort ($f''(e) < 0$). Since the production function is specific to a particular task for a particular worker, we will make no further assumptions regarding it. In some instances, where output cannot be reliably measured or observed, the production function may be augmented with some noise ε. If there is output can be perfectly monitored, then $\varepsilon = 0$.

Workers decide on the level of effort to exert in order to produce their desired output. However, effort is costly to the worker. The cost of effort $c(e)$ is assumed to be increasing and convex in effort. Higher effort is more costly than less ($c'(e) > 0$) and effort becomes increasing costly to exert ($c''(e) > 0$). We consider workers of different ability to have different marginal effort costs, whereby marginal costs are lower for higher ability workers.

As q is the production associated with individual workers, it also represents their level of performance. The ratio of performance to effort q/e therefore measures the productivity of workers. Since effort is typically not empirically observable, the distinction between performance and productivity is much less relevant in practice. As such, we will use the terms 'performance' and 'productivity' interchangeably. Where it is relevant, we will infer effort from observed performance, drawing from the assumption that higher effort is necessary for higher performance.

Workers choose the amount of effort to exert in order to maximise their utility, which is the payment received from the firm according to a particular pay scheme less the cost associated with the effort exerted. Without loss of generality, this will yield a first order condition which suggests the worker should exert effort such that the incremental pay associated with (the output generated from) a marginal increase in effort is equal to their marginal cost of effort.

[1] For simplicity, we assume the firm is perfectly competitive in the output market, so it has no influence over the market price.

[2] Here we refer to effort in terms of its intensity. As such, we do not consider the duration of time that a worker works for as their effort.

Anticipating the effort exerted by the worker and the associated output, for a given set of pay scheme parameters (we explain this below), the firm chooses these parameters in order to maximise their profit. We will not discuss the firm's optimisation further, since we primarily focus on worker performance.

2.1.2. Taxonomy of Pay Schemes

We now define and classify different pay schemes.[3] The pay schemes that we discuss here differ along three dimensions: 1) whether or not the pay scheme depends on performance; 2) if it depends on performance, whether the pay scheme depends on absolute or relative performance; and 3) whether the pay scheme is discrete or continuous.

Fixed salaries (or wages)[4] are a common pay scheme that is invariant to performance. Workers receive the same pay whether they exert high or low effort, since the marginal pay with respect to output – and therefore effort – is zero. Since effort is costly and is not rewarded, workers would be expected to exert zero effort under salaries. This is the textbook example of moral hazard, whereby workers shirk under fixed wages once they have been hired.

In contrast, piece rates depend on performance. Total remuneration depends continuously on output, and is invariant to other factors. Piece rates can be expressed as:

$$Piece\ Rate\ = a + bq$$

where a is the fixed component and b is the marginal return on output q. a is non-negative while b is strictly positive. The fixed component a can be thought of as a base salary which is earned even if output is zero. b is sometimes itself referred to as the piece rate, representing the rate of earnings. We will however refer to b as the power of the piece rate, representing the stakes at play. The piece rate is low-powered if b is small, or high-powered if b is large.

From the perspective of workers, they will exert effort such that the marginal cost of effort is equal to the piece rate earnings associated with the marginal unit of effort. This means that

[3] See Prendergast (1999) for a review of various incentive schemes and related themes.

[4] We will use the terms 'wage' and 'salary' interchangeably in this thesis.

worker effort is positive and is typically considered to be high. In other words, in theory, piece rates motivate workers to perform. Piece rates therefore serve as the benchmark pay scheme.

Distinction can also be made between pay schemes which are either continuous or discrete. The aforementioned piece rate is an example of a continuous pay scheme, where the marginal payoffs from increased output is strictly positive for *any level* of output. A discrete pay scheme can be thought of as a lump-sum payment that is payable only when output exceeds a particular threshold. Performance bonuses can be considered discrete performance pay if a fixed amount is paid to workers whose performance exceeds some predefined standard. A discrete performance pay scheme will be defined to be one where the marginal payoffs from increased performance is zero for *at least some interval* of output, and is only strictly positive at the relevant threshold.

The motivating effects from discrete performance-based pay schemes are somewhat complex. For simplicity, a two-tiered discrete pay scheme takes the form:

$$Discrete\ Performance\ Pay = \begin{matrix} a & if\ \ q < \hat{q} \\ b & if\ \ q \geq \hat{q} \end{matrix} \quad where\ a < b$$

With this discrete payment scheme, there are no incentives for the worker to produce output beyond the threshold level $\hat{q}$, since the excess output $(q - \hat{q})$ does not lead to any additional monetary reward while is costly in terms of effort. Similarly, there are no incentives for workers to produce output levels in the open interval between 0 and the threshold output $\hat{q}$, since they will always be better off producing output of either 0 or $\hat{q}$, depending on their cost of effort. Consider some level of output less than $\hat{q}$. A worker will only exert sufficient effort to reach $\hat{q}$ by producing $(\hat{q} - q)$ output if the incremental reward $(b - a)$ is sufficient to at least compensate for the additional cost of effort. If the incremental reward is sufficient to compensate for the increase in effort required for attaining the higher level of payment, then it follows that the worker increases his output to reach the threshold $\hat{q}$. Otherwise, the worker is better off producing output of 0, since any excess output incurs unnecessary effort which is not compensated for. A rational worker should only be observed to produce output of either 0 or $\hat{q}$, depending on their cost of effort.

There are clear similarities between the motivating effects of discrete performance pay and salaries in that the marginal incentives to perform is zero over a range of output levels. In fact, salaries which have a performance requirement would be considered a discrete performance pay scheme, since the flat payment is received only when performance meets or exceeds a prescribed level. This type of scheme is most reminiscent to salaries in the real world, where there is a tacit understanding that workers will be laid off – earning nothing – if their performance is falls short of some standard.

Tournament pay schemes can be considered a special form of discrete performance-based pay, where the performance standard is endogenised. Rank-order tournaments (Lazear & Rosen, 1981; Green & Stokey, 1983; Nalebuff & Stiglitz, 1983) pay fixed monetary 'prizes' depending on how people perform relative to others, whereby a higher prize is paid to those with a higher performance rank. The key performance motivator in tournaments is the prize spread, the difference between the winning and losing prizes. The wider the prize spread, the greater the incentives for winning relative to losing, better motivating players to perform.

Theoretical tournament models show that they elicit a level of effort from workers analogous to that under piece rates. We shall refer to this as the Piece Rate Equivalence property of tournaments. In a lab experiment, Bull, Schotter, and Weigelt (1987) confirm Piece Rate Equivalence, where agents exert a similar level of effort under both tournaments and piece rates, although there is higher variation of effort choices under tournaments.

As with discrete performance pay, a unit improvement in performance under tournaments does not lead to additional earnings if it is insufficient to increase the rank of an agent. By contrast, continuous relative pay schemes reward the agent for their own performance, while penalises their pay against the performance of others. Examples of continuous relative pay schemes can be found in Knoeber and Thurman (1994) and Bandiera, Barankay, and Rasul (2005), where individual performance in these schemes are evaluated against the average performance of all workers in an additive or multiplicative manner, respectively.

Pay schemes that depend on relative performance inherently feature negative productivity spillovers. As relative performance increases, it will adversely affect the (expected) earnings of others. In a tournament setting, suppose an agent improves his rank and receives a higher prize. Due to the zero sum nature of performance ranks, this implies that the rank of someone else will fall, reducing their prize. Similarly in continuous relative pay schemes, the pay of others will be penalised by an increase in individual performance, ceteris paribus. If agents have concerns over the welfare of others, perhaps with fairness considerations (Fehr & Schmidt, 1999), then they may choose to withhold their performance to mitigate the negative effects they impose on others (see Bandiera et al., 2005).

Different pay schemes are also unique in their sorting efficiency. When there is a distribution of worker abilities, piece rates are shown both theoretically and empirically to attract and retain the high ability workers (Lazear, 1986, 2000; Dohmen & Falk, 2011). The performance improvements that arise from sorting are due to average performance increasing when the low performance group drop out, rather than due to improvements in individual performance per se. In our study we focus on worker-level productivity associated with different pay schemes, and for this reason the effects that arise from sorting and selection are not looked at in this thesis.

2.2. Extrinsic Incentives versus Intrinsic Motivation

We have previously discussed the extrinsic incentives associated with different pay schemes. These extrinsic incentives suggest that performance pay schemes are effective in motivating performance. On the other hand, fixed salaries do not incentivise performance. We will discuss empirical studies comparing these pay schemes shortly. This literature is contentious, with a number of papers showing that performance pay schemes are actually counter-productive in that they reduce performance.[5]

These papers draw on the notion of intrinsic motivation and suggest that extrinsic incentives 'crowd out' a person's innate desire to perform. A person is considered to be intrinsically motivated to perform a particular task if there is no apparent reward at stake for doing it – the only reward is attained by doing the task itself (Deci, 1972). Conceptually, intrinsic motivation is closely related to the effort that people exert.

[5] Although the terms "performance" and "output" are used repeatedly here to refer to the basis of "performance pay", it should be noted that payment schemes have applications in contexts outside the workplace. "Performance" should be more accurately thought of as the outcomes that are tied to the performance pay schemes, although we will refer to the former for ease of exposition.

Extrinsic incentives can reduce a person's intrinsic motivation through various channels. For example, the Overjustification Hypothesis (Lepper, Greene, & Nisbett, 1973)[6] posits that if a person is initially intrinsically motivated to perform a task in the absence of pay, but are subsequently offered a piece rate on his performance, then he faces too many reasons to perform. As a result, his intrinsic motivation falls to balance such 'overjustification'.

In other words, while extrinsic incentives can induce effort, they can also reduce it by negatively impacting intrinsic motivation. Whether performance pay schemes motivate performance or not depends on the relative magnitudes of the extrinsic incentive and intrinsic motivation effects.

In the following section, we draw on the relevant literature to discuss empirical papers relevant to our first research question, which asks which pay schemes of piece rates, tournaments and fixed salaries elicit the highest level of performance from people.

2.3. Evidence on Pay Schemes

2.3.1. Performance Incentives and No Pay

To study whether performance pay schemes are effective in incentivising behaviour, we first review studies that compare the performance of people when they are not paid to when they are paid according to a performance pay scheme. In many circumstances it is not the norm to be paid to do certain activities, for example when people are volunteers. If people are motivated by the performance pay schemes, then we would expect them to outperform those who are not paid at all for the same work.

Gneezy and Rustichini (2000b) studied how student volunteers respond to financial incentives. In Israel a number of 'donation days' are publicised each year where high school students would go door-knocking around the community to solicit charitable donations. The authors offered to pay one group of student volunteers a piece rate of 1% of the donations they collected on the day, and another group was offered 10% of the collection total. It was made

[6] We discuss other channels for which intrinsic motivation can be crowded out in Section 2.3.4.

clear that this percentage payment to them was not deducted from the donation money, but was rather paid out from separate funds available to the authors. A third group of students served as the control and did not receive any money: they collected money out of their own free will. For volunteers who were paid 1% of the donation money, the average donations solicited was lower than the control group, who were not paid for their services. Those who were paid 10% solicited a similar amount of money as the control group, though higher than those paid 1%.

Fryer (2011) found that financial payments made to high school students had no effect on their academic performance. In a large scale field experiment involving numerous schools in New York City and Chicago, he compared the academic performance of students who were incentivised to study with control groups who were not paid. In New York City, participating students received $5 U.S. for completing each test and a piece rate on performance in each; students could earn up to $25 for each of ten tests. The incentives were high powered, with the average student receiving $140 out of a maximum of $244. Students performed no differently to the control group who were not paid for their study.

Fryer also paid students in Chicago based on the letter grades in each of five courses issued in five-weekly reports. The discrete performance pay scheme paid students $50 for each A-grade, $35 for each B, $20 for each C and nothing for grades D and below. On average, these Chicago students received $696 out of the maximum $1875 that could be earned. Despite the sizeable incentives that were offered, the average grades of treatment students were similar to those at control schools, who were not paid.

In an experimental setting, Charness and Gneezy (2009) analysed whether a discrete performance-based pay scheme was effective in influencing students' long term gym-visiting behaviour. 120 students from the University of Chicago, who received gym membership as part of their tuition fees, were recruited for the study. Each student participant was asked to sign a consent form which allowed the authors to access their access card records for past and future gym attendance. Forty students served as the control group and were not paid at all, but were informed of the benefits of exercise. Forty other students were paid $25 if they visited the gym at least once in the upcoming week. The last group of forty students were paid $25 to visit the gym once in the following week, as per the previous group, but when they arrived to collect this initial payment, were offered an additional payment of $100 if they went to the gym 8 additional times across the next four weeks. The two treatment groups who were incentivised for gym attendance were also provided information about the benefits of exercise.

Participants were drawn to these incentives and increased their gym attendance accordingly to the level required. The pre-intervention level of gym attendance averaged 0.7 times per week. In the first week after the incentives were offered, when they had to visit the gym once, gym attendance increased to 1.7 times in the one-time group and 1.5 times for the eight-times group. The eight-times group were required to visit the gym eight more times over the following 4 weeks in order to collect an additional payment. Gym attendance increased once again to a level of 2.3 times a week. By comparison, there was no change in the average gym visits for the control group who were not paid. This suggests that the incentives have been effective in driving this behaviour.

By and large, the studies in this section found that performance pay schemes have mixed results when they are implemented. Gneezy and Rustichini (2000b) showed adverse effects when a low powered incentive is introduced. Fryer (2011) found that high powered payments tied to academic performance had no effect on student achievement. Charness and Gneezy (2009), on the other hand, found that financial incentives were effective in encouraging students to visit the gym.

The common element in each of these papers is that people are provided incentives for something that they usually would not expect payment from, and their behaviours compared to a control group who were not provided these incentives.[7] The notion of being paid in this context could potentially reduce people's intrinsic motivation, as it changes the nature of the task or activity, reducing their motivation to do what they were willing to do in the absence of financial incentives. We elaborate on this, and other possible explanations in Section 2.3.4.

[7] There are also papers that look at the effect of performance-based penalties to deter behaviour – the counterpart to incentives to encourage behaviour. See Gneezy and Rustichini (2000a) and Holmås, Kjerstad, Lurås, and Straume (2010). These studies find that performance-based penalties are counter-productive and do not have their desired effect.

2.3.2. Performance Incentives and Fixed Salaries

While the previous studies showed mixed effects when performance pay schemes were compared to instances when people were not paid, we can also look at the effect of performance pay schemes by comparing them to fixed salaries. Although the situation of no-pay can be considered an instance of a zero salary, we make a distinction between these since there are differences with respect to whether people expect to be paid or not. There is, therefore, a different reference point when the pay scheme changes or is introduced.

Lazear (2000) reported on the productivity improvements Safelite Glass Corporation realised when they decided to shift its employees from hourly wages to piece rates during 1994-5. The new piece rate paid workers for every windscreen they installed, while guaranteeing them a minimum wage if their weekly pay under piece rates fell short. Workers' productivity, measured by the number of windshields installed per day increased 44% after piece rates replaced the wages; half of this productivity gain was attributed to the incentives brought about by piece rates.

In a field experiment with Canadian tree-planters, Shearer (2004) finds that they are more productive under paid a piece rate than a fixed wage. The work requires physical effort: digging a hole, planting a seedling and filling the hole back. In the industry, workers are usually paid piece rates on the number of seedlings they plant, without any base wage. These piece rates are only made known to workers at the start of each work day, as it varies with the physical terrain of the land and how difficult it is to plant on. However, on occasion when unexpected circumstances arise, workers are paid a fixed wage for the day's work, with total remuneration similar to what can be earned under piece rates. In their experiment, the output of 9 male tree-planters was tracked over a number of days. Workers were assigned to work under piece rates or a fixed wage at the start of each planting day, such that they planted an equal number of days under each pay scheme. On average, workers planted 1256 trees per day under a piece rate, compared to 1037 trees under paid fixed wages. This represents 21% higher performance under piece rates.

A number of lab experiments have also looked at how performance pay schemes compare to salaries. Gneezy and Rustichini (2000b) reports on a real-effort experiment with students working on an IQ task. In their experimental control, participants were paid a flat $60. In other treatments, participants were paid a piece rate of 10c, $1 or $3 for each correct answer on top of the base $60. They found that those who were paid the 10c piece rate performed worse than the control group who only received the flat $60, while the treatment groups who were paid the $1 and $3 piece rate performed better than the baseline control. Similarly Bellemare, Lepage, and Shearer (2010) found that participants in a data entry task were more productive when they were paid a piece rate compared to when they were paid a salary.

These studies all show that piece rates are superior to salaries in motivating performance. The empirical findings regarding performance pay being introduced to people who were not previously paid is less clear, though there is evidence suggesting that the incentives need to be sufficiently high powered for the motivating effects to work.

2.3.3. Psychology Studies on Performance Pay Schemes

The effects of performance pay schemes have also been studied extensively in the field of psychology.[8] This literature is presented separately because the methodology and focus of these studies is different to those by economists. One key difference is the focus on measures of intrinsic motivation rather than on measures of performance. Conceptually, intrinsic motivation is closely related to the effort that people exert.

Most studies follow the paradigm laid out by Deci (1971), which considers the proportion of a participant's 'free choice' time they devote to the task as the primary measure of intrinsic motivation. In his study, student participants[9] were asked to reproduce configurations with various pieces of puzzle blocks for three rounds. In the control, participants were not paid for any of the rounds. In the treatment group, participants were paid a piece rate of $1 for each correct configuration in the second round; they were not paid for the first or third rounds. At the end of each round, the experimenter left the room for a couple of minutes, informing the participant that he will be in another room preparing materials for the experiment. The

[8] The psychology literature on intrinsic motivation is voluminous. The papers reviewed here are selective – see Deci et al. (1999) for an extensive meta-review of the literature.

[9] Unlike economic experiments, it is not the norm to pay participants in psychology experiments. The participants are usually university students enrolled in psychology courses who are required to participate in experiments to fulfil course requirements. In these studies, participants are usually experimented on individually rather than as a group.

participant was told he could do anything he wanted to, including reading the magazines or working on other puzzle configurations that were placed on the participant's desk beforehand. While the experimenter was away, he was actually observing the participant through a one-way window from another room, recording the duration the participant had spent working on the puzzles. The idea is that if the participant chooses to work on the puzzle in his 'free choice time' when he is not required to, and in the presence of alternative options such as reading a magazine, then he must be intrinsically motivated to work on the puzzle. Deci (1971) found that the proportion of free choice time that participants spent on the puzzle increased significantly after round 2, when piece rates were provided. The free choice time spent on the puzzle fell significantly after round 3, after the piece rate was removed, to a level even lower than in the first round. By comparison, the free time spent on the puzzle in the control group was higher after round 3 than after round 1.

It was unclear why intrinsic motivation increased immediately after piece rates were provided but fell below the pre-piece rate level after they were removed. Deci (1972) anecdotally noted from his 1971 study that participants who were paid felt they had received a lot of money for relatively little work. He posited that by receiving more money than they felt they deserved – being overpaid relative to their expectations – people would continue to perform even if the pay was discontinued[10], in order to correct the 'inequity' of being overpaid. This would not be the case if people were paid according to their expectations.

Deci (1972) followed up with a similar design. With the same task, his 1972 study was one-shot, consisting of a single round followed by the free choice period. Participants were either not paid at all, or paid a piece rate on configurations solved during the round. For the participants who were paid, payment was made either before the free choice period or afterwards. The experimental design with payment made before/after the free choice period was aimed to distinguish the motivating effects of incentives from that of equity restoration. The evaluation of overpayment was thought to occur at the moment that participants are paid, so those who

[10] This idea is similar to that of the Fair Wage Hypothesis (Akerlof & Yellen, 1990), that posits people will shirk if they are paid wages lower than their expectations such that the effective wage, relative to effort exerted, will be equalised to a level as if they were paid according to their expectation but did not shirk.

were paid after the free choice period would not have felt they were overpaid in the free choice period that preceded it. Indeed, participants who were paid the same piece rates before the free choice period were more intrinsically motivated than those who were paid afterwards, giving weight to the hypothesis that participants are more intrinsically motivated in order to reduce their feelings of guilt.

Those who were paid after the free choice period, who would not have felt they earned more than they deserved, were less intrinsically motivated than those who were not paid at all. This suggests that people's intrinsic motivation decreases when performance incentives are present, after ruling out people's concerns of pay equity.

Ryan, Mims, and Koestner (1983) employed a similar design looking at the effect of discrete performance based pay. Participants who were paid according to some performance criterion showed less intrinsic motivation than a similar group who received no pay. They also showed that participants who were paid salaries to be less motivated than those who were not paid at all. These results confirm those of Harackiewicz (1979), who rewarded participants with pens and a notebook instead of with money.

The studies cited here have been selected from, but is representative of, a large literature in psychology which shows that monetary and in-kind rewards undermines intrinsic motivation. See Cameron and Pierce (1994) and Deci et al. (1999) for opposing perspectives of the intrinsic motivation literature.

2.3.4. Theory and Explanations

It is interesting to compare the economics and psychology literatures. In economics, there is clear evidence that performance pay schemes improve performance over fixed salaries, while there is mixed evidence that performance pay schemes are successful in motivating performance when introduced to situations where people do not expect to be paid for their work. In the psychology literature, performance pay schemes have been shown to 'crowd out' intrinsic motivation, which presumably adversely affects performance. While this dichotomy is difficult to explain, we discuss some explanations as to why performance pay schemes might not always work to motivate performance.[11] In what follows, we discuss Cognitive Evaluation Theory, the Overjustification Hypothesis, and how image motivation and framing affects the effectiveness of performance pay schemes.

Cognitive Evaluation Theory

There are two key aspects to Cognitive Evaluation Theory (Deci & Ryan, 1985). It posits that intrinsic motivation is affected by underlying psychological needs for competency and autonomy. Events or actions that affect a person's perceptions of competency or autonomy will affect their intrinsic motivation. If someone is provided positive feedback that he is performing well in an activity, then he is made to feel more competent. According to Cognitive Evaluation Theory, he should be more intrinsically motivated to perform the activity. Conversely, negative feedback is expected to reduce intrinsic motivation. There is an emphasis on how informative interventions are in terms of how well someone is performing.

The other aspect of Cognitive Evaluation Theory is that of autonomy. People prefer to work free of constraints without being controlled by others or their environment. Drawing heavily from deCharms's (1968) notion of the locus of causality, people would rather be in control of their actions than be controlled by others – that their locus of causality is internal and not controlled by external factors. When people face deadlines, for example, they are usually less intrinsically motivated to work since they are considered to have a controlling effect on one's behaviour. Ryan et al. (1983) show that people are more intrinsically motivated when they are provided feedback of an informative nature than when they are provided feedback of a controlling nature. Falk and Kosfeld (2006) also provide empirical support that people are averse to control.

With the framework prescribed by Cognitive Evaluation Theory, pay schemes can be analysed in terms of their information content and how controlling they are perceived to be. Observing the amount earned, rank-order tournaments provide more information about an agent's

[11] See Frey and Jegen (2001), Gneezy, Meier, and Rey-Biel (2011) and Bowles and Polanía-Reyes (2012) for surveys on the role of performance pay schemes and intrinsic motivation from an economics perspective. From a psychology perspective, see Cameron and Pierce (1994) and Deci et al. (1999) for opposing perspectives. Bruno (2013), Promberger and Marteau (2013) and Festré and Garrouste (2015) for reviews of both economics and psychology literatures.

competency than discrete performance-based pay, since an agent would be able to infer their competency relative to others under tournaments but not under discrete performance pay. Discrete performance-based pay schemes are more informative than piece rates since agents are able to assess their level of competency against the performance thresholds under discrete pay, while piece rates do not provide such benchmark. Piece rates are in turn more informative than salaries since one cannot make out their level of competency based on pay that is invariant with performance.

In terms of control, the pay schemes that are most informative about performance also rank to be most controlling. Tournaments are most controlling since an agent is required to outperform others in order to earn more money, followed by discrete output based pay where agents need to outperform a predefined performance threshold. Piece rates are less controlling since there is no performance criteria to satisfy in order to earn more money. Salaries are least controlling as the flat pay structure means that there is no pressure to improve performance, since it will not increase pay.

Since pay schemes that are more informative about competency are also more controlling, Cognitive Evaluation Theory makes no solid predictions a priori about how various pay schemes affect intrinsic motivation. The overall effect on intrinsic motivation depends on both the relative strength of these effects as well as the emphasis an agent places on them.

Cognitive Evaluation Theory predicts that higher powered incentives should be more controlling while carrying the same competency feedback as lower incentives, so larger incentives would unambiguously be expected to lower intrinsic motivation. The theoretical model by James (2005) is consistent with this, suggesting that motivation crowding out worsens as the total level of compensation increases.

Overjustification Hypothesis

The Overjustification Hypothesis (Lepper et al., 1973) suggests that agents will reduce their level of intrinsic motivation in the presence of extrinsic rewards as the agent faces too many reasons to engage in the activity. In the absence of any incentives, an agent will perform the activity out of his innate desire to do so and he attributes his actions to it. If he is subsequently provided a monetary incentive to perform the same task, he places greater weight on the extrinsic incentive as his reason to perform the task, and lesser weight on his intrinsic motivation. This lower weight placed on intrinsic motivation is realised, resulting in the agent being less intrinsically motivated to perform in the presence of monetary incentives.

This can explain why people are less intrinsically motivated under fixed salaries than if they are not paid at all (Harackiewicz, 1979; Ryan et al., 1983). The mere fact that people are paid provides extra justification to perform over and above that when they are not paid to perform, crowding out intrinsic motivation. The Overjustification Hypothesis provides a better explanation of this result than Cognitive Evaluation Theory given that salaries carry no informative value about competency and are not considered to be controlling.

The Overjustification Hypothesis suggests that larger stakes and higher powered incentives have a stronger undermining effect as this amplifies the extent of overjustification. This contradicts the result from Gneezy and Rustichini (2000b) where small incentives are counterproductive while higher powered incentives induce performance.

Image Motivation

Bénabou and Tirole (2006) models motivation crowding out through the adverse effect incentives have on one's image and reputation. Workers are assumed to derive utility from three key sources: from money, from their innate desire to work and from their reputation. Individuals assign separate weights to each of these sources to reflect the relative importance of each. Intrinsic motivation is modelled in the utility function by assuming that workers derive utility from their own choice of effort, despite it also being costly to them. Reputation reflects how other people perceive the agent, for which the agent cares since he does not want to be negatively judged. Agents want to be seen as intrinsically motivated rather than greedy.

A key point made by Bénabou and Tirole is that the power of incentives provide information to outsiders allowing them to judge how intrinsically motivated or greedy agents are, which is private information to agents themselves. If people are observed to work in the absence of any financial reward, then they cannot be working out of greed, so they must either be working out of intrinsic motivation or for the reputational benefit they derive. When incentives are provided or if the power of the pay scheme increases, it becomes less clear to the outsider whether people are working for the monetary reward or because they are intrinsically motivated to do so. The weight outsiders place on greed increases while the weight on motivation decreases in the presence of higher incentives, diminishing the image they hold of the agent. Though agents benefit from receiving greater income, their reputation falls as they are also deemed more to be greedy. If agents place greater weight on their reputation than on monetary incentives, the provision of monetary incentives (or its increase) would lead to lower effort exerted by the agent.

The theory of image motivation receives empirical support from Ariely, Bracha, and Meier (2009), who conducted a lab experiment where participants took part in a real-effort key-pressing task. The higher the task performance by participants, the greater the donations that the experimenters pledged to charity. The donations were made according to a piece rate on each key-press, with the rate declining with higher output thresholds. The treatments varied across two dimensions: whether or not each participant also received individual incentives and whether or not the performance and payment individuals received were made public. For those who were paid, they were paid according to the same scheme that was applied for donations – the incentives received were separate to the donations. Participants in public condition had their performance level, donation amount and incentive receipts revealed to all other participants in the session, while all this was kept private in the private condition.

In the public condition, performance was higher when there were no personal incentives compared to when piece rates were paid to participants. On the other hand, performance was higher in the presence of personal incentives when pay and performance were kept private. The publicity component suggests that people care about how they are perceived by others, and reduces their performance in the presence of incentives to mitigate others' perception that they are greedy. Image concerns are irrelevant when all information is private, and participants are shown to be motivated by the incentives. This is consistent with the theory of image motivation.

Framing

Motivation crowding out can also be explained by changes in a person's mindset.[12] When monetary incentives are introduced, the task or action may be perceived differently and with a different context compared to when these incentives were not in place, and this may adversely affect intrinsic motivation and subsequently performance. The framing effect explains why a monetary penalty issued to late-arriving parents actually increased the incidence of late-arrivals at Israeli daycare centres, since the parents merely viewed the financial penalty as a price and treated lateness as a marketable good for which they were willing to pay (Gneezy & Rustichini, 2000a).

The framing effect can explain the phenomena Bowles and Polanía-Reyes (2012) refer to as the 'categorical effect' of incentives, where low powered incentives are inadequate to compensate for the reduction of intrinsic motivation. In Gneezy and Rustichini (2000b) the introduction of a piece rate reduces the amount volunteers collect when the stakes are small, but only begins to induce performance as the power of the incentives increases. The initial reduction in performance can be attributable to framing as volunteers reconsider the nature of volunteer work and reduce their motivation to collect money. Only when incentives are sufficiently large do they begin to outweigh the negative categorical effect.

2.3.5. *Relative Performance Pay Schemes*

Having discussed the motivating effects of performance pay schemes, we now look at how pay schemes that depend on relative performance affect a workers' performance. Earlier on we discussed the Piece Rate Equivalence property of tournaments, where tournaments elicit the same level of effort as piece rates do (Lazear & Rosen, 1981). Bull et al. (1987) verified this claim in a lab experiment with the caveat that effort choices under tournaments are more variable than under piece rates. This section reviews studies which show the motivating effects of the relative incentives inherent in tournaments[13] and continuous relative performance pay schemes.

[12] See Bowles and Polanía-Reyes (2012) for an elaborate discussion and review.

[13] See Dechenaux, Kovenock, and Sheremeta (2015) for a survey of experimental research on tournaments.

The incremental nature of rank incentives have been shown to induce performance. Eriksson (1999) showed with Danish personnel data that various features of the corporate hierarchy are consistent with a tournament. He found that a larger prize spread improves firm performance, presumably from higher effort put forth by individual workers. Rank-based incentives have also been shown to induce behaviour in sport contests. Ehrenberg and Bognanno (1990) showed that a larger prize spread motivates performance in golf tournaments. Knoeber and Thurman (1994) has shown with data for chicken growers, who are paid tournaments, that increases in the magnitude of the rank prizes have no effect on performance if the prize spread if left unchanged.

On the other hand, Delfgaauw, Dur, Non, and Verbeke (2014) do not find that tournaments have any effect. Tournament bonuses were offered to every employee and manager of treatment stores in a retail chain. These bonuses were awarded if the retail store had higher sales growth than three other stores, who served as the control and were not offered these bonuses. The authors found that these bonuses had no effect on the sales of treatment stores over and above that of the control stores.

We now look at studies which compare relative performance pay schemes with other pay schemes. Comparing the performance of tournaments and a fixed salary in a real-effort decoding task, Masclet, Peterle, and Larribeau (2015) find that better performance is attained under tournaments.

Comparing a continuous relative performance pay scheme to piece rates, Bandiera et al. (2005) find piece rates induce better performance from U.K. fruit pickers. Under the relative pay scheme, workers received higher pay with higher individual performance, but was penalised by higher average performance of all workers. With the introduction of the piece rate, despite being approximately 12% lower on a per unit basis than the continuous relative pay scheme, the quantity of fruit picked increased by between 50-70%. The authors attribute a large portion of this effect to workers withholding performance under the relative pay scheme.

2.4. Productivity Spillovers and Monitoring

Most of the prior emphasis in this literature review has been on the effects of pay schemes when production externalities are absent. As previously mentioned, relative performance pay schemes exhibit negative production externalities where the unilateral increase in an agent's performance will reduce the expected payoffs of others. Productivity spillovers also occur in other circumstances, and quite often the externality affects others favourably rather than adversely. For example positive productivity spillovers occur when pay schemes depend on the joint production of many agents (for example see Fryer, 2013), under profit sharing arrangements (Kandel & Lazear, 1992), and are sometimes inherent in the job itself (Mas & Moretti, 2009). When the spillovers are positive, the agent only receives a portion of the aggregate benefit derived from a marginal increase in output – since it is shared with other agents – while incurring marginal costs in full.[14] Each agent have the incentive to free-ride off the output of others, reducing it below the efficient level.

If others are able to monitor the performance of the agent, the imposition of appropriate sanctions may be able to influence the performance of the agent. This is the essence behind Kandel and Lazear's (1992) model of peer pressure, where they find that agents' use of peer pressure as a means of social sanction is effective in manipulating the performance of their peers.

In Bandiera et al.'s (2005) study of fruit pickers, they attributed the low performance of relative pay schemes to workers' internalisation of the negative externality that they impose onto others. Workers are thought to restrict performance so not to reduce the rate of pay that their peers receive. For a particular type of fruit whereby the tall shrubs prevent pickers from observing the performance of their peers, they find that there are no differences in the performance under piece rates and relative pay. This suggests that peer monitoring is effective in influencing others' performance.

Peer monitoring has also been found to be effective in affecting performance when positive production externalities are present. Mas and Moretti (2009) found that supermarket checkout operators became more productive as higher productivity workers arrive on shift. The nature of checkout work is such that if one shirks, they impose a higher workload onto other workers; conversely, someone working harder reduces the workload of others. The productivity of workers improved when they were being observed by their co-workers, being in their line of sight –

[14] This is analogous to the problem of public goods provision.

suggesting that peer pressure is at play. Workers who monitored others but were not being monitored themselves were not subjected to peer pressure and their performance did not increase when higher productivity workers entered a shift.

In Falk and Ichino (2006), experimental student workers were asked to stuff envelopes, either alone in a room by themselves or in the presence of a partner. The task was individual and workers were paid a fixed wage for their work. The authors found that those who worked alone had lower performance than those who worked in the presence of a partner. Furthermore, the variation of output within each pair is lower than between different pairs, suggesting that each partner mutually motivated the other to perform at a similar level. This suggests that peer effects comes about from a desire to conform to the actions of others.[15]

2.5. *Decomposition of Tournament Effects*

2.5.1. *Relative Performance Feedback*

The literature review has so far focused on the motivating effects associated with different pay schemes – relevant to our first research question. We now review papers that are related to our second research question about the effects of relative performance feedback and rank-dependent payoffs. We first focus on papers that investigate what effect the provision of relative performance feedback has on performance. Relative performance feedback is used to invoke competition between workers, independently from payoffs.

Relative performance feedback may affect performance through two distinct channels. The first is where people improve their performance in order to increase their chances of winning, where they derive utility from being compared favourably with others relative to unfavourable comparisons.[16] For ease of exposition, we refer to favourable comparisons as 'winning' and the converse as 'losing'. In a tournament setting, Kräkel (2008) modelled the emotions associated with winning and losing a tournament. He shows that the desire to win increases effort, regardless

[15] See also Kuhn, Kooreman, Soetevent, and Kapteyn (2011) and Card, Mas, Moretti, and Saez (2012).

[16] The desire to win may be derived from preferences for status and respect. See Ellingsen and Johannesson (2007) for a review. See also Kosfeld and Neckermann (2011).

of whether one experiences positive emotions from winning or negative emotions from losing. While winning increases utility, agents also want to exert more effort in order to avoid losing, since losing reduces it. Similar to how the prize spread induces effort in tournament theory, emotions increase the benefit from winning relative to losing, making the perceived prizes larger than the actual monetary prizes, in turn incentivising agents to increase their effort. Since this first channel reflects a person's innate desire to win, the motivating effect would take place even before the participant has received any feedback on relative performance feedback.

In contrast, the second channel which relative performance feedback could influence performance occur after the feedback has been received. Cognitive Evaluation Theory (Deci & Ryan, 1985) suggests that relative performance feedback will improve a person's intrinsic motivation, and therefore performance. Relative performance feedback will provide richer information about an agent's performance vis à vis others which should improve their feelings of competency if the feedback is favourable. Unfavourable relative feedback would be expected to reduce perceptions of competency, in turn reducing intrinsic motivation and performance.

Many studies have found that the provision of relative feedback to agents leads to higher performance. A number of experimental studies (Hannan, Krishnan, & Newman, 2008; Cadsby, Engle-Warnick, Fang, & Song, 2010; Kuhnen & Tymula, 2012; Charness, Masclet, & Villeval, 2014; Azmat & Iriberri, 2016) have found that different forms of relative feedback improve performance. In the field, Azmat and Iriberri (2010) and Tran and Zeckhauser (2012) find that students' academic performance improves when relative performance feedback is provided, allowing students to compare their academic performance to others in their class. Blanes i Vidal and Nossol (2011) found that the productivity of German warehouse workers improved when they were provided rank information in their payslips. The authors find evidence for both anticipation and revelation effects, where worker productivity improved both in the post-announcement period before feedback was provided, and in the period after initial feedback has been provided.

There are, however, a few papers that find relative performance feedback to have no effect on performance. In a lab experiment, Eriksson, Poulsen, and Villeval (2009) provided either discrete or continuous feedback to participants working on a number-adding task. Participants were paid either piece rates or tournaments. The authors found that, regardless of how they were paid, players who received either form of relative feedback did not perform differently to their counterparts who did not receive any relative feedback.

Several papers have found relative performance feedback to have mixed effects, depending on the pay scheme. Hannan et al. (2008) find that relative feedback improves performance under piece rates, but reduces performance under tournaments. In Bellemare et al. (2010), relative feedback has no effect under piece rates but reduces performance under fixed salaries. Azmat and Iriberri (2016) find that relative performance feedback improves performance under piece rates, while having no effect under salaries. Since there are no consistent results from these studies, it is unclear what effect different pay schemes have on the effectiveness of relative performance feedback.

This literature, by and large, shows that relative performance feedback improves performance. The literature provides several insights into how relative feedback works to improve performance. The first insight is that it makes no difference whether feedback is delivered publicly or privately (Tran & Zeckhauser, 2012; Cadsby et al., 2010). This implies that relative performance feedback improves performance not necessarily because agents strive for status (Ellingsen & Johannesson, 2007) or external recognition (Bénabou & Tirole, 2006), since external recognition would require the revelation of one's relative position to others. We posit that the effect of relative performance feedback is driven by an innate desire to do better than others. The utility attained from doing better than others is independent to the utility attained from 'showing off' this piece of information to peers.

A second insight is that agents respond not only to the relative performance feedback after its release, but also in the period leading up to the release of this information (Blanes i Vidal & Nossol, 2011). Knowledge that relative feedback will be provided at some point in the future is sufficient to motivate performance, even before any feedback has been provided. The anticipation of relative feedback is sufficient to motivate people to improve their performance, as they strive for a favourable comparison with others. Kuhnen and Tymula (2012) reinforces this point when they found that players who were merely told that they *might* receive rank feedback performed better than those who were told they will definitely not receive such feedback. This suggests that agents are motivated by the desire to do better than their peers and will exert additional effort attempting to improve their relative performance, even before any information can be extracted from the feedback itself.

Few studies mentioned so far look at the effect of favourable and unfavourable feedback. Studies by psychologists shed some light, whereby they manipulate the feedback that is provided.[17] Harackiewicz (1979) looked at the effect of favourable relative performance feedback on various measures of intrinsic motivation, including the amount of time participants spent on the task during a free choice period. The favourable feedback that was provided told participants that they performed better than the average high school student. This comparison was chosen to be artificially low so that all participants would receive favourable feedback. Participants who received favourable feedback were either not rewarded at all, or were rewarded markers and a notebook for participation in the experiment. Intrinsic motivation was higher amongst those who received favourable relative feedback compared to groups who did not receive any feedback at all, whether they were rewarded or not. In similar fashion, Epstein and Harackiewicz (1992) found that favourable relative feedback improved participants' sense of competency compared to those who received unfavourable feedback. Reeve and Deci (1996) reported higher intrinsic motivation amongst participants who received favourable feedback compared to those who received unfavourable feedback. These findings are consistent with Cognitive Evaluation Theory.

2.5.2. Rank-Dependent Payoffs of Tournaments

We now review papers that look at the motivating effects associated with rank-dependent payoffs. Here we are mainly interested in the tournament pay mechanism. In order to study the effect of rank-dependent payoffs inherent in tournaments, it is necessary to control for the relative performance feedback that also feature in tournaments. While there are many papers which focus on the effect of relative performance feedback, there are few papers that focus on the rank-dependent payoffs characteristic of tournaments. To our knowledge, there are only two papers that distinguish and decompose these two effects.

[17] In these studies, the feedback that is provided is exogenously determined by the experimenters, and is not necessarily true.

Eriksson et al. (2009) conducted a lab experiment where they paid participants either piece rates or tournaments, and provided them with either relative feedback or not. In the absence of relative performance feedback, piece rates and tournaments perform similarly to one another – supporting the property of Piece Rate Equivalence. Relative performance feedback is found to have no effect on performance. Proceeding with the decomposition, the authors find that tournaments perform no differently to piece rates when relative feedback is controlled for. This suggests that the rank-dependent payoffs of tournaments serve as the main motivator, with relative feedback playing no role.

Hannan et al. (2008) conduct a similar experiment, whereby they decompose the effects of relative feedback and payoffs. Their findings are, however, quite different to those of Eriksson et al. (2009). When relative feedback is absent, tournaments outperformed piece rates. Relative performance feedback improves performance under piece rates, but reduces performance under tournaments. When relative performance feedback is controlled for, they find that piece rates perform better than tournaments. In Hannan et al's study, relative feedback plays a greater role in tournaments than payoffs do.

2.6. Learning

Our third research question focuses on the dynamics of learning – how do different pay schemes and relative performance feedback affect learning? The process of learning is multi-dimensional: involving heuristics and rules (Roth & Erev, 1995; Erev & Roth, 1998; Charness & Levin, 2005), feedback (Rick & Weber, 2010) and observation (Merlo & Schotter, 2003; Cardella, 2012), and payoffs (Merlo & Schotter, 1999). To our knowledge, however, there has been no prior research on how pay schemes affect learning.

In terms of how relative performance feedback affects learning, previous papers which study the effect of relative performance feedback have not focused on learning. To our knowledge, only Kuhnen and Tymula (2012) and Azmat and Iriberri (2016) mention learning in their papers. In both these papers, the authors study the effect of relative feedback in an arithmetic task across a number of rounds. In analyses, the authors of these papers control for a linear time trend and find that, in general, learning occurs over time. However, the learning here is common for all participants and not disaggregated by treatment – so we do not know how learning is affected by relative feedback.

2.7. Contribution to Literature

In this literature review, we have discussed many papers that relate to pay schemes, relative performance feedback and learning. While some areas of research has reached a consensus on findings, other areas of research remain contentious, whereby the results of many papers do not conform to those from other papers. For example, there is much disagreement amongst economists as to whether performance pay schemes are effective in motivating performance and whether motivation crowding out exists or not. With respect to the literature on relative performance feedback, while the majority of studies show that the provision of relative feedback improves performance, a non-negligible number of papers also show that relative feedback either has no effect or is counterproductive. As such, it is difficult to interpret the literature as a whole and there remains a number of key questions that have been left unanswered.

Although some of our research questions have been studied before by others, we nevertheless add to the literature in several ways. First, we are able to test the robustness of previous findings to see whether or not they are replicable. Second and more importantly, we look into the underlying channels that influence people's motivation to perform under different pay schemes or in the presence of relative performance feedback.

We utilise controlled lab experiments to investigate these underlying effects.[18] Laboratory experiments facilitate us in isolating various elements that could possibly affect the results. We further facilitate the study of underlying mechanics by asking participants to fill out questionnaires which elicit various factors such as their level of interest, effort, and anxiety associated with the experimental task. These factors allow us to narrow down the core effects that are at play.

[18] See Falk and Fehr (2003) for a discussion of the merits of lab experiments as a means to study labour market issues.

One of the main effects that we postulate to play a crucial role is Cognitive Evaluation Theory. While it has been studied widely in the field of psychology, previous studies have focused on how the two dimensions of control and competency associated with pay schemes have affected people's intrinsic motivation, rather than their performance. Our study fills this void by focusing on performance.

Our experiment utilises a cognitively difficult real-effort task which replicates the work environment. Participants work on a stock forecasting task, and are paid according to various pay schemes. In some treatments, participants are provided information about whether they performed better than a random partner or not. By manipulating different pay schemes and relative feedback, we are able to answer our three research questions. The details of the experiment are provided in the following chapter.

To our knowledge, our experiment is unique in using a cognitively challenging task. The experimental studies that we have reviewed use tasks that require little thought and cognition. These tasks – such as simple arithmetic, key-pressing or data entry – are mechanical in nature and require little skill. By contrast, our stock forecasting task is difficult. To do well in our task, players need to exert cognitive effort and process the various pieces of information provided to them to assist them to make their forecast. Our task is reminiscent of white-collar jobs, where the skills of reasoning and problem-solving are valued.

Since our task is difficult, and is repeated over a number of rounds, we are able to also study how pay schemes and feedback affect people's learning over time. There has been little work done in this area. Those prior studies that have indeed tried to study learning in the context of pay schemes and relative feedback do not look at learning in a meaningful way due to their choice of experimental task. Similar to the point made in the previous paragraph, there is nothing to 'learn' per se in simple mechanical tasks.

3. Experimental Design

3.1. The Task

This thesis uses an experiment that is based on a multiple cue probabilistic learning task19, where participants are required to determine the value of a variable x based on the observation of two numerical cues provided to the participant. This variable x can be thought of as the underlying price of a stock; the cues as variables that affect the value of the stock; and the task at hand as one of forecasting stock prices. The actual value of the stock is determined by the underlying equation:

$$x_t = 10 + 0.3 \times Cue\ A_t + 0.7 \times Cue\ B_t + \varepsilon_t \tag{3.1}$$

where x_t is the actual value of the stock participants are required to predict, $Cue\ A_t$ and $Cue\ B_t$ are the values of the two numerical cues provided to the participant, and ε_t is a random variable which is uniformly, but discretely, distributed within $[-5, 5]$ in round t. Although participants know that the actual stock's value is determined by an underlying relationship with cue values for every round, participants do not know anything about this relationship, including its functional form.

Two variants of the task are used in the experiment. In one of these, Cue A is fixed at the value of 150 for each of the 20 rounds, while Cue B changes every round. This 'single cue' task is designed to be less difficult than the 'dual cue' task, where both cue values change round by round. Both the single cue and dual cue tasks are employed in parallel to determine treatment effects across different levels of task difficulty. The cues, though randomly determined, are predefined for each of the single and dual cue tasks. As such, the cue values for a particular round are identical for each participant in each treatment for the respective task difficulty, facilitating direct comparison of performance across treatments.

Participants are given 5 minutes to study ten examples of cue values and actual stock prices prior to the first round. The examples for each version of the task are shown in Table 3.1. This

[19] Multiple cue probability learning tasks are common in the field of psychology to study cognitive learning and reinforcement (see Balzer, Doherty, & O'Connor, 1989). In economics, see Brown (1995, 1998), Vandegrift and Brown (2003, 2005) and Vandegrift, Yavas, and Brown (2007) for examples of the multiple cue probabilistic learning task used in experiments.

provides them with an opportunity to familiarise themselves with the task at hand. The provision of these examples also mitigates the effect that 'wild guesses' have on early rounds of play.

In each of the rounds, having observed the cues presented to them, participants have 90 seconds to submit their forecast of the stock's price. After all participants have entered their forecasts, they are presented with a table of information, including the values of Cue A and Cue B, the underlying actual value that corresponds with the cues, their submitted forecast, the absolute forecast error and earnings for the prior round, as well as a tally of their cumulative earnings. The absolute forecast error (hereafter, forecast error) is the absolute difference between the actual stock value and that forecasted by the participant. The forecast error measures the accuracy of the forecast, and serves as our primary measure of performance.

Table 3.1 Example Cues

Single Cue Task			Dual Cue Task		
Cue A	Cue B	Actual Price	Cue A	Cue B	Actual Price
150	92	117	12	64	54
150	143	157	372	63	162
150	379	321	179	109	137
150	373	313	415	146	240
150	240	220	116	186	175
150	285	256	355	223	275
150	187	188	145	286	255
150	143	153	199	356	317
150	191	185	439	354	372
150	361	311	73	442	345

3.2. *Pay schemes*

All participants are guaranteed to earn at least a $5 show-up fee for their participation.[20] In addition to this, they will earn money based on piece rates, tournaments or salaries depending on

[20] All dollar values are expressed in New Zealand Dollars.

the treatment they are in. All earnings are accrued in an earnings account and the balance is paid out in cash at the end of the session.

Piece rates pay people proportionately according to their performance. In each round, each participants' piece rate earnings are calculated to be $1 minus their forecast error (expressed as cents). For example, for a forecast error of 18 in a particular round, the corresponding piece rate earnings is $0.82 (1.00 – 0.18). Greater performance, represented by lower forecast errors, increases monetary earnings. A participant who makes a perfect forecast, with forecast error of zero, would earn $1 for the round. If the forecast error is greater than 100, earnings are boosted up to zero to avoid negative earnings.

Tournaments pay people according to relative performance. Since tournaments require participants to be ranked amongst others, the performance of each participant in each round will be benchmarked against a random and anonymous partner, who is rematched every round. The better performer amongst each pair receives $1 while his counterpart receives nothing in the round.[21] This scheme is a simple two player winner-takes-all tournament.

Salaries pay people a flat amount that is independent of performance. Participants are told upfront that they will receive $20 for their participation before they begin the task, which is to be paid out at the end of the session. This amount includes the $5 show-up fee.

These three pay schemes are inherently different in nature, and they accordingly differ in terms of the hypothetical maximum and minimum that could be earned under each pay scheme. For example, the least that can be earned in any particular round is zero under piece rates (if forecast errors are 100 or greater) and tournaments (if participants lose), while earnings are guaranteed in salaries to be an equivalent of $0.75 ($20 guaranteed payment, less $5 show up fee, divided by 20 rounds). Despite these inherent differences, we facilitate comparison of these three pay schemes by calibrating the average amount that is earned under each scheme. As a result, as

[21] If the forecast errors for a particular pair are identical, then the winner and loser is determined randomly.

can be seen in Table 3.8, the average earnings in each pay scheme are not statistically different from one another, in the vicinity of $20.

3.3. Treatments

For each version of the forecasting task, the experiment consists of five treatments that differ in terms of how participants are paid and the feedback that is provided to them. In all treatments other than the Salary treatment, participants are paid piece rates on their forecast errors for each of the first five rounds of the 20 round game.[22] These first five rounds allow us to benchmark participants' underlying ability. After each round they are provided feedback on how they have performed, getting to see the actual stock price, the corresponding forecast error and earnings for that round. Participants are told that there might be a change in how the game is played in rounds 6 to 20. After round 5, participants are informed whether or not there are any changes to how the game is to be played thereafter.

3.3.1. Piece Rate Treatment

In the Piece Rate treatment participants are told that there is no change in how the game is played beyond round 5. Participants continue to receive piece rates on their forecasts for each round and there is no change to the feedback that they are provided with at the end of each round. Table 3.2 shows a hypothetical example, based on hypothetical forecasts, of the information a player in the single cue Piece Rate treatment would observe at the end of round 10.

[22] In the Salary treatment, participants are told at the start of the experiment that they will be paid a flat $20 for their participation. Piece rates therefore cannot be applied to the first five rounds. Though participants in the Salary treatment are not paid piece rates for the first five rounds, they see the same feedback – including piece rate earnings – as participants would in any other treatment.

Table 3.2 Onscreen Information in Piece Rate Treatment

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round
1	150	201	179	192	13	$0.87
2	150	263	221	243	22	$0.78
3	150	88	131	117	14	$0.86
4	150	248	219	232	13	$0.87
5	150	201	197	200	3	$0.97
6	150	196	182	194	12	$0.88
7	150	353	298	305	7	$0.93
8	150	173	169	173	4	$0.96
9	150	270	231	248	17	$0.83
10	150	243	233	222	11	$0.89

3.3.2. Piece Rate Win Lose Treatment

The Piece Rate Win Lose treatment also pays participants piece rates after round 5, but additional information is provided to participants after each round regarding how they have performed compared to a random partner. They are randomly and anonymously paired and subsequently re-matched every round. The participant with the smaller forecast error will receive the feedback of 'win' while his partner will receive feedback of 'lose'[23] – an example is shown in Table 3.3. Participants are not provided information about the forecast errors of their partners, so they do not know the margin for which they win or lose by. Irrespective of whether a participant wins or loses, their round earnings are based on piece rates. In other words, the winning/losing feedback is decoupled from monetary payoffs.

This additional winning/losing information brings about a sense of competition. In the Piece Rate Win Lose treatment, participants' earnings after round 5 are independent of the outcome of winning or losing – contrasting with tournaments where earnings depend solely on whether one wins or loses. If competition that is independent of pay has an effect on performance, then

[23] If the forecast errors are identical, then the tie is broken by random draw.

this competition for rank and status plays a role in motivating people to perform. Such competition refers solely to people's innate desire for favourable comparison, so they would be expected to exert more effort and increase their performance in order to improve their prospects of winning.

We randomly rematch participants so that they are able to play with people of varying ability throughout the experiment, though they are blind to who they are matched with. This keeps the spirit of competition alive. If a low ability participant is matched with the same superior opponent over the course of session, he might decide to give up if he believes he will not be able to beat his opponent. Similarly, the superior opponent might choose to exert less effort since he can expect to win easily while matched with an inferior opponent. Random rematching alleviates this scenario – which we later refer to as 'bifurcation' – and spurs competition over time as people are matched with different people. We will elaborate on these dynamics in Chapter 6.

While both Piece Rate and Piece Rate Win Lose treatments pay participants piece rates after round 5, the difference lies in the additional winning/losing information that is displayed to participants in the Piece Rate Win Lose treatment. Since both treatments apply the same pay scheme, any observable differences in the performance of participants would be attributable to the winning/losing feedback in the Piece Rate Win Lose treatment, suggesting that psychological competition influences performance.

Table 3.3 Onscreen Information in Piece Rate Win Lose Treatment

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round	Win or Lose
1	150	201	179	192	13	$0.87	
2	150	263	221	243	22	$0.78	
3	150	88	131	117	14	$0.86	
4	150	248	219	232	13	$0.87	
5	150	201	197	200	3	$0.97	
6	150	196	182	194	12	$0.88	LOSE
7	150	353	298	305	7	$0.93	WIN
8	150	173	169	173	4	$0.96	WIN
9	150	270	231	248	17	$0.83	LOSE
10	150	243	233	222	11	$0.89	WIN

3.3.3. Tournament Treatment

The Tournament treatment is based on a winner-takes-all tournament with a large prize for the winner and a prize of zero for the loser. After round 5, participants are randomly paired with someone else and are paid fixed amounts depending on how participants have performed relative to their partners. At the end of each round, participants are provided the same winning/losing information as in the Piece Rate Win Lose treatment, but their earnings depend on their performance relative to their partner. Participants either earn $1 if their forecasts are more accurate than their partner, or they do not earn anything for the round if their forecasts are worse. Due to the all-or-nothing nature of the rank-dependent payoffs, all participants receive an extra $4 after round 5 in order to align total earnings closer to the $20 average that was announced during the recruitment process. Table 3.4 shows an example of the feedback players receive. Notice that when the tournament scheme kicks in from round 6, earnings are no longer paid according to piece rates but are now tied to winning or losing.

Table 3.4 Onscreen Information in the Tournament Treatment

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round	Win or Lose
1	150	201	179	192	13	$0.87	
2	150	263	221	243	22	$0.78	
3	150	88	131	117	14	$0.86	
4	150	248	219	232	13	$0.87	
5	150	201	197	200	3	$0.97	
6	150	196	182	194	12	$0.00	LOSE
7	150	353	298	305	7	$1.00	WIN
8	150	173	169	173	4	$1.00	WIN
9	150	270	231	248	17	$0.00	LOSE
10	150	243	233	222	11	$1.00	WIN

Both Piece Rate Win Lose and Tournament treatments simulate competition by allowing participants to learn how they perform relative to a random partner, though they differ in terms of how participants are paid. The Piece Rate Win Lose treatment pays piece rates which do not depend on how participants perform relative to their partners. On the other hand, the

Tournament treatment pays participants based on the outcome of the competition. Differences that are observed across these treatments therefore are due to differences in the payment scheme.

3.3.4. Tournament-No-Info Treatment

The Tournament-No-Info treatment is also based upon a winner-takes-all tournament after round 5. Participants are paired and subsequently rematched at the end of each round. They earn either $1 if they perform better than their partner or nothing if they do worse. After each round participants are informed only of their forecast errors and are unaware of how they have performed relative to their partners. This means that winning/losing feedback and the associated earnings feedback is withheld over rounds 6 to 20. The information pertaining to relative standing are released to players at the end of the game at the end of round 20, meaning that participants only know of their absolute performance and not their relative performance whilst they are engaged in the tournament. Like the Tournament treatment, all participants are paid an extra $4 after round 5.

Panel A of Table 3.5 shows an example of the information available to Tournament-No-Info participants. After round 5 when the tournament scheme applies, participants are oblivious to how well they are doing compared to their partners and do not know their earnings after each round. After round 20 participants see a summary of their performance, including the previously omitted winning/losing and earnings information for all rounds: Panel B shows an example of this.

Both Tournament and Tournament-No-Info treatments pay participants the same way but differ in terms of the feedback that is provided. The Tournament treatment notifies participants of their relative performance after every round, while the Tournament-No-Info treatment only releases this information after the final round. Any differences in participants forecast errors are attributable to the role that relative performance feedback has on participants. It will be particularly interesting to see how learning is affected in the Tournament-No-Info treatment when participants are blind to their earnings and how they perform relative to others.

Table 3.5 Onscreen Information in the Tournament-No-Info Treatment

Panel A: Before Round 20

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round	Win or Lose
1	150	201	179	192	13	$0.87	
2	150	263	221	243	22	$0.78	
3	150	88	131	117	14	$0.86	
4	150	248	219	232	13	$0.87	
5	150	201	197	200	3	$0.97	
6	150	196	182	194	12		
7	150	353	298	305	7		
8	150	173	169	173	4		
9	150	270	231	248	17		
10	150	243	233	222	11		

Panel B: After Round 20

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round	Win or Lose
1	150	201	179	192	13	$0.87	
2	150	263	221	243	22	$0.78	
3	150	88	131	117	14	$0.86	
4	150	248	219	232	13	$0.87	
5	150	201	197	200	3	$0.97	
6	150	196	182	194	12	$0.00	LOSE
7	150	353	298	305	7	$1.00	WIN
8	150	173	169	173	4	$1.00	WIN
9	150	270	231	248	17	$0.00	LOSE
10	150	243	233	222	11	$1.00	WIN
11	150	60	105	102	3	$1.00	WIN
12	150	320	256	274	18	$0.00	LOSE
13	150	340	304	289	15	$1.00	WIN
14	150	361	287	311	24	$0.00	LOSE
15	150	321	280	285	5	$1.00	WIN
16	150	361	311	309	2	$1.00	WIN
17	150	148	149	155	6	$0.00	LOSE
18	150	309	246	275	29	$0.00	LOSE
19	150	135	143	145	2	$1.00	WIN
20	150	142	147	156	9	$0.00	LOSE

3.3.5. Salary Treatment

The Salary treatment offers participants a flat payment of $20, including the $5 show up fee, for their participation in the experiment. This is announced at the start of the experiment before they begin the task. Although earnings do not depend on performance, at the end of each round participants are shown their forecast errors and feedback on earnings as if they were paid piece rates.

We have chosen to provide piece rate earnings information to Salary participants so we are able to perfectly match feedback with that present in the Piece Rate treatment, given that the main comparison of the Salary treatment is with the Piece Rate treatment.

Prior research has found that feedback on absolute performance itself motivates people to perform (Bandiera, Larcinese, & Rasul, 2015). Since piece rates are based on performance, information about piece rate earnings makes a participant's performance more salient, providing a greater source of motivation to those who are paid piece rates. If such piece rate earnings is not controlled for in the Salary treatment, this effect would impede our analyses of the effect of extrinsic incentives.

The Piece Rate and Salary treatments both provide the same information to participants, but varies by payment scheme. Piece rates pay participants for higher performance while salaries are performance-invariant. If participants are primarily motivated by money, those in the Piece Rate treatment would be expected to perform better than those in the Salary treatment. If participants in the Salary treatment performs at least as well as those in the Piece Rate treatment, then it can be inferred that participants in the Salary treatment are intrinsically motivated by factors other than financial rewards.

Each of these five treatments differ in terms of pay scheme and whether or not winning/losing feedback is provided. These treatments are applied to both the single and dual cue tasks. The differing levels of task difficulty allows for analysis of how task difficulty affects competitiveness and intrinsic motivation. The different treatments are summarised in Table 3.7.

Only one session of the Tournament-No-Info treatment was conducted with the single cue task. There did not appear to be differences in forecast errors between the Tournament and

Tournament-No-Info treatments in the dual cue task, and also in the single cue task with the data for the sole session of Tournament-No-Info treatment. As a result, further single cue Tournament-No-Info treatments were not conducted due to financial constraints. Due to the small sample size from one session, analyses hereafter will not include the single cue Tournament-No-Info treatment. We will present analyses of the dual cue Tournament-No-Info treatment separately from the other treatments.

Table 3.6 Onscreen Information in the Salary Treatment

Round	Cue A	Cue B	Forecast	Actual Price	Forecast Error	Earnings this Round
1	150	201	179	192	13	$0.87
2	150	263	221	243	22	$0.78
3	150	88	131	117	14	$0.86
4	150	248	219	232	13	$0.87
5	150	201	197	200	3	$0.97
6	150	196	182	194	12	$0.88
7	150	353	298	305	7	$0.93
8	150	173	169	173	4	$0.96
9	150	270	231	248	17	$0.83
10	150	243	233	222	11	$0.89

For the purpose of exposition, we will abbreviate the treatment names. PR denotes the Piece Rate treatment; PRWL for the Piece Rate Win Lose treatment; T for the Tournament treatment; TNI for the Tournament-No-Info treatment; and S for the Salary treatment. In what follows, only the treatments are abbreviated. When we refer to pay schemes generically rather than the treatment itself, we will write it out in full.

Table 3.7 Summary of Treatments

Single Cue Task

	Payoffs	Relative Feedback	Number of Sessions	Number of Subjects
Piece Rate (PR)	Piece rate	No	2	42[1]
Piece Rate Win Lose (PRWL)	Piece rate	Yes	2	42
Tournament (T)	Tournament	Yes	2	40
Tournament-No-Info (TNI)	Tournament	No	1	20
Salary (S)	Salary	No	2	42

Dual Cue Task

	Payoffs	Relative Feedback	Number of Sessions	Number of Subjects
Piece Rate (PR)	Piece Rate	No	2	39[2]
Piece Rate Win Lose (PRWL)	Piece Rate	Yes	2	35[3]
Tournament (T)	Tournament	Yes	2	38
Tournament-No-Info (TNI)	Tournament	No	2	38
Salary (S)	Salary	No	2	34

1. There was initially a third single cue Piece Rate session with 22 participants. We decided not to include this session in the analyses because a glitch in the experimental software meant that a) there are only 17 rounds of data in this session and b) the data for one participant is missing. The results do not vary substantially whether this session is included in analyses or not.
2. We recruited 20 people for a particular session, but one person did not show up.
3. For reasons out of our control, a participant had left halfway through a session. We do not use data pertaining to this participant.

3.4. Questionnaires

Two questionnaires were distributed to participants over the course of the experiment: one prior to the task and the other afterwards. The main aim of the questionnaires is to elicit psychological variables, many relating to intrinsic motivation, which can be used to complement forecast errors as our central measure of performance. Questionnaires of this nature are quite common in the field of psychology.

The pre-task questionnaire (see Appendix A2.6) is used to elicit each participants' trait anxiety level. Trait anxiety measures how prone people are to stress and situations that make them anxious. It differs from state anxiety in that it is not situation- or context-specific, but rather is specific to individuals like a personality trait. This is why the trait anxiety is elicited in the pre-task questionnaire before the forecasting task.

We elicit trait anxiety as a means to control for differences in the competitiveness of participants in our sample. Exogenous differences in how competitive participants are could affect their performance in some of our treatments – particularly those in the PRWL, T and TNI treatments that feature an element of competition.[24] For example, those who have performed poorly in these treatments may have done so as a means of 'shying away' due to low preferences for competition, rather than performing poorly as a result of the treatment intervention. The elicitation of trait anxiety allows us to disentangle these effects by controlling for participants' aversion to competition. The link between trait anxiety and competitiveness was established by Segal and Weinberg (1984), where they propose trait anxiety as a proxy for competition avoidance.

The questionnaire that we use to elicit trait anxiety is adopted from the State-Trait Anxiety Inventory (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983). There are twenty statements describing various feelings and emotions for which participants are asked to rate on a 4-point scale how accurate each statement reflects their general feelings. Some example statements include "I am inclined to take things hard", "I wish I could be as happy as others seem to be", "I

[24] Aversion to competition may be especially salient in women compared to men. See Niederle and Vesterlund (2007).

lack self-confidence" and "I worry too much over something that doesn't really matter". These statements reflect various dimensions of trait anxiety.

A trait anxiety variable is constructed by adding the elicited scores associated with each question, reversing the statements which are presented with a negative frame. The trait anxiety variable is calculated as:

$$\begin{aligned} Trait\ Anxiety_j \\ &= \left(5 - ta_{1j}\right) + ta_{2j} + ta_{3j} + ta_{4j} + ta_{5j} + (5 - ta_{6j}) \\ &+ (5 - ta_{7j}) + ta_{8j} + ta_{9j} + (5 - ta_{10j}) + ta_{11j} + ta_{12j} \\ &+ (5 - ta_{13j}) + ta_{14j} + ta_{15j} + (5 - ta_{16j}) + ta_{17j} + ta_{18j} \\ &+ (5 - ta_{19j}) + ta_{20j} \end{aligned} \tag{3.2}$$

where ta_{ij} is the self-evaluated score for the i'th statement in the trait anxiety questionnaire for participant j. A higher value indicates a higher level of trait anxiety. The constructed variable is ordinal in nature allowing for comparisons across individuals, even though no intrinsic meaning is attached to the value itself. A priori, it is expected that a participant who is more trait anxious will perform worse than another who has lower trait anxiety, ceteris paribus. This effect should be larger in the treatments that feature competition: the PRWL, T and TNI treatments.

A second questionnaire was handed out at the conclusion of the twenty round forecasting task. The post-task questionnaire (see Appendix A2.7) elicits information on intrinsic motivation, relatedness and participant demographics. The questionnaire begins with 23 statements for which participants rate on a 7 point scale. These statements are associated with how interested participants were in the forecasting task, how competent they felt they were, how much effort they felt they put in, and how anxious they were during the task. Some questions include: "This activity was fun to do", "I was pretty skilled at this activity", "I tried very hard on this activity" and "I felt pressured while doing this activity". This questionnaire is similar to that used by Ryan (1982). In a similar way to how trait anxiety was calculated before, variables reflecting interest, competency, effort and tension are calculated as follows:

$$Interest_j = im_{1j} + im_{5j} + (8 - im_{9j}) + (8 - im_{13j}) + im_{17j} + im_{21j} + im_{23j} \tag{3.3}$$

$$Competency_j = im_{2j} + im_{6j} + im_{10j} + im_{14j} + im_{18j} + (8 - im_{22j}) \tag{3.4}$$

$$Effort_j = im_{3j} + (8 - im_{7j}) + im_{11j} + im_{15j} + (8 - im_{19j}) \tag{3.5}$$

$$Tension_j = (8 - im_{4j}) + im_{8j} + (8 - im_{12j}) + im_{16j} + im_{20j} \tag{3.6}$$

where im_{ij} is the self-evaluated score for statement i of the intrinsic motivation section of the post-questionnaire for participant j. Of these, the competency and tension variables relate to the competency and control components of Cognitive Evaluation Theory. Someone who feels more competent would be more intrinsically motivated to perform while someone with higher tension levels should be less intrinsically motivated to perform. The other variables, interest and effort also relate to intrinsic motivation.

The second part of the post-task questionnaire elicits a measure of how socially distant participants felt they were to their peers. This sense of relatedness is considered a psychological need and should foster engagement through higher levels of effort and interest, and reduced anxiety (Baumeister & Leary, 1995). Relatedness is considered a central element for productive learning (see Ryan & Powelson, 1991; Furrer & Skinner, 2003). This variable is again constructed from ratings, on a 7 point scale, that participants assign to 8 statements. It is constructed as:

$$\begin{aligned} Relatedness_j = \left(8 - r_{1j}\right) + (8 - r_{2j}) + r_{3j} + r_{4j} + (8 - r_{5j}) + (8 - r_{6j}) + r_{7j} \\ + r_{8j} \end{aligned} \tag{3.7}$$

The elicitation of the psychological variables potentially allow us to identify specific factors that drive intrinsic motivation. However, with the exception of trait anxiety, these will not be used to relate directly to forecasting performance. Since the intrinsic motivation and relatedness variables are elicited at the completion of the task, the way participants respond to the statements could have been influenced by how they performed in the task. Yet the insights from psychology suggests that these factors actually influence performance. There is likely to be endogeneity between these factors and forecast errors that cannot be disentangled. Nevertheless there is value in analysing these, so we study these independently to forecast errors.

The remaining part of the post-task questionnaire elicits demographic information such as gender, age, course of study, ethnicity and country of birth.

3.5. Experimental Procedure

The experiment was conducted at the University of Auckland with undergraduate students recruited mainly from the Faculty of Business and Economics. An email was sent out to students in various courses about an economic decision-making experiment and that it was seeking participation (see Appendix 1). Students were told that there will be financial remuneration for participation where they can expect to earn \$20, including a \$5 show up fee, in a session which lasts approximately 90 minutes. The actual earnings depend on the decisions they make during the session. If they wished to participate, they were asked to sign up for their session of choice via a website link. Students are informed that they cannot sign up for and participate in more than one session.

At the start of the experiment, after all participants have been seated, the experimenters first distributed the trait anxiety questionnaire for participants to complete (see Section 3.4 for details). After participants have filled out this trait anxiety questionnaire, the experimenters distributed a copy of the instructions explaining the forecasting task and it is read out loud. These instructions[25] explained to participants that they will be presented with cues and are required to make a prediction of the actual stock value, determined by an underlying relationship involving the cues. Their forecast accuracy is represented by their absolute forecast error, the absolute difference between their prediction and the underlying stock value. Participants were told that they will be paid a piece rate based on their forecast errors for the first five rounds of the task[26], and that there might be a change in the way the game is played at the end of round 5, in which

[25] There are two versions of the general instructions: one version that is based around paying participants piece rates for the first 5 rounds, and another version that announces that they will be receiving a flat salary. The latter version has additional emphasis that participants will only receive \$20 for participation, despite observing information that suggests otherwise. The overall description of the task is identical in both versions. See Appendix A2.1 for these instructions and differences between treatments is noted in brackets.

[26] In all treatments other than the S treatment, participants receive piece rates on their forecasts for the first 5 rounds. Piece rates do not apply in the S treatment as the flat pay had been announced at the beginning of the session.

case further instructions will be provided at that point. They were provided an opportunity to ask questions relating to the task after these instructions have been read.

The experimenters then handed out instructions for participants about how to login to the server which hosted the experimental software. Once in, participants were presented with 10 examples of cues and corresponding stock prices to study before the rounds begin. After providing them 5 minutes to review these examples, participants start the first round of play. They have 90 seconds to make their forecasts. After all participants have submitted their forecasts for the round, the computer software calculates each participant's absolute forecast error and their corresponding piece rate earnings. Participants are provided feedback about these along with the cue values for the round; this information remains on the computer screen for the rest of the experiment. Tables 3.2 to 3.6, presented earlier, showed examples of the on-screen feedback participants observe after each round. Subsequent rounds proceed in the same manner. After round 4, participants are reminded that there may be a change in the way the game is played after round 5.

After round 5 participants are told by the experimenters how the game would proceed from round 6. For the PR treatment, a verbal announcement is made that gameplay would continue as usual with piece rates being applied to forecast errors. For the PRWL, T and TNI treatments, a new set of written instructions were handed out and read aloud by the experimenters (see Appendices A2.2 to A2.4). These instructions explain that participants would now be paired with another participant in the session and would be rematched each round. For the PRWL and T treatments, each participant would be shown winning and losing feedback after each round, while the same information for each round will only be shown at the conclusion of the final round in the TNI treatment. Rank-dependent payoffs are applied to T and TNI treatments after round; participants in these treatments also receive a further $4 lump sum payment. The intervention after round 5 does not apply to the S treatment, since participants are told from the outset that they will be paid a fixed sum at the end of the experiment and that there is no change in the information that is provided to them. After reading the appropriate instructions, participants are given the opportunity to ask questions relating to the modified gameplay before continuing to round 6.

With no changes to the forecasting task, rounds 6 through 20 proceed in a manner similar to the first five rounds except with changes, if any, to the method of payment or to the feedback that is displayed. At the end of the twenty rounds, participants are asked to fill out another questionnaire which elicits various measures of intrinsic motivation and basic demographic information.

After filling out the questionnaire, each participant was called up and paid the balance of their earnings account. Participants were asked to leave the premise after collecting their earnings. Table 3.8 shows the average earnings of participants by treatment.

Table 3.8 Average Earnings by Treatment

	Single Cue Task		Dual Cue Task
PR	$22.81		$19.70
PRWL	$22.96		$20.11
T	$20.85		$19.90
TNI	$20.75 *		$19.98
S	$20		$20

* The average earnings in the single cue TNI treatment are computed with 20 participants since only one session was run.

3.6. Hypotheses

Having laid out the experiment, we now formulate some working hypotheses that we structure our analyses upon. These hypotheses are based on our three research questions. Our first research question asks which of the three pay schemes that we study – piece rates, tournaments and fixed salaries – brings about the best performance from workers.

Comparing the piece rate and tournament pay schemes, we would expect the two pay schemes to perform similarly. This comes from the property of Piece Rate Equivalence, where rank-order tournaments are theoretically shown to elicit the same amount of effort as piece rates do (Lazear & Rosen, 1981). This theoretical property was later confirmed in a lab experiment by Bull et al. (1987). According to Piece Rate Equivalence, we would expect the PR treatment to perform no differently to the T treatment.

H1. *According to Piece Rate Equivalence, we expect forecast errors in the PR treatment to be no different to those in the T treatment.*

Comparing piece rates and salaries, there is some ambiguity about how these pay schemes should perform relative to one another. On the one hand, economic intuition would suggest fixed salaries to perform poorly as players shirk in order to minimise their effort costs. Piece rates would perform well as players exert a high level of effort in order to maximise their monetary payoffs. Conventional intuition would therefore suggest piece rates to outperform salaries. This effect is also known as that brought about by extrinsic incentives.

On the other hand, and as noted earlier, there is a large literature which shows that performance based incentives are counter-productive, as they crowd out the intrinsic motivation of people to perform. See Deci et al. (1999), Frey and Jegen (2001), Gneezy et al. (2011), Bowles and Polanía-Reyes (2012) and Festré and Garrouste (2015) for a selection of recent surveys. According to Cognitive Evaluation Theory, Deci and Ryan (1985) suggest that people view performance-based incentives as controlling, going against their preferences for autonomy. As a result, people's intrinsic motivation falls when they are paid for performance, reducing their productivity as a result.

The issue of motivation crowding out is a contentious one, with evidence supporting opposing conclusions. Our experimental setup provides a clean test of extrinsic versus intrinsic motivation by perfectly controlling for the feedback that people receive. In the S treatment, in addition to feedback of their own forecast errors, players get to see hypothetical feedback as if they were paid according to a piece rate. This allows us to examine the effects of extrinsic and intrinsic motivation, ruling out feedback as a confounding factor.

How piece rates perform relative to salaries depends on the relative magnitudes of the aforementioned effects. If standard economic theory holds, we would expect extrinsic incentives to be dominant, with little to no crowding out of motivation. In this instance, we would expect piece rates to unambiguously outperform salaries. Since this is the benchmark case, we can infer motivation crowding out even if we observe no difference in performance between piece rates and salaries. A stronger form of crowding out occurs if salaries outperform piece rates. We therefore express Hypothesis 2 in two parts:

H2a. *If motivation crowding out does not occur, extrinsic incentives suggest that forecast errors in the PR treatment are smaller than in the S treatment.*

H2b. *If motivation crowding out occurs, the motivating effect associated with extrinsic incentives is overpowered by the reduction in intrinsic motivation. We expect forecast errors in the PR treatment to be greater than or equal to those in the S treatment.*

The final pairwise comparison from the three pay schemes is of tournaments and fixed salaries. Since Piece Rate Equivalence posits that performance under tournaments is no different to that under piece rates, the two-part hypothesis presented as H2 would also apply to tournaments. Tournaments motivate performance, while salaries do not. However since tournaments are controlling, intrinsic motivation would be crowded out in tournaments. The magnitude of the crowding out effect, if it occurs at all, determines how tournaments perform relative to salaries. While adopted directly from hypothesis 2, we present Hypothesis 3 below in two parts:

H3a. *If motivation crowding out does not occur, extrinsic incentives suggest that forecast errors in the T treatment are smaller than in the S treatment.*

H3b. *Under motivation crowding out, we would expect forecast errors in the T treatment to be greater than or equal to forecast errors in the S treatment.*

The following research question relates to the decomposition of tournaments into feedback and payoff components. Accordingly, we split analyses into two parts. To look at the effect of providing relative performance feedback, we compare the PR and PRWL treatments. These treatments are identical, except that the latter provides additional information at the end of each round about whether subjects performed better or worse than their partners. Since subjects are paid piece rates in both treatments, the relative feedback does not impact earnings, allowing us to isolate the effect associated with the relative feedback.

If people are motivated solely by monetary earnings, then we would not expect this relative feedback to have any effect on their performance. If, on the other hand, people derive utility from winning (disutility from losing), we would expect them to exert greater effort to improve their chances of winning (reduce their chances of losing). As such we would expect performance to improve. This effect may consist of an ex ante anticipation effect (Blanes i Vidal & Nossol, 2011) that is associated with preferences for status and respect (see Ellingsen & Johannesson, 2007 for a review), or an ex-post revelation effect when players respond to the feedback received. Both effects suggest better performance when relative performance feedback is provided.

H4. *The provision of relative performance feedback improves performance. We expect forecast errors to be lower in the PRWL treatment than in the PR treatment.*

The other part of the second research question relates to the rank-dependent payoffs inherent in tournaments. We compare the effectiveness of rank-dependent payoffs in motivating performance with piece rates. This is addressed by comparing the forecast errors of the PRWL and T treatments. Both treatments feature rank feedback, but differ in terms of incentives: piece rates and rank-dependent prizes respectively.

Piece Rate Equivalence suggests that tournaments perform similarly to piece rates. However this property is not directly applicable to our research question, which involves feedback-augmented piece rates (as in the PRWL treatment). If standard piece rates perform no differently to tournaments (H1), and feedback on relative performance induces higher performance from players (H4), then we would accordingly expect the feedback-augmented piece rates to perform better than tournaments. Both Hannan et al. (2008) and Eriksson et al. (2009) provide evidence indicative of this, although not statistically significant.

H5. *Piece rates perform better than rank-dependent payoffs of tournaments, when relative performance feedback is controlled for. We would expect forecast errors in the PRWL treatment to be lower than in the T treatment.*

Our final research question relates to learning. Which of our treatments brings about the fastest rate of learning? We posit that the rate of learning is highest amongst treatments that feature an element of competition. This is because competition would continuously motivate people to perform. According to Dutcher, Balafoutas, Lindner, Ryvkin, and Sutter (2015), tournament players increase their effort immediately following a loss, since they want to reduce their chances of losing in subsequent rounds. The competition provides the impetus to perform.

Since competition features in both the PRWL and T treatments, we would expect learning to be more pronounced in these treatments compared to the PR and S treatments.

Between the PRWL and T treatments, we would expect learning to be more salient in the T treatment than in the PRWL treatment. While competition is common to both treatments, the precise nature of it is different. In the PRWL treatment, winning or losing itself has no effect on monetary payoffs, since players are paid according to their absolute performance. On the other hand, winning and losing has financial implications in the T treatment. Winners receive a positive payoff, while losers receive nothing. Since the notion of competition is reinforced by payoffs in the T treatment, we would expect learning to occur at a faster rate in the T treatment than in the PRWL treatment.

We summarise our learning hypothesis below:

H6. *The rate of learning should be most pronounced in the T treatment, followed by the PRWL treatment. Learning occurs at a slower rate in the PR and S treatments.*

These six hypotheses presented here form the core of our study, addressing our three primary research questions. The presentation of results in the upcoming chapter will be structured around these hypotheses.

4. Results: Treatment Effects

4.1. First Five Rounds

From the design of our experiment, participants' earnings in the first five rounds of play are based on piece rates in all treatments, except the S treatment. Before we proceed with between-subjects comparisons of forecast errors in rounds 6 to 20, we begin by confirming that there are no differences in performance across treatments in these pre-intervention rounds, since the treatments should be identical to each other ex ante. This check is necessary to ensure that the forthcoming between-subjects analyses is not undermined by systematic differences in participants' performance.

One way to analyse pre-intervention differences in performance is to look at each participant's median forecast error across the first five rounds of play. This serves as a metric for the overall ability of each participant. A close look at the data for individual participants show that those who make unusually large forecast errors in any of the first five rounds do not necessarily perform poorly overall – it may be due to players making the odd mistake, or my simply reflect a trial-and-error strategy being played. In this regard the median statistic is an appropriate measure of ability, since it is unaffected by unusual forecast error values that do not necessarily reflect an individual's overall ability. As with forecast errors, a smaller value represents higher ability. The average of these are shown in Table 4.1.

Two obvious features of the data are apparent from Table 4.1. Firstly, forecast errors tend to be larger in the dual cue task than in the single cue task in every treatment; this shows that the dual cue task is more difficult than the single cue task. Secondly, there is a large spread in forecast errors even within each treatment. These two features can also be observed in the distribution plots in Figure 4.1.

From Table 4.1 we can see that the average ability of participants are reasonably similar across each of the single cue treatments, with ability values ranging from 9.1 to 11.9 across treatments. A Kruskal Wallis test shows that there are no cross treatment differences in ability across participants in the single cue task ($\chi^2(3) = 1.60$, $p = 0.659$, $n = 166$). Across the dual cue treatments, a Kruskal Wallis test shows that there are no significant differences in the distribution of participant ability ($\chi^2(3) = 2.04$, $p = 0.564$, $n = 146$).

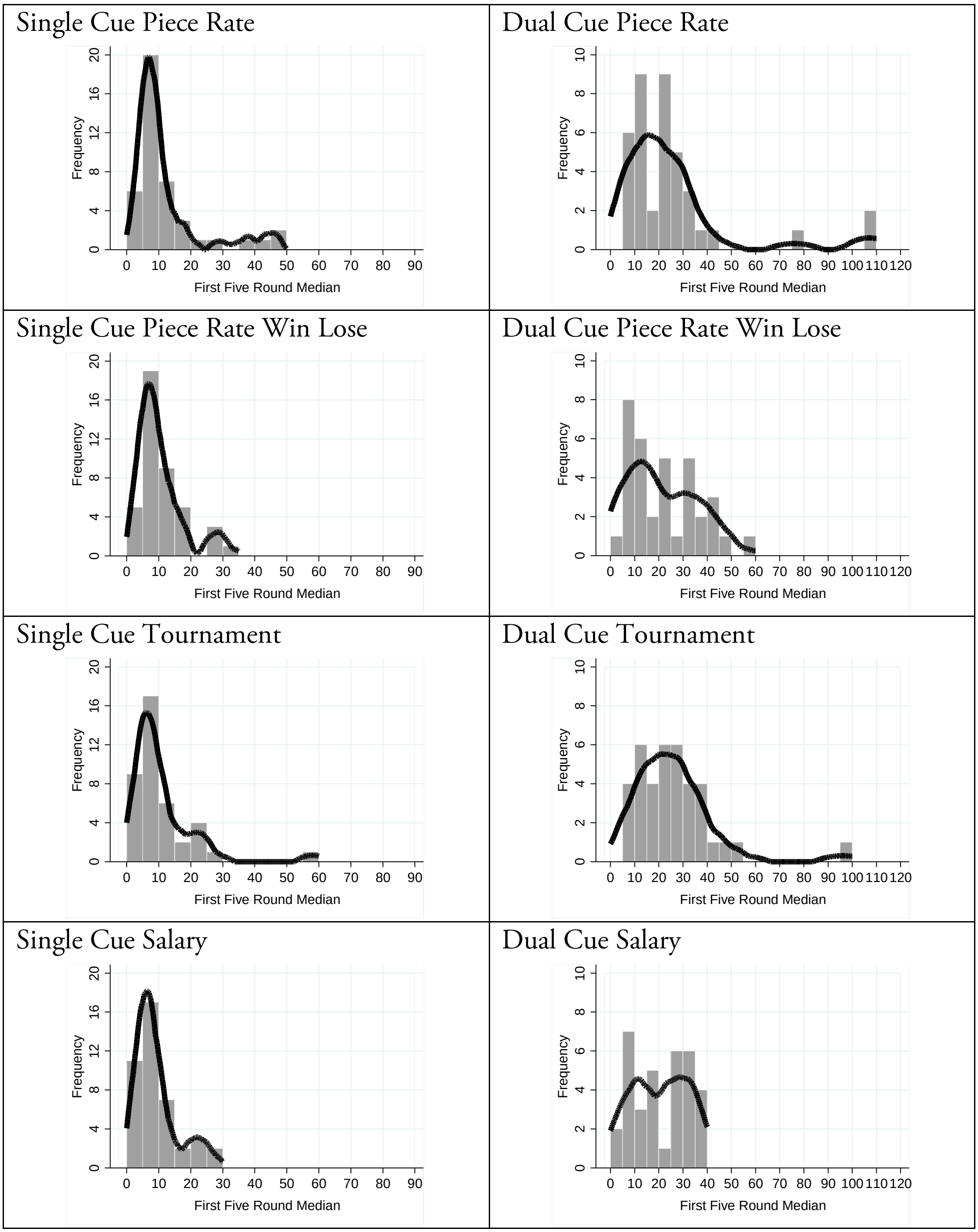
Table 4.1 Participant Ability by Treatment

	Single Cue	Dual Cue
Piece Rate	11.93 (11.04)	25.23 (23.36)
Piece Rate Win Lose	10.38 (7.10)	21.94 (13.93)
Tournament	10.90 (9.99)	25.89 (16.27)
Salary	9.14 (6.63)	20.79 (11.25)
Aggregated	10.58 (8.85)	23.58 (17.02)

Ability is defined to be the median forecast error of each participant across the first five rounds. Smaller values indicate higher ability. Table shows mean ability; standard deviation of ability are in parentheses.

We have now established that there are no significant differences in participants' ability in the first five rounds of play. This finding enables us to analyse the effects of experimental interventions between subjects, and allows us to make causal inferences without worrying about how exogenous differences in participants' ability and characteristics undermine the results.

Figure 4.1 Distributions of Participant Ability in Rounds 1 to 5, by Treatment



Single Cue Piece Rate

Dual Cue Piece Rate

Single Cue Piece Rate Win Lose

Dual Cue Piece Rate Win Lose

Single Cue Tournament

Dual Cue Tournament

Single Cue Salary

Dual Cue Salary

Ability is defined to be the median forecast error of each participant across the first five rounds. Since they are measured in forecast error units, ability is also truncated at zero. The histograms show the number of participants in each treatment whose ability fall within each category of width 10. The solid line shows the kernel density.

4.2. Descriptive Statistics

In this section we compare participants' forecast errors in rounds 6 to 20 across treatments. To start off, descriptive statistics of forecast errors in rounds 6 to 20 are presented in Table 4.2, while Figure 4.2 depicts the distribution of forecast errors across treatments. Prima facie, the shape of the forecast error distributions are similar across the single cue treatments, with the peak showing 80-90% of the forecast error observations taking a value of 10 or less. The right tail of the PRWL distribution represents a single observation where the forecast error took the value of 299.

The mean forecast errors in Table 4.2 show that there are no obvious differences in performance across the single cue treatments. The median forecast errors are practically identical. A series of non-parametric Wilcoxon Ranksum tests of pairwise treatment differences in forecast errors are tabulated in Table 4.3. The unit of observation is the median forecast error of each participant across the 15 post-intervention rounds, yielding a single observation for each participant.[27] These pairwise tests confirm that there are no significant post-intervention differences in performance across single cue treatments.

In the dual cue task, the T treatment stands out from the others. The mean forecast error of 30.7 in the T treatment appear to be much larger than forecast errors of the other treatments, with mean forecast errors in the vicinity of 25. The distribution plots in Figure 4.2 show that despite the T treatment having a longer right tail than the PRWL and S treatments, much of the difference lies in that the peak of the distribution does not rise nearly as high as in the other treatments. The higher median forecast error in the T treatment also reflects this. However, when we look at the rank-sum tests presented in Panel B of Table 4.3, we do not observe the T treatment to perform significantly worse than other treatments at conventional levels. The rank-sum test between the dual cue PRWL and T treatments, however, only misses out at the 10% significance level.

[27] Due to the panel nature of the data, we cannot conduct rank-sum tests with raw observations, since the raw observations are not truly independent – since they are correlated for each participant over the dimension of time. We run the rank-sum tests with the median forecast error for each participant across rounds 6 to 20 to circumvent the independence requirement.

Table 4.2 Average Forecast Errors in Rounds 6 to 20, by Treatment

	Single Cue		Dual Cue	
	Mean (Std Dev)	Median	Mean (Std Dev)	Median
PR	10.16 (14.70)	5	26.56 (33.63)	15
PRWL	9.60 (16.83)	5	24.03 (25.91)	15
T	10.01 (14.40)	5	30.74 (36.55)	19
S	9.01 (11.84)	6	25.06 (27.00)	16

Forecast errors in rounds 6 to 20 averaged across participants and time for each treatment.

Table 4.3 Pairwise Ranksum Differences in Forecast Errors in Rounds 6-20

Panel A: Single Cue Task

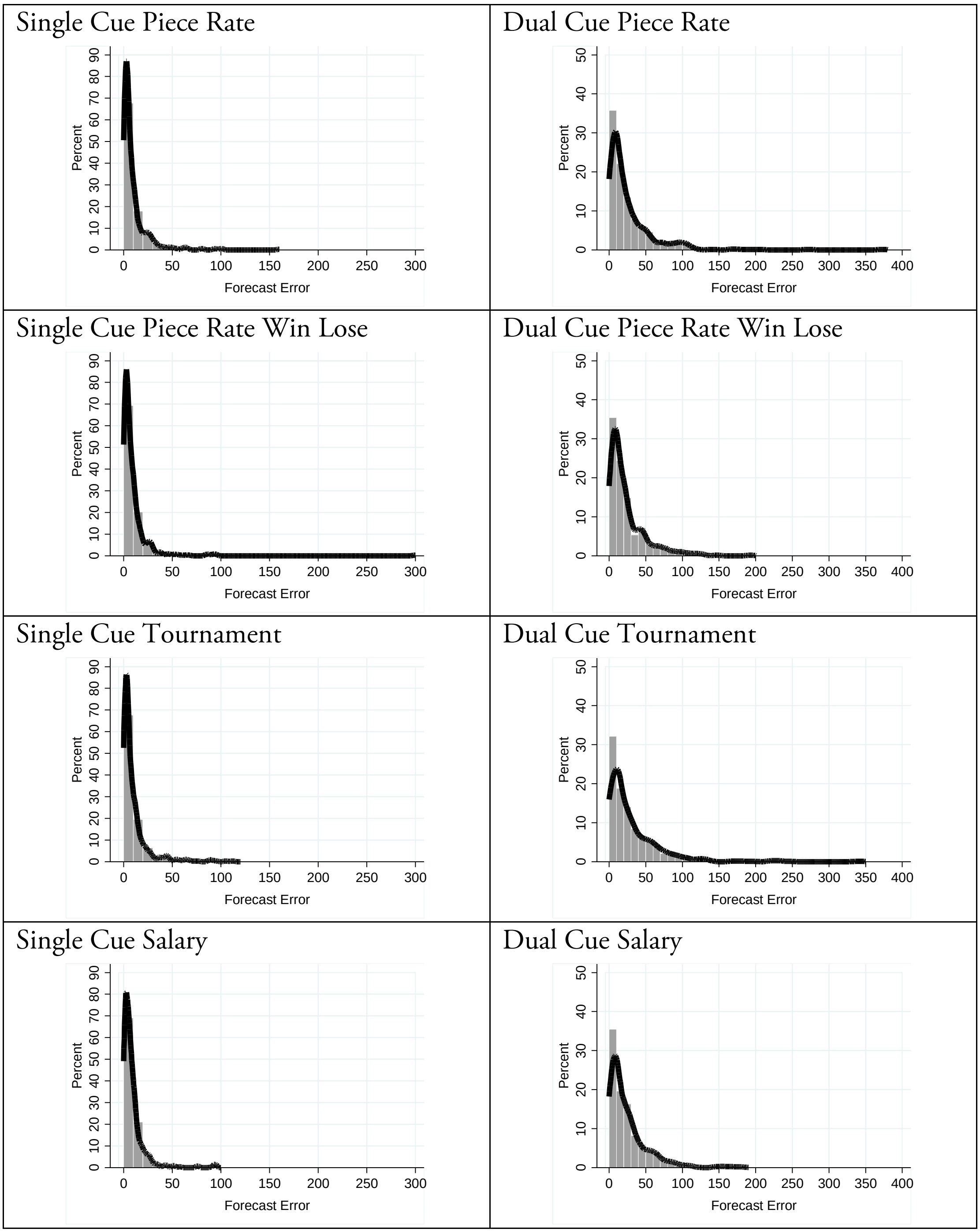
	PRWL (n = 42)	T (n = 40)	S (n = 42)
PR (n = 42)	|z| = 0.10 p = 0.917	|z| = 0.78 p = 0.438	|z| = 0.93 p = 0.352
PRWL (n = 42)		|z| = 0.53 p = 0.600	|z| = 0.83 p = 0.404
T (n = 40)			|z| = 1.57 p = 0.116

Panel B: Dual Cue Task

	PRWL (n = 35)	T (n = 38)	S (n = 34)
PR (n = 39)	|z| = 0.72 p = 0.474	|z| = 0.63 p = 0.531	|z| = 0.08 p = 0.934
PRWL (n = 35)		|z| = 1.63 p = 0.104	|z| = 0.78 p = 0.435
T (n = 38)			|z| = 0.89 p = 0.372

Unit of observation is the median forecast error over the rounds 6 to 20 for each participant..

Figure 4.2 Distribution of Forecast Errors in Rounds 6-20, by Treatment



Single Cue Piece Rate

Dual Cue Piece Rate

Single Cue Piece Rate Win Lose

Dual Cue Piece Rate Win Lose

Single Cue Tournament

Dual Cue Tournament

Single Cue Salary

Dual Cue Salary

Forecast errors are expressed as absolute values, so the distributions are truncated at zero. The histogram shows the percentage of forecast error observations that fall within each category of width 10. The solid line is the kernel density.

Initial analyses shows that there are no significant differences in forecast errors between treatments for both the single and dual cue tasks. One problem inherent in the data is that there is a great deal of variability in forecast errors both across participants and across time. It is quite possible that the variability, unaccounted for, may be masking some differences across treatments.

We now proceed with regression analyses, which allows us to explain some of this noise by controlling for other factors that might influence performance. Panel data models are also able to account for some of the unobserved time-invariant differences across participants. Regression analyses also allow us to answer more complex questions like how the process of learning is affected by the treatment interventions.

4.3. Baseline Results

In this section, we present our baseline regressions, which we will use to base our results upon. These regressions are run across rounds 6 to 20, the post-intervention rounds, and are estimated with random effects generalised least squares. We opt for random effects models over fixed effects models here because fixed effects models are unable to estimate variables which are time invariant. This means that the treatment dummies, which we use to estimate the effects arising from experimental interventions, are inestimable since participants who are assigned to a treatment remain in that treatment for the entire duration of the task. Due to this limitation of the fixed effects model, random effects estimation is used by default without consideration to the underpinning assumptions relating to the individual fixed effects. The reported standard errors are clustered by participants to control for the within-participant correlation of standard errors across rounds.

Our baseline regressions in Tables 4.4 to 4.6 regress forecast errors in rounds 6 to 20 while controlling for a series of different variables. Each of these tables show the same regression specifications, but with different datasets: Table 4.4 runs the regressions with the single and dual cue data pooled together, while Tables 4.5 and 4.6 are repeated with single and dual cue treatments individually. The main variables of interest are a complete series of dummy variables identifying each treatment, with the PR treatment serving as the reference category. The estimated coefficients on these treatment dummies will reveal how the different treatments perform vis à vis others, quantifying the effect of the various treatment interventions. Panel B of each regression table presents the results from Wald chi-squared tests of pairwise treatment differences.

We also control for learning over time with the 'Round' variable, which represents a linear time trend. The linear time trend itself reveals the general pattern of how forecast errors change over the 15 post-intervention rounds. If the trend is negative, it means that there is evidence for learning over time, where forecast errors improve over time. In the regression tables, models 1 and 2 estimate a common trend for all participants. In models 3 and 4, we allow for treatment-specific trends by interacting the linear time trend by treatment. The interactions include the PR treatment, so there is no reference category to interpret the trends against.[28] In Chapter 5, we study the notion of learning in greater detail.

Models 2 and 4 of the regression tables also incorporate a number of control variables. These controls include the trait anxiety and gender of participants. Trait anxiety scores measure how prone participants are to stress and situations that may make them anxious. We use this as a proxy to control for participants' competitiveness – see Segal and Weinberg (1984). Although we expect people who are more trait anxious to perform worse, trait anxiety does not seem to significantly influence forecast errors in our baseline regressions.

Gender is also included in the regression models. The forecasting task that we use is cognitively challenging and we expect women to perform worse than men, since men have been shown to be more capable in terms of abstract problem solving (Hyde, Fannema, & Lamon, 1990). The gender dummy is therefore included to capture these gender differences. Gender differences in performance are only present in the dual cue task, and not in the simpler single cue task. Chapter 7 studies gender performance differences in greater detail.

We proceed with discussing the findings of our baseline regressions in Tables 4.4 to 4.6, systematically addressing our main first two research questions and associated hypotheses. After the main results have been highlighted, a series of robustness checks are presented. Evidence suggests that the PRWL and S treatments perform significantly better than the PR and T treatments, with no differences between the PRWL and S treatments, and between the PR and T treatments. We discuss the results in detail below.

[28] As such, the coefficients on each of these treatment-round interactions are interpreted as the slopes of the estimated trend line, rather than the marginal change in slope relative to that of the reference treatment.

Table 4.4 Baseline Regressions of Forecast Errors: Pooled

Panel A: Regression Results

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-1.897 (2.185)	-4.235 * (2.399)	-3.275 (2.939)	-6.114 * (3.185)
Tournament	2.052 (2.738)	0.661 (2.944)	4.984 (3.177)	3.339 (3.381)
Salary	-1.865 (2.182)	-4.185 * (2.329)	-2.558 (2.684)	-5.749 ** (2.848)
Trait Anxiety		0.080 (0.121)		0.080 (0.121)
Female		8.173 *** (1.536)		8.173 *** (1.536)
Round	-0.113 * (0.063)	-0.126 * (0.068)		
PR*Round			-0.096 (0.129)	-0.142 (0.146)
PRWL*Round			0.010 (0.147)	0.003 (0.153)
T*Round			-0.322 *** (0.109)	-0.348 *** (0.119)
S*Round			-0.043 (0.116)	-0.021 (0.118)
Constant	19.53 *** (1.849)	13.83 *** (4.925)	19.31 *** (2.017)	14.03 *** (5.413)
Observations	4680	4245	4680	4245
Participants	312	283	312	283
R^2	0.004	0.036	0.005	0.036
Wald χ^2	6.87	40.91	13.60	50.50
$p > \chi^2$	0.143	0.000	0.059	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.75$ p = 0.385	**$\chi^2(1) = 3.12$** **p = 0.078**	$\chi^2(1) = 1.24$ p = 0.265	**$\chi^2(1) = 3.68$** **p = 0.055**
PRWL = T	**$\chi^2(1) = 2.71$** **p = 0.099**	**$\chi^2(1) = 4.07$** **p = 0.044**	**$\chi^2(1) = 6.44$** **p = 0.011**	**$\chi^2(1) = 7.71$** **p = 0.006**
T = 0	$\chi^2(1) = 0.56$ p = 0.454	$\chi^2(1) = 0.05$ p = 0.822	$\chi^2(1) = 2.46$ p = 0.117	$\chi^2(1) = 0.98$ p = 0.323
S = 0	$\chi^2(1) = 0.73$ p = 0.393	**$\chi^2(1) = 3.23$** **p = 0.072**	$\chi^2(1) = 0.91$ p = 0.341	**$\chi^2(1) = 4.08$** **p = 0.044**
T = S	$\chi^2(1) = 2.67$ p = 0.102	**$\chi^2(1) = 4.18$** **p = 0.041**	**$\chi^2(1) = 6.20$** **p = 0.013**	**$\chi^2(1) = 8.82$** **p = 0.003**
PR*Round = PRWL*Round			$\chi^2(1) = 0.30$ p = 0.587	$\chi^2(1) = 0.47$ p = 0.494
PRWL*Round = T*Round			**$\chi^2(1) = 3.28$** **p = 0.070**	**$\chi^2(1) = 3.27$** **p = 0.071**
PR*Round = T*Round			$\chi^2(1) = 1.79$ p = 0.182	$\chi^2(1) = 1.20$ p = 0.274
PR*Round = S*Round			$\chi^2(1) = 0.09$ p = 0.758	$\chi^2(1) = 0.41$ p = 0.521
T*Round = S*Round			**$\chi^2(1) = 3.06$** **p = 0.080**	**$\chi^2(1) = 3.77$** **p = 0.052**

Bold typeface indicates statistical significance at the 10% level or better.

Table 4.5 Baseline Regressions of Forecast Errors: Single Cue

Panel A: Regression Results

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win-Lose	-0.557 (1.493)	-1.127 (1.734)	-1.317 (2.973)	-2.692 (3.335)
Tournament	-0.145 (1.656)	-0.313 (1.957)	0.769 (2.766)	0.193 (3.202)
Salary	-1.149 (1.349)	-1.910 (1.571)	-0.665 (2.432)	-2.452 (2.774)
Trait Anxiety		-0.063 (0.098)		-0.063 (0.098)
Female		1.796 (1.167)		1.796 (1.168)
Round	-0.111 * (0.061)	-0.123 * (0.067)		
PR*Round			-0.100 (0.177)	-0.158 (0.140)
PRWL*Round			-0.042 (0.177)	-0.038 (0.181)
T*Round			-0.170 * (0.090)	-0.197 ** (0.100)
S*Round			-0.137 * (0.077)	-0.116 (0.077)
Constant	11.61 *** (1.493)	14.07 *** (4.587)	11.46 *** (2.103)	14.53 *** (5.077)
Observations	2490	2220	2490	2220
Participants	166	148	166	148
R^2	0.002	0.009	0.002	0.009
Wald χ^2	4.77	12.58	8.63	15.07
$p > \chi^2$	0.312	0.050	0.280	0.089

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.14$ p = 0.709	$\chi^2(1) = 0.42$ p = 0.516	$\chi^2(1) = 0.20$ p = 0.658	$\chi^2(1) = 0.65$ p = 0.420
PRWL = T	$\chi^2(1) = 0.08$ p = 0.776	$\chi^2(1) = 0.30$ p = 0.583	$\chi^2(1) = 0.57$ p = 0.451	$\chi^2(1) = 0.99$ p = 0.319
T = 0	$\chi^2(1) = 0.01$ p = 0.930	$\chi^2(1) = 0.03$ p = 0.873	$\chi^2(1) = 0.08$ p = 0.781	$\chi^2(1) = 0.00$ p = 0.952
S = 0	$\chi^2(1) =0.73$ p = 0.394	$\chi^2(1) = 1.48$ p = 0.224	$\chi^2(1) = 0.07$ p = 0.785	$\chi^2(1) = 0.78$ p = 0.377
T = S	$\chi^2(1) = 0.60$ p = 0.440	$\chi^2(1) = 1.44$ p = 0.230	$\chi^2(1) = 0.44$ p = 0.509	$\chi^2(1) = 1.38$ p = 0.240
PR*Round = PRWL*Round			$\chi^2(1) = 0.08$ p = 0.783	$\chi^2(1) = 0.28$ p = 0.599
PRWL*Round = T*Round			$\chi^2(1) = 0.42$ p = 0.516	$\chi^2(1) = 0.59$ p = 0.442
PR*Round = T*Round			$\chi^2(1) = 0.23$ p = 0.634	$\chi^2(1) = 0.05$ p = 0.822
PR*Round = S*Round			$\chi^2(1) = 0.07$ p = 0.791	$\chi^2(1) = 0.07$ p = 0.794
T*Round = S*Round			$\chi^2(1) = 0.08$ p = 0.779	$\chi^2(1) = 0.41$ p = 0.524

Bold typeface indicates statistical significance at the 10% level or better.

Table 4.6 Baseline Regressions of Forecast Errors: Dual Cue

Panel A: Regression Results

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win-Lose	-2.533 (3.434)	-4.523 (3.495)	-4.659 (4.708)	-6.889 (4.935)
Tournament	4.173 (4.416)	2.294 (4.577)	9.229 * (4.747)	7.067 (4.895)
Salary	-1.498 (3.405)	-4.716 (3.669)	-3.653 (4.418)	-7.684 (4.714)
Trait Anxiety		0.090 (0.193)		0.090 (0.193)
Female		10.08 *** (2.375)		10.08 *** (2.376)
Round	-0.115 (0.117)	-0.129 (0.123)		
PR*Round			-0.092 (0.236)	-0.127 (0.246)
PRWL*Round			0.072 (0.243)	0.055 (0.261)
T*Round			-0.481 ** (0.201)	-0.494 ** (0.212)
S*Round			0.074 (0.241)	0.101 (0.251)
Constant	28.06 *** (2.941)	19.84 *** (7.941)	27.76 *** (2.990)	19.81 ** (8.680)
Observations	2190	2025	2190	2025
Participants	146	135	146	135
R^2	0.007	0.034	0.008	0.035
Wald χ^2	5.15	25.13	12.11	31.80
$p > \chi^2$	0.273	0.000	0.097	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.54$ p = 0.461	$\chi^2(1) = 1.67$ p = 0.196	$\chi^2(1) = 0.98$ p = 0.322	$\chi^2(1) = 1.92$ p = 0.162
PRWL = T	**$\chi^2(1) = 3.30$** **p = 0.069**	**$\chi^2(1) = 3.52$** **p = 0.061**	**$\chi^2(1) = 7.19$** **p = 0.007**	**$\chi^2(1) = 6.47$** **p = 0.011**
T = 0	$\chi^2(1) = 0.89$ p = 0.345	$\chi^2(1) = 0.25$ p = 0.616	**$\chi^2(1) = 3.78$** **p = 0.052**	$\chi^2(1) = 2.08$ p = 0.149
S = 0	$\chi^2(1) = 0.19$ p = 0.660	$\chi^2(1) = 1.65$ p = 0.199	$\chi^2(1) = 0.68$ p = 0.408	$\chi^2(1) = 2.66$ p = 0.103
T = S	$\chi^2(1) = 2.40$ p = 0.122	**$\chi^2(1) = 3.52$** **p = 0.061**	**$\chi^2(1) = 6.86$** **p = 0.009**	**$\chi^2(1) = 8.29$** **p = 0.004**
PR*Round = PRWL*Round			$\chi^2(1) = 0.23$ p = 0.630	$\chi^2(1) = 0.26$ p = 0.612
PRWL*Round = T*Round			**$\chi^2(1) = 3.06$** **p = 0.080**	$\chi^2(1) = 2.66$ p = 0.103
PR*Round = T*Round			$\chi^2(1) = 1.58$ p = 0.209	$\chi^2(1) = 1.28$ p = 0.258
PR*Round = S*Round			$\chi^2(1) = 0.24$ p = 0.623	$\chi^2(1) = 0.42$ p = 0.516
T*Round = S*Round			**$\chi^2(1) = 3.13$** **p = 0.077**	**$\chi^2(1) = 3.28$** **p = 0.070**

Bold typeface indicates statistical significance at the 10% level or better.

4.3.1. Pay Schemes

Piece Rates and Tournaments

Our first research question asks which of the three pay schemes that we look at – piece rates, tournaments and fixed salaries – induces the best performance from workers. This research question can be broken down into three pairwise comparisons.

The first pairwise comparison is of piece rates and tournaments. Drawing from the property of Piece Rate Equivalence, hypothesis H1 – laid out earlier in Section 3.6 – posited that forecast errors in the PR and T treatment would not be statistically different from one another. From the regressions in Tables 4.4 to 4.6, this hypothesis would be represented by an insignificant T treatment dummy, since the PR treatment serves as the reference category.

Overall, we find that tournaments do not perform differently to piece rates. In the baseline pooled regressions in Table 4.4, the coefficients on the T treatment dummy are positive, but with large standard errors. As a result, none of the coefficients in these models are statistically significant at conventional levels. When we look only at the single cue treatments, Table 4.5 again shows that the dummy variable for the T treatment is statistically insignificant in each of the four regression models.

In the dual cue regressions in Table 4.6, we see that the T dummy is positive in each of the regression models, although insignificant in models 1, 2 and 4. In regression model 3, forecast errors are higher in the T treatment than in the PR treatment by an average of 9.23 points, and is statistically significant with a p-value of 0.052. While model 3 shows that tournaments perform significantly worse than piece rates in the dual cue task, this finding no longer holds when controls of trait anxiety and gender are included in model 4. It should be noted that while the forecast errors in the T treatment are large in regression models 3 and 4, it is compensated by a greater rate of learning in the T treatment than in the PR treatment. We will elaborate on the notion of learning at a later point in the following chapter.

The insignificant T dummy in our pooled, single and dual cue baseline regressions in Tables 4.4 to 4.6 lends support to hypothesis H1. Tournaments do not perform differently to piece rates. This verifies the property of Piece Rate Equivalence. Our first result is stated below:

Result 4.1. *Piece Rate Equivalence is verified. Tournaments do not perform any differently to piece rates.*

Piece Rates and Salaries

Our second pairwise comparison of pay schemes is between piece rates and fixed salaries. As discussed earlier, the effects of extrinsic incentives and motivation crowding out play out in different directions. The relative magnitude of these effects determines how piece rates perform relative to salaries. Hypothesis H2 came in two parts. According to H2a, if the motivation crowding out effect plays little or no role and the extrinsic effect dominates, then we would expect the PR treatment to have smaller forecast errors than the S treatment. On the other hand if H2b holds, then forecast errors in the PR treatment are greater than or equal to forecast errors in the S treatment, indicating motivation crowding out.

We find evidence to support the existence of motivation crowding out. The pooled regressions in Table 4.4 show the S treatment dummy to have a negative coefficient in each of the four specifications. The S dummy is statistically significant in regression models 2 and 4 which incorporate the controls of trait anxiety and gender. In the single cue regressions in Table 4.5, we observe the coefficients on the S treatment dummy in each regression model to be negative, although none are statistically significant. It is similar for the dual cue task in Table 4.6, where again the S treatment dummy is negative but insignificant in each specification. The S dummy in regression model 4 of Table 4.6, however, is only insignificant at the margin, with a p-value of 0.103.

It is interesting that the signs on the S dummy coefficients in each of the pooled, single and dual cue regression models are negative, but is only statistically significant in the pooled regressions of Table 4.4. This is likely due to the reduction in statistical power in the single and dual cue regressions as the sample size reduces through disaggregation. Nevertheless, the results point towards salaries performing better than piece rates, providing support for hypothesis H2b over the alternative H2a, that piece rates crowd out intrinsic motivation.[29] Accordingly, Result 4.2 is stated below:

Result 4.2. *Piece rates perform worse than fixed salaries. The reduction in intrinsic motivation is more salient than the extrinsic incentives of piece rates.*

Tournaments and Salaries

Our final pairwise comparison of pay schemes is of tournaments and salaries. Given our first two results that a) piece rates perform similarly to tournaments, and that b) piece rates perform worse than fixed salaries, we would accordingly expect tournaments to perform worse than salaries. This is expressed as hypothesis H3b, which we would expect to hold over H3a.

The various regression tables show that the S treatment performs significantly better than the T treatment. From the pooled regressions in Table 4.4, we see that forecast errors in the S treatment are estimated to be smaller than those for the T treatment – Wald tests show the difference to be highly significant in regression models 2 to 4, and only marginally insignificant in model 1 with p = 0.102. The pattern of results is replicated when we run identical regressions for the dual cue task, where we see again that the S treatment performs better than the T treatment in models 2 to 4 of Table 4.6. For the single cue task in Table 4.5, although the differences are insignificant, the estimated coefficients for T and S treatments indicate that the S treatment performs better than the T treatment.

Our baseline regressions lend support for H3b, that salaries perform better than tournaments. Drawing upon Cognitive Evaluation Theory, this is attributable to the higher degree of control inherent in tournaments, which people may be averse to. In order for the participant to earn any money under tournaments, they are required to outperform their random partner. On the other hand, there are no performance requirements for the participant to fulfil in order to earn money under salaries. The greater degree of control in tournaments reduces participants' autonomy

[29] Motivation crowding out can also be inferred even if piece rates perform similarly to salaries, as we find in the single and dual cue baseline regressions. This is because if extrinsic incentives are dominant, then piece rates should unambiguously *outperform* salaries.

relative to salaries, and as a result their intrinsic motivation falls, leading to lower forecast performance.

This explanation can be supported more directly by looking at the various psychological measures of intrinsic motivation which we have elicited from the post-task questionnaire (see Section 3.4 for details). Table 4.7 presents the mean and standard deviation of four self-reported psychological variables – interest, competency, effort and tension – for the S and T treatments when the single and dual cue observations are pooled together. To validate our explanation about the greater degree of control in the T treatment, we focus on the variable of 'tension'. Tension is indicative of control and participants' loss of autonomy. For example, the strict requirement of winning to earn money in tournaments brings about pressure for participants to perform, for which they should report higher levels of tension relative to salaries.

From Table 4.7, we see that on average S participants report tension levels of 13.10 points, whereas T participants report higher levels of tension at 16.59 points. According to a rank-sum test, the difference is highly significant (|z| = 3.58, p = 0.000, n = 149). This supports the explanation that tournaments perform worse than salaries due to its more controlling nature. Of the other psychological variables, there is no difference in the reported levels of interest and effort between the S and T treatments, although we do find that S participants rate themselves to be more competent than T participants.

We can narrow down our comparison of tournaments and salaries. Cognitive Evaluation Theory posits that intrinsic motivation improves as people receive favourable feedback regarding their level of competency, while motivation falls as they receive unfavourable feedback. This applies to tournaments, where feedback provides rich information to players, allowing them to gauge their competency against others. Since the winning or losing feedback is binary, there is no ambiguity in interpreting whether the feedback received is favourable or not.[30]

[30] Compare this to the absolute performance feedback of forecast errors in the PR and S treatments. Although smaller forecast errors indicate better performance than larger forecast errors, it is difficult for players to assess whether a particular forecast error represents 'good' or 'bad' performance since there is no benchmark.

Table 4.7 Psychological Measures of Intrinsic Motivation: T and S Treatments

Panel A: Means and Standard Deviations

	Salary	Tournament	Tournament Winners	Tournament Losers
Interest	34.72 (8.54)	34.23 (7.82)	37.20 (6.89)	30.95 (7.56)
Competency	29.04 (5.23)	25.19 (8.02)	29.30 (5.59)	20.49 (7.83)
Effort	23.05 (6.26)	22.49 (4.52)	22.80 (4.54)	22.14 (4.55)
Tension	13.10 (5.40)	16.59 (6.14)	16.20 (5.96)	17.06 (6.40)

Tournament winners are defined to be those in the T treatment who wins 8 or more of the 15 tournament rounds, while Tournament losers are those who wins 7 rounds or fewer. Means of the psychological variables are presented; standard deviations are in parentheses.

Panel B: Wilcoxon Rank-Sum Tests

	Salary = Tournament	Salary = Tournament Winners	Salary = Tournament Losers	Winners = Losers

Interest	|z| = 0.14
	p = 0.891
	n = 152
		|z| = 1.59
		p = 0.111
		n = 115
			|z| = 1.94
			p = 0.053
			n = 111
				|z| = 3.51
				p = 0.001
				n = 78

Competency	**|z| = 2.94**
	p = 0.003
	n = 150
		|z| = 0.17
		p = 0.865
		n = 115
			|z| = 5.21
			p = 0.000
			n = 110
				|z| = 4.60
				p = 0.000
				n = 75

Effort	|z| = 0.42
	p = 0.675
	n = 150
		|z| = 0.12
		p = 0.905
		n = 114
			|z| = 0.58
			p = 0.559
			n = 110
				|z| = 0.53
				p = 0.598
				n = 76

Tension	**|z| = 3.58**
	p = 0.000
	n = 149
		|z| = 2.78
		p = 0.006
		n = 114
			|z| = 3.10
			p = 0.002
			n = 108
				|z| = 0.53
				p = 0.594
				n = 76

These elicited variables are participant-specific, so statistics and tests use a sample equal to the number of participants in these treatments (sub-groups). Bold typeface indicates statistical significance at 10% or better.

In the last two columns of Panel A of Table 4.7, we distinguish T participants according to whether they have won or lost more than half the time. For simplicity, participants who wins more often than loses will be referred to as "winners", while those who loses more often than wins will be referred to as "losers". Of the 78 participants in the Tournament treatment across both single and dual cue tasks, there are 41 winners and 37 losers according to this classification. Naturally, the median performance of each player in the first five rounds differ significantly between T winners and losers (rank-sum |z| = 2.36, p = 0.019, n = 78).

There are some clear differences in these measures of intrinsic motivation between winners and losers in the T treatment. At the end of the game, T losers' self-reported scores suggest that they are significantly less interested and less competent than T winners. The lower competency of losers relative to winners is consistent with Cognitive Evaluation Theory. There are no differences in the self-reported effort and tension scores between winners and losers in the T treatment.

We now compare these measures of intrinsic motivation for T winners and losers with participants in the S treatment. While the aggregated level of interest in the T treatment is similar to that in the S treatment, losers in the T treatment show significantly less interest than S participants. T winners appear to be more interested than S participants, though not significantly so with a p-value just short of the 10% threshold (p = 0.111).

Competency levels are similar between T winners and S participants. On the other hand, T losers feel less competent than S participants.

Tension levels are significantly higher for both T winners and losers compared to S participants. Reported tension levels are similar across T winners and losers. This is highly suggestive of the Cognitive Evaluation Theory prediction that people feel that tournaments are more controlling than fixed salaries, where the rank-dependent payoffs impose greater pressure on them to perform. The dimension of control does not vary with the level of competency.

Although we have established that those competing under tournaments – both winners and losers – face more tension than those who are paid salaries, and that tournament losers have lower levels of interest and feel less competent about their ability than those facing salaries, we do not know what the net effect is on forecast performance. Since T losers show less interest, feel less competent and face greater tension than participants in the S treatment, we postulate that T losers have higher forecast errors than S participants, despite the extrinsic incentives to perform. On the other hand, it is not so clear how T winners would perform relative to S participants.

Although they face higher tension from tournaments which lessens intrinsic motivation, there are also strong monetary incentives to perform.

Table 4.8 Regressions of Forecast Errors: Salary and Tournament Winners and Losers

Dep Var: Forecast Errors	Tournament Winners		Tournament Losers	
	Model 1	Model 2	Model 3	Model 4
Salary	(base)	(base)	(base)	(base)
Tournament Winners	-1.150 (2.002)	1.985 (3.065)		
Tournament Losers			10.51 *** (3.746)	15.82 *** (4.445)
Trait Anxiety	0.032 (0.148)	0.032 (0.148)	-0.005 (0.194)	-0.005 (0.194)
Female	6.308 *** (1.889)	6.308 *** (1.889)	9.607 *** (2.366)	9.607 *** (2.366)
Round	-0.101 (0.092)		-0.159 (0.104)	
S*Round		-0.021 (0.119)		-0.021 (0.119)
T*Round		-0.262 * (0.136)		-0.430 ** (0.195)
Constant	12.23 * (6.287)	11.20 * (6.370)	12.77 (8.364)	10.98 (8.363)
Observations	1590	1590	1605	1605
Participants	106	106	107	107
R^2	0.027	0.027	0.074	0.074
Wald χ^2	18.35	19.08	34.03	38.62
$p > \chi^2$	0.001	0.002	0.000	0.000
S*Round = T*Round		$\chi^2(1) = 1.79$ p = 0.181		**$\chi^2(1) = 3.22$** **p = 0.073**

Tournament winners are defined to be those in the T treatment who wins 8 or more of the 15 tournament rounds, while Tournament losers are those who wins 7 or fewer rounds. Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

We previously reported that forecasting performance is higher in the S treatment compared to the T treatment as a whole. We repeat regression analyses of forecast errors while splitting T winners from losers. Table 4.8 presents these regressions which only include observations from the S treatment and either winners or losers from the T treatment. These regressions are pooled and do not distinguish between single and dual cue tasks.

Table 4.8 shows that T winners perform no differently to S participants. The previous finding that, overall, salaries perform better than tournaments is attributed entirely to the significantly lower performance of T losers. T losers on average have forecast errors 10.51 points higher than those in the S treatment. This difference widens in model 4 when we allow for treatment differences in time trend, for which we observe that T losers improve performance over time, both standalone and relative to S participants.

To sum up, overall, we find that fixed salaries perform better than tournaments. Broken down, this is due to the S treatment performing better than T players who loses more frequently than wins. On the other hand, those T players who wins frequently performs similarly to S players. These findings are reinforced by self-reported measures indicative of intrinsic motivation. T losers are less interested, feel less competent and faces greater tension than S participants. T winners face greater tension than S participants, while reporting similar levels of interest, competency and effort. The patterns of forecast errors and of these psychological measures are commensurate with Cognitive Evaluation Theory. We state this as Result 4.3 below:

Result 4.3. *Tournaments perform worse than fixed salaries. This is due to comparable performance by tournament winners alongside lower performance by tournament losers. These findings are consistent with Cognitive Evaluation Theory.*

While the psychological measures complement our analysis, we are careful not to interpret a causal relationship from these. For example, while it is plausible to suggest that greater tension brought about by greater control in tournaments reduces intrinsic motivation, which in turn contributes to higher forecast errors in the T treatment relative to the S treatment, we cannot empirically draw this inference. This is because the intrinsic motivation variables reported in Table 4.7 are elicited at the end of the game, after players are able to observe their own performance – and in the case of the T treatment, a record of their winning – which may influence how they self-report these measures. For example, it is unclear whether T losers report low levels of competency because they were faced with a large number of losses, or because the losses were brought about by lower intrinsic motivation from low competency. Although the direction of causality cannot be disentangled, mindful of this caveat, the correlation between these measures is nevertheless useful for us to present coherent explanations.

4.3.2. Tournament Decomposition

While our first research question asked how different pay schemes performed relative to one another, our second research question relate to the decomposition of tournament schemes. This decomposition involves separating tournaments into components which relate to competition for rank or competition for payoffs, seeing how each influences performance. Rank competition is modelled with the provision of rank feedback that is not tied in with monetary payoffs.

The first part of our decomposition isolates the effect of rank feedback by controlling for payoffs. We go about this in two ways: a) comparing forecast errors in the PR and PRWL treatments, and b) by comparing them for the Tournament (T) and Tournament-No-Info (TNI) treatments. In the PR and PRWL treatments, both pay participants piece rates for their performance, but the PRWL treatment provides additional information after round 5 as to whether they performed better or worse than a random partner. A between-subjects comparison of these two treatments allow us to infer the effect feedback on relative performance has on performance itself.[31]

In the T and TNI treatments, we focus on tournament pay schemes while manipulating relative feedback. In the T treatment, players observe relative feedback about whether they have won or lost. By comparison, the TNI treatment suppresses this feedback, so that players do not know whether or not they have won the prior rounds. Comparison of forecast errors between

[31] In principle, the effect of relative performance feedback can also be inferred from a within-subjects comparison of participants in the PRWL treatment by comparing forecast errors in rounds 1 to 5, without feedback, to rounds 6 to 20 which provides it. However this comparison is difficult to make due to the changing cue values in each round, and also due to the learning that might occur.

the T and TNI treatments allow us to ascertain the effect of relative performance feedback, independently to the comparison of the PR and PRWL treatments.

The second part of our decomposition compares tournament payoffs to piece rates while controlling for rank feedback. Comparing the PRWL and T treatments, both provide identical winning and losing feedback to participants, but differ by pay scheme. The PRWL treatment pays participants according to piece rates on their forecast errors, while the T treatment pays rank-dependent prizes to participants.

With these two comparisons, we can determine, in a controlled manner, the relative importance that relative performance feedback has on tournament performance vis à vis rank dependent payoffs. This decomposition exercise most closely resembles Eriksson et al. (2009) who have piece rate and tournament treatments with and without relative feedback.

Relative Performance Feedback

We begin by looking at the effect of relative performance feedback by comparing forecast errors of the PR and PRWL treatments in our baseline regressions. In the baseline regressions in Tables 4.4 to 4.6, the coefficient on the PRWL treatment dummy is consistently negative in each of the four regression models in each of the pooled, single and dual cue regressions respectively. Since the base category is the PR treatment, this negative dummy shows that participants in the PRWL treatment make more accurate forecasts than those in the PR treatment, suggesting that the winning/losing feedback in the PRWL treatment has a motivating effect on performance. However, the PRWL dummy is only statistically significant in the pooled data regressions in Table 4.4 in models 2 and 4, where the control variables of trait anxiety and gender are controlled for. The PRWL treatment dummy is not significant in analogous regressions when run individually in the single or dual cue tasks.

Two things can be made out from the pattern of significance of the PRWL dummy across the regressions. The first is that the treatment dummy is only significant when the controls of trait anxiety and gender are incorporated. When these variables are controlled for, they increase the magnitude of the coefficient on the PRWL dummy. This is mainly brought about by the inclusion of both trait anxiety and gender. When either trait anxiety or gender is excluded from the regression, the PRWL dummy is no longer significant at conventional levels (regressions not presented). For this reason, we include the trait anxiety variable even though it is not significant itself.

The second insight here relates to the effect of pooling both single and dual cue data. As mentioned, the PRWL dummy is insignificant in each regression model in Tables 4.5 and 4.6 for the single and dual cue tasks. Despite this, the coefficients are negative in all single and dual cue regressions. By pooling the data, the treatment dummy is now significant with p-values of 0.078 and 0.055 in models 2 and 4 of Table 4.4. It appears that the results are masked by small samples, for which pooling the data assists with by increasing the sample size.

Bearing these two caveats in mind, we find that relative performance feedback itself motivates participants to perform. This is expressed as Result 4.4:

Result 4.4. *Under piece rates, the provision of relative performance feedback improves performance. Rank competition, independent from payoffs, improves performance.*

Comparing Result 4.4 to the findings from previous studies of the effect of relative feedback on performance under piece rates, it differs from Eriksson et al. (2009) who find that relative performance feedback has no effect on productivity, but is in line with Blanes i Vidal and Nossol (2011) who find that German factory workers are more productive when they are provided rank feedback. Result 4.4 is also consistent with the numerous studies that find relative performance feedback motivates performance when provided to people who are paid salaries.

The other way we can analyse the effect of relative performance feedback is by comparing the performance of the T and TNI treatments. In Table 4.9, we look for forecast error differences between the T and TNI treatments. The regressions are run only with observations from the dual cue T and TNI treatments.[32] Regression model 1 presents a basic regression with the TNI treatment dummy (with the T treatment as the reference category) as well as a time trend

[32] This is because we only ran a single session of the TNI treatment with the single cue task. Regressions of this session of the single cue TNI treatment compared to the single cue T treatment does not yield results any different to those presented in Table 4.9 with the dual cue task. To simplify both presentation and analyses, we will focus on the dual cue T and TNI treatments.

common to both treatments. Model 2 includes controls of trait anxiety and gender. Regression models 3 and 4 repeat the previous regressions, but include treatment interacted time trends instead of the common trend.

In each of the four regression specifications, the TNI treatment performs no differently to the T treatment. This suggests that relative performance feedback does not have any effect in the context of tournaments. On the face of things, it seems inconsistent with Result 4.4 that relative performance feedback improves performance when applied to piece rates, although it could simply suggest that relative feedback has different effects under different pay schemes.[33] Similar to our finding here, Eriksson et al. (2009) also finds that relative feedback under tournaments have no effect on performance.

[33] An alternative explanation for why the T and TNI treatments do not perform differently despite differing in terms of feedback is due to the smaller sample with only dual cue data. Notice that in Tables 4.5 and 4.6, for the single and dual cue tasks separately, the PR and PRWL treatments were found to perform similarly to one another.

Table 4.9 Regression of Forecast Errors: Dual Cue T and TNI Treatments

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Tournament	(base)	(base)	(base)	(base)
Tournament No Info	-3.000 (4.337)	-4.042 (4.842)	-2.808 (5.587)	-4.659 (5.959)
Trait Anxiety		0.412 (0.287)		0.412 (0.287)
Female		9.070 ** (4.418)		9.070 ** (4.420)
Round	-0.488 *** (0.144)	-0.471 *** (0.153)		
T*Round			-0.481 ** (0.202)	-0.494 ** (0.213)
TNI*Round			-0.495 ** (0.205)	-0.447 ** (0.220)
Constant	34.64 *** (3.292)	13.26 (13.07)	36.99 *** (3.698)	13.57 (13.38)
Observations	1140	1065	1140	1065
Participants	76	71	76	71
R^2	0.006	0.027	0.006	0.027
Wald χ^2	13.47	15.50	13.47	16.06
$p > \chi^2$	0.001	0.004	0.004	0.007
T*Round = TNI*Round			$\chi^2(1) = 0.00$ p = 0.959	$\chi^2(1) = 0.02$ p = 0.877

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table.

Piece Rates and Rank-Dependent Payoffs

The second part of our tournament decomposition looks at how piece rate and tournament incentives compare to one another once competition has been controlled for, with rank feedback simulating such competition. Drawing from our previous findings, we expect the PRWL treatment to perform better than the T treatment. Result 4.1 found support for Piece Rate

Equivalence, that tournaments perform equally to standard piece rates – that is that the PR and T treatments have similar performance. Result 4.4 found that relative feedback provided to piece rates improved performance, with better performance in the PRWL treatment compared to the PR treatment. Jointly considering both results, we would expect the PRWL treatment to perform better than the T treatment. This is the essence of hypothesis H5.

The baseline pooled regressions in Table 4.4 and the corresponding Wald hypothesis tests shows that the PRWL treatment performs significantly better than the T treatment in every regression model. The differences are marginally significant at the 10% level in regression model 1, but significance improves in latter models, with significance at around 1% in models 3 and 4 when treatment-round interactions are included. Results from the dual cue task in Table 4.6 echo these. Although forecast errors in the single cue PRWL treatment are not statistically different to those in the T treatment, we nevertheless observe the coefficients for the PRWL dummy to be consistently smaller than those for the T dummy.

Having controlled for rank competition, we find that piece rates perform better than the rank-based incentives of tournament. Hypothesis H5 receives support. This is stated as Result 4.5.

Result 4.5. *When relative feedback has been controlled for, piece rates perform better than tournaments.*

The decomposition of tournament schemes is now complete. From Result 4.1, we found that tournaments perform similarly to piece rates. This Piece Rate Equivalence, however, no longer holds when competition is introduced to piece rates. Result 4.5 found that when competition is controlled for, piece rates actually perform better than the rank-dependent payoffs inherent in tournaments. This is explained by Result 4.4, that competition for rank motivates performance in its own right. Piece Rate Equivalence, therefore, relies on the motivating effect of competition in tournaments to perfectly offset its inferior incentives relative to piece rates.

4.4. Checks for Robustness

The five headline results stated earlier were determined from our baseline regressions presented in Tables 4.4 to 4.6. We now reinforce these results with a series of robustness checks which addresses several points that could potentially undermine them.

The first two robustness checks controls for temporal variability in forecast errors that could arise not only through learning, but through the cue values that change exogenously every round. The fact that cue values change every round is potentially problematic in two ways. First, the inherent difficulty of the forecasting task changes with the cue values. The wider the two cues are from one another, the more difficult it is to forecast the underlying value. Second, as a result of the previous point, the variation of forecast errors over time is large, which explains to some extent why our baseline regressions had poor fit. Temporal variability is controlled for in two ways: by controlling for an array of round dummies, and also by standardising our forecast errors so that it has common properties in every round.

The third robustness check repeats the baseline regressions of Tables 4.4 to 4.6 while removing the random error term from the relationship of cue values to the underlying stock price in Equation 3.1. Relative to the average forecast errors in both the single and dual cue tasks, such noise has a relatively large range. To assess the impact of this random term, we recalculate forecast errors based on the noise-free equivalent relationship. We run regressions analogous our earlier regressions and compare the results.

4.4.1. Round Dummies

The first robustness check involves controlling for temporal variability with a series of round dummies in our regressions. Tables 4.10 to 4.12 include a complete array of dummy variables that represent the 15 post-intervention rounds – 14 round dummies plus the reference category. By including these dummies, the round-averaged forecast errors, which vary every round, are calibrated by the respective round dummy. The approach of using round dummies can be thought of as including a unique constant term for each round. This purges time effects that are common to all participants, including the effects of the changing cue values. As a consequence, regressors that vary only over the time dimension are no longer estimable since they can be reconstructed by the series of round dummies in a multi-collinear manner. The linear time trend can no longer be estimated. Tables 4.10 to 4.12 show these regressions for each of the pooled, single and dual cue datasets respectively. Like the previous baseline regressions, regression model 1 includes only the treatment and round dummies as regressors, while model 2 additionally controls for trait anxiety and gender. Models 3 and 4 include interactions between each round dummy with each treatment dummy, without and with these controls respectively. Due to the large number of regressors in these regressions, the coefficients on the round dummies and their treatment interactions are suppressed.

We first check whether the regressions lend support to Result 4.1 regarding Piece Rate Equivalence. From the pooled regressions in Table 4.10, we see that the T treatment dummy is insignificant in all four regression models, indicating that forecast errors are no different from the reference PR treatment. The single and dual cue regressions with round dummies in Tables 4.11 and 4.12 also show the T treatment to perform no differently to the PR treatment, affirming Result 4.1.

There is also support for Result 4.2. The S treatment dummy is consistently negative in every regression model across Tables 4.10 to 4.12, suggesting that forecast errors are lower in the S treatment than the reference PR treatment. It is significant in model 2 of the pooled regressions in Table 4.10, and is almost significant in model 4. The round dummy regressions also provide additional support to Result 4.3, although much weaker than in the previous baseline regressions in Tables 4.4 to 4.6. The coefficient for the S treatment dummy is smaller than that for the T dummy in each of the pooled, single and dual cue regressions, though the differences are only significant in model 2 of the pooled and dual cue regressions.

In terms of our tournament decomposition, the round dummy regressions also provide support to Results 4.4 and 4.5. The PRWL treatment, as before, performs significantly better than the PR treatment when both single and dual cue tasks are pooled, suggesting that rank competition improves performance. Comparing the effectiveness of piece rate and tournament payoffs while controlling for competition, we again see that the PRWL treatment outperforms the T treatment.

Table 4.10 Regression of Forecast Errors with Round Dummies: Pooled

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-1.897 (2.188)	-4.235 * (2.402)	-4.192 * (2.423)	-5.669 ** (2.527)
Tournament	2.052 (2.742)	0.661 (2.948)	-0.482 (2.963)	-2.386 (3.028)
Salary	-1.865 (2.185)	-4.185 * (2.333)	-2.592 (2.522)	-4.106 (2.598)
Trait Anxiety		0.080 (0.121)		0.080 (0.122)
Female		8.173 *** (1.538)		8.173 *** (1.546)
Constant	14.43 *** (1.836)	8.384 * (4.744)	15.80 *** (1.957)	9.498 * (5.026)
Round Dummies	Yes	Yes	Yes	Yes
Treatment-Round Dummy Interactions	No	No	Yes	Yes
Observations	4680	4245	4680	4245
Participants	312	283	312	283
R^2	0.052	0.086	0.057	0.097
Wald χ^2	222.2	260.3	388.2	402.5
$p > \chi^2$	0.000	0.000	0.000	0.000
PRWL = 0	$\chi^2(1) = 0.75$ p = 0.386	**$\chi^2(1) = 3.11$** **p = 0.078**	**$\chi^2(1) = 2.99$** **p = 0.084**	**$\chi^2(1) = 5.03$** **p = 0.025**
PRWL = T	**$\chi^2(1) = 2.70$** **p = 0.100**	**$\chi^2(1) = 4.06$** **p = 0.044**	$\chi^2(1) = 1.97$ p = 0.161	$\chi^2(1) = 1.52$ p = 0.218
T = 0	$\chi^2(1) = 0.56$ p = 0.454	$\chi^2(1) = 0.05$ p = 0.823	$\chi^2(1) = 0.03$ p = 0.871	$\chi^2(1) = 0.62$ p = 0.431
S = 0	$\chi^2(1) = 0.73$ p = 0.393	**$\chi^2(1) = 3.22$** **p = 0.073**	$\chi^2(1) = 1.06$ p = 0.304	$\chi^2(1) = 2.50$ p = 0.114
T = S	$\chi^2(1) = 2.66$ p = 0.103	**$\chi^2(1) = 4.17$** **p = 0.041**	$\chi^2(1) = 0.59$ p = 0.441	$\chi^2(1) = 0.40$ p = 0.527

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Table 4.11 Regression of Forecast Errors with Round Dummies: Single Cue Task

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-0.557 (1.497)	-1.127 (1.739)	-1.762 (1.550)	-2.219 (1.775)
Tournament	-0.145 (1.660)	-0.313 (1.963)	-1.027 (1.512)	-1.731 (1.793)
Salary	-1.149 (1.352)	-1.910 (1.576)	-1.548 (1.473)	-1.934 (1.710)
Trait Anxiety		-0.063 (0.098)		-0.063 (0.099)
Female		1.796 (1.171)		1.796 (1.182)
Constant	7.334 *** (1.188)	9.564 ** (4.206)	7.952 *** (1.303)	10.21 ** (4.572)
Round Dummies	Yes	Yes	Yes	Yes
Treatment-Round Dummy Interactions	No	No	Yes	Yes
Observations	2490	2220	2490	2220
Participants	166	148	166	148
R^2	0.069	0.076	0.081	0.089
Wald χ^2	223.6	240.1	624.7	682.8
$p > \chi^2$	0.000	0.000	0.000	0.000
PRWL = 0	$\chi^2(1) = 0.14$ p = 0.710	$\chi^2(1) = 0.42$ p = 0.517	$\chi^2(1) = 1.29$ p = 0.256	$\chi^2(1) = 1.56$ p = 0.211
PRWL = T	$\chi^2(1) = 0.08$ p = 0.777	$\chi^2(1) = 0.30$ p = 0.584	$\chi^2(1) = 0.41$ p = 0.522	$\chi^2(1) = 0.16$ p = 0.687
T = 0	$\chi^2(1) = 0.01$ p = 0.930	$\chi^2(1) = 0.03$ p = 0.873	$\chi^2(1) = 0.46$ p = 0.499	$\chi^2(1) = 0.93$ p = 0.334
S = 0	$\chi^2(1) = 0.72$ p = 0.395	$\chi^2(1) = 1.47$ p = 0.225	$\chi^2(1) = 1.10$ p = 0.294	$\chi^2(1) = 1.28$ p = 0.258
T = S	$\chi^2(1) = 0.59$ p = 0.441	$\chi^2(1) = 1.43$ p = 0.231	$\chi^2(1) = 0.25$ p = 0.617	$\chi^2(1) = 0.04$ p = 0.850

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Table 4.12 Regression of Forecast Errors with Round Dummies: Dual Cue Task

Dep Var: Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-2.534 (3.444)	-4.523 (3.506)	-6.142 (4.237)	-6.561 (4.034)
Tournament	4.173 (4.429)	2.295 (4.592)	-0.099 (5.254)	-2.370 (5.109)
Salary	-1.498 (3.415)	-4.716 (3.681)	-2.639 (4.412)	-4.673 (4.496)
Trait Anxiety		0.090 (0.193)		0.090 (0.195)
Female		10.08 *** (2.382)		10.08 *** (2.408)
Constant	22.01 *** (3.074)	13.38 *** (7.621)	24.26 *** (3.344)	15.10 * (8.030)
Round Dummies	Yes	Yes	Yes	Yes
Treatment-Round Dummy Interactions	No	No	Yes	Yes
Observations	2190	2025	2190	2025
Participants	146	135	146	135
R^2	0.108	0.140	0.120	0.154
Wald χ^2	341.6	437.1	618.8	687.9
$p > \chi^2$	0.000	0.000	0.000	0.000
PRWL = 0	$\chi^2(1) = 0.54$ $p = 0.462$	$\chi^2(1) = 1.66$ $p = 0.197$	$\chi^2(1) = 2.10$ $p = 0.147$	$\chi^2(1) = 2.64$ $p = 0.104$
PRWL = T	**$\chi^2(1) = 3.28$** **$p = 0.070$**	**$\chi^2(1) = 3.49$** **$p = 0.062$**	$\chi^2(1) = 1.58$ $p = 0.210$	$\chi^2(1) = 0.80$ $p = 0.371$
T = 0	$\chi^2(1) = 0.89$ $p = 0.346$	$\chi^2(1) = 0.25$ $p = 0.617$	$\chi^2(1) = 0.00$ $p = 0.985$	$\chi^2(1) = 0.22$ $p = 0.643$
S = 0	$\chi^2(1) = 0.19$ $p = 0.661$	$\chi^2(1) = 1.64$ $p = 0.200$	$\chi^2(1) = 0.36$ $p = 0.550$	$\chi^2(1) = 1.08$ $p = 0.299$
T = S	$\chi^2(1) = 2.38$ $p = 0.123$	**$\chi^2(1) = 3.50$** **$p = 0.061$**	$\chi^2(1) = 0.26$ $p = 0.609$	$\chi^2(1) = 021$ $p = 0.648$

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

4.4.2. Standardised Forecast Errors

Our second check of robustness involves transforming the forecast errors so that they have common properties in each round, purging the variability that occurs over time. We construct standardised forecast errors so that they have a mean of zero and a standard deviation of one for each treatment in each round. Since the mean and standard deviations of the standardised statistics are identical for every round, the time effects common to all participants are removed.

This standardisation procedure is used by Brown (1995, 1998) to facilitate comparison of forecast errors in each round, removing heteroscedasticity associated with the changing cue values. Following Brown, the transformation subtracts the mean forecast error for each round from each individual's forecast error, and then divides this mean-deviation by the standard deviation of forecast errors in the corresponding round. It is represented by:

$$Standardised\ Forecast\ Error_{it} = \frac{e_{it} - \bar{e}_t}{\sigma_t(e)}$$

where e_{it} denotes the forecast error of participant i in round t. $\bar{e}_t$ is the mean forecast error across all participants for round t, and $\sigma_t(e)$ is the corresponding standard deviation. Since the mean and standard deviation of forecast errors are calculated for each round, cross-round comparison of the standardised forecast errors are interpreted relative to their respective means.

The standardised forecast errors vary over the dimensions of participants and time. For any given round, standardised forecast errors have a mean of zero and a standard deviation of one.[34] Standardised forecast errors have an interpretation akin to mean-deviation – its numerator – where a positive (negative) standardised forecast error means that the participant performed worse (better) than the average in that round. From the formula, the division of the mean-deviation by the standard deviation means that the standardised forecast errors are measured in standard deviation units away from the mean.[35]

[34] See Appendix 3 for a proof of these properties.

[35] An example is Fryer (2011, 2013), who reports the effects of incentives in schools in standard deviation units.

Table 4.13 presents the regressions of standardised forecast errors. Regression models 1 and 2 are estimated for the pooled series; models 3 and 4 for the single cue series; and models 5 and 6 for the dual cue series. The odd-numbered regression models include only the treatment dummies, with the PR treatment serving as the reference category. The even-numbered models also incorporate the controls of trait anxiety and gender. Time series regressors, such as the time trends, are not included since their effect has been purged due to the properties of the standardised forecast errors.

The regressions in Table 4.13 provide additional support for our five results. From the regressions, the T dummy is insignificant in each of the regression models, lending support to Result 4.1 that tournaments perform no differently to piece rates. Salaries perform better than piece rates, supporting Result 4.2. Comparing the T and S treatments, we see that the salaries perform significantly better than tournaments in regression models 1, 3 and 6, supporting Result 4.3. Results 4.4 and 4.5 are also supported. In model 2, we see that forecast errors in the PRWL treatment are 0.18 standard deviations smaller than in the PR treatment, and is significant with a p-value of 0.07. The PRWL treatment also performs better than the T treatment, both in the pooled and dual cue regressions.

Table 4.13 Regression of Standardised Forecast Errors

Dep Var: Standardised Forecast Errors	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PR	(base)	(base)	(base)	(base)	(base)	(base)
PRWL	-0.084 (0.090)	-0.180 * (0.099)	-0.080 (0.111)	-0.125 (0.132)	-0.087 (0.121)	-0.155 (0.123)
T	0.078 (0.112)	0.024 (0.122)	0.005 (0.131)	-0.016 (0.156)	0.131 (0.151)	0.068 (0.158)
S	-0.084 (0.089)	-0.180 * (0.096)	-0.085 (0.107)	-0.151 (0.125)	-0.058 (0.118)	-0.168 (0.127)
Trait Anxiety		0.003 (0.005)		-0.004 (0.007)		0.004 (0.006)
Female		0.320 *** (0.063)		0.109 (0.090)		0.329 *** (0.083)
Constant	0.022 (0.075)	-0.218 (0.194)	0.040 (0.093)	0.216 (0.317)	-0.000 (0.104)	-0.312 (0.254)
Observations	4680	4245	2490	2220	2190	2025
Participants	312	283	166	148	146	135
R^2	0.005	0.037	0.002	0.009	0.007	0.037
Wald χ^2	3.75	27.76	1.25	4.06	3.26	17.89
$p > \chi^2$	0.290	0.000	0.742	0.540	0.354	0.003
PRWL = 0	χ^2 = 0.88 p = 0.349	**χ^2 = 3.29** **p = 0.070**	χ^2 = 0.52 p = 0.471	χ^2 = 0.89 p = 0.345	χ^2 = 0.52 p = 0.473	χ^2 = 1.58 p = 0.208
PRWL = T	**χ^2 = 2.77** **p = 0.096**	**χ^2 = 4.26** **p = 0.039**	χ^2 = 0.60 p = 0.440	χ^2 = 0.90 p = 0.343	**χ^2 = 3.00** **p = 0.083**	**χ^2 = 3.20** **p = 0.074**
T = 0	χ^2 = 0.48 p = 0.489	χ^2 = 0.04 p = 0.845	χ^2 = 0.00 p = 0.969	χ^2 = 0.01 p = 0.920	χ^2 = 0.75 p = 0.387	χ^2 = 0.19 p = 0.664
S = 0	χ^2 = 0.89 p = 0.346	**χ^2 = 3.53** **p = 0.060**	χ^2 = 0.63 p = 0.428	χ^2 = 1.47 p = 0.226	χ^2 = 0.24 p = 0.627	χ^2 = 175 p = 0.186
T = S	**χ^2 = 2.81** **p = 0.094**	**χ^2 = 4.52** **p = 0.033**	χ^2 = 0.72 p = 0.397	χ^2 = 1.52 p = 0.217	χ^2 = 2.34 p = 0.126	**χ^2 = 3.53** **p = 0.060**

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

4.4.3. Deterministic Forecast Errors

As previously stated in Section 3.1 when the forecasting task was outlined, the underlying stock price is based on the underlying relationship earlier stated as Equation 3.1:

$$x_t = 10 + 0.3 \times Cue\ A_t + 0.7 \times Cue\ B_t + \varepsilon_t \tag{3.1}$$

where x_t is the stock price and ε_t is a random term that is uniformly and discretely distributed within $[-5, +5]$. This underlying relationship is probabilistic due to this noise. It could be argued that such noise is excessively large in light of the average forecast error values. The ε has a range of 10, which is large compared to the mean post-intervention forecast errors across single (dual) cue treatments of 10 (23). Since forecast errors incorporate the effect of such noise, to what extent does randomness influence our underlying results?

We test our results for robustness by seeing whether they continue to hold once the randomness has been stripped out. This is done by first working with a hypothetical deterministic relationship:

$$\tilde{x}_t = 10 + 0.3 \times Cue\ A_t + 0.7 \times Cue\ B_t \tag{4.1}$$

where $\tilde{x}_t$ is the hypothetical stock price whereby the noise ε_t does not feature. Since this relationship now depends only upon the two cue values, and is not affected by random noise, this relationship is now a deterministic one. From these hypothetical stock price values, we can construct a new series of forecast errors based on these. The hypothetical forecast error is the absolute difference between the participant's forecast and this hypothetical stock price. This is expressed below:

$$\tilde{e}_{it} = |\hat{x}_{it} - \tilde{x}_t| \tag{4.2}$$

where $\tilde{e}_{it}$ is the hypothetical forecast error based on participant i's forecast $\hat{x}_{it}$. Notice how the noise ε_t is not reflected by these forecast errors. We shall refer to these hypothetical forecast errors as deterministic forecast errors, since these arise from the deterministic relationship. Since we are only removing the noise from the forecast errors, both variables are highly and almost perfectly correlated, with a correlation coefficient of 0.993.

It should be noted that although these deterministic forecast errors are noise-free, they are merely hypothetical. The standard forecast errors which incorporate such noise are recognised publicly as the measure of performance, where they are displayed as feedback round by round, and are used to calculate payoffs under piece rates. In light of this, the deterministic forecast errors are only useful to test the robustness of our previous results.

Tables 4.14 to 4.16 replicate the baseline regressions in Tables 4.4 to 4.6 for each of the pooled, single and dual cue series. Comparing both sets of analogous regressions, we see the estimated coefficients do not differ considerably, and each of our results receives the same level of support from the new regressions.

In terms of pay schemes, the deterministic forecast errors are no different between the PR and T treatments in each regression model in Tables 4.14 to 4.16 with the T treatment dummy being insignificant. This once again confirms Result 4.1 of Piece Rate Equivalence. There is also support for Results 4.2 and 4.3, whereby salaries perform better than both piece rates and tournaments respectively.

In terms of tournament decomposition, we find once again that the provision of relative performance feedback improves performance, where the PRWL treatment performs better than the PR treatment. Comparing piece rates and tournaments when feedback has been controlled for, the PRWL treatment performs better than the T treatment. Results 4.4 and 4.5 are supported once more.

Table 4.14 Regressions of Deterministic Forecast Errors: Pooled

Panel A: Regression Results

Dep Var: Deterministic Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-2.005 (2.217)	-4.370 * (2.438)	-3.825 (2.963)	-6.592 ** (3.206)
Tournament	1.884 (2.768)	0.486 (2.979)	4.684 (3.204)	3.097 (3.410)
Salary	-2.093 (2.209)	-4.449 * (2.360)	-3.359 (2.697)	-6.584 ** (2.861)
Trait Anxiety		0.081 (0.123)		0.081 (0.123)
Female		8.163 *** (1.562)		8.163 *** (1.563)
Round	-0.143 ** (0.063)	-0.157 ** (0.068)		
PR*Round			-0.148 (0.126)	-0.192 (0.142)
PRWL*Round			-0.008 (0.147)	-0.021 (0.153)
T*Round			-0.363 *** (0.110)	-0.393 *** (0.121)
S*Round			-0.051 (0.116)	-0.027 (0.118)
Constant	19.79 *** (1.873)	14.07 *** (4.977)	19.85 *** (2.020)	14.53 *** (5.435)
Observations	4680	4245	4680	4245
Participants	312	283	312	283
R^2	0.005	0.037	0.005	0.038
Wald χ^2	8.95	42.98	16.90	54.27
$p > \chi^2$	0.063	0.000	0.018	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.82$ p = 0.366	**$\chi^2(1) = 3.21$** **p = 0.073**	$\chi^2(1) = 1.67$ p = 0.197	**$\chi^2(1) = 4.23$** **p = 0.040**
PRWL = T	$\chi^2(1) = 2.56$ p = 0.110	**$\chi^2(1) = 3.89$** **p = 0.049**	**$\chi^2(1) = 6.65$** **p = 0.010**	**$\chi^2(1) = 7.87$** **p = 0.005**
T = 0	$\chi^2(1) = 0.46$ p = 0.496	$\chi^2(1) = 0.03$ p = 0.871	$\chi^2(1) = 2.14$ p = 0.144	$\chi^2(1) = 0.83$ p = 0.364
S = 0	$\chi^2(1) = 0.90$ p = 0.343	**$\chi^2(1) = 3.56$** **p = 0.059**	$\chi^2(1) = 1.55$ p = 0.213	**$\chi^2(1) = 5.30$** **p = 0.021**
T = S	$\chi^2(1) = 2.69$ p = 0.101	**$\chi^2(1) = 4.23$** **p = 0.040**	**$\chi^2(1) = 6.89$** **p = 0.009**	**$\chi^2(1) = 9.72$** **p = 0.002**
PR*Round = PRWL*Round			$\chi^2(1) = 0.52$ p = 0.470	$\chi^2(1) = 0.67$ p = 0.413
PRWL*Round = T*Round			**$\chi^2(1) = 3.74$** **p = 0.053**	**$\chi^2(1) = 3.64$** **p = 0.056**
PR*Round = T*Round			$\chi^2(1) = 1.65$ p = 0.198	$\chi^2(1) = 1.16$ p = 0.281
PR*Round = S*Round			$\chi^2(1) = 0.32$ p = 0.570	$\chi^2(1) = 0.79$ p = 0.375
T*Round = S*Round			**$\chi^2(1) = 3.81$** **p = 0.051**	**$\chi^2(1) = 4.67$** **p = 0.031**

Bold typeface indicates statistical significance at the 10% level or better.

Table 4.15 Regressions of Deterministic Forecast Errors: Single Cue Task

Panel A: Regression Results

Dep Var: Deterministic Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-0.659 (1.518)	-1.228 (1.758)	-1.730 (2.934)	-2.984 (3.268)
Tournament	-0.390 (1.703)	-0.548 (2.002)	-0.166 (2.692)	-0.708 (3.102)
Salary	-1.270 (1.382)	-2.049 (1.608)	-1.767 (2.410)	-3.506 (2.717)
Trait Anxiety		-0.066 (0.101)		-0.066 (0.101)
Female		1.726 (1.203)		1.726 (1.204)
Round	-0.162 *** (0.060)	-0.171 *** (0.066)		
PR*Round			-0.188 * (0.112)	-0.242 * (0.130)
PRWL*Round			-0.106 (0.174)	-0.107 (0.179)
T*Round			-0.206 ** (0.087)	-0.230 ** (0.098)
S*Round			-0.150 * (0.083)	-0.130 (0.084)
Constant	11.83 *** (1.486)	14.46 *** (4.648)	12.18 *** (2.023)	15.38 *** (5.076)
Observations	2490	2220	2490	2220
Participants	166	148	166	148
R^2	0.003	0.010	0.004	0.010
Wald χ^2	8.80	16.04	13.03	19.23
$p > \chi^2$	0.066	0.014	0.071	0.023

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.19$ p = 0.664	$\chi^2(1) = 0.49$ p = 0.485	$\chi^2(1) = 0.35$ p = 0.555	$\chi^2(1) = 0.83$ p = 0.361
PRWL = T	$\chi^2(1) = 0.03$ p = 0.859	$\chi^2(1) = 0.19$ p = 0.659	$\chi^2(1) = 0.32$ p = 0.572	$\chi^2(1) = 0.62$ p = 0.433
T = 0	$\chi^2(1) = 0.05$ p =0.819	$\chi^2(1) = 0.07$ p = 0.784	$\chi^2(1) = 0.00$ p = 0.951	$\chi^2(1) = 0.05$ p = 0.820
S = 0	$\chi^2(1) = 0.84$ p = 0.358	$\chi^2(1) = 1.62$ p = 0.203	$\chi^2(1) = 0.54$ p = 0.463	$\chi^2(1) = 1.66$ p = 0.197
T = S	$\chi^2(1) = 0.41$ p = 0.521	$\chi^2(1) = 1.14$ p = 0.286	$\chi^2(1) = 0.53$ p = 0.468	$\chi^2(1) = 1.49$ p = 0.222
PR*Round = PRWL*Round			$\chi^2(1) = 0.16$ p = 0.691	$\chi^2(1) = 0.37$ p = 0.541
PRWL*Round = T*Round			$\chi^2(1) = 0.26$ p = 0.609	$\chi^2(1) = 0.36$ p = 0.546
PR*Round = T*Round			$\chi^2(1) = 0.01$ p = 0.903	$\chi^2(1) = 0.01$ p = 0.940
PR*Round = S*Round			$\chi^2(1) = 0.08$ p = 0.784	$\chi^2(1) = 0.52$ p = 0.470
T*Round = S*Round			$\chi^2(1) = 0.21$ p = 0.644	$\chi^2(1) = 0.60$ p = 0.438

Bold typeface indicates statistical significance at the 10% level or better.

Table 4.16 Regressions of Deterministic Forecast Errors: Dual Cue Task

Panel A: Regression Results

Dep Var: Deterministic Forecast Errors	Model 1	Model 2	Model 3	Model 4
Piece Rate	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-2.610 (3.429)	-4.542 (3.507)	-5.394 (4.849)	-7.615 (5.048)
Tournament	4.078 (4.392)	2.237 (4.572)	9.603 ** (4.820)	7.489 (4.958)
Salary	-1.812 (3.390)	-4.982 (3.662)	-4.110 (4.465)	-8.254 (4.741)
Trait Anxiety		0.092 (0.192)		0.092 (0.192)
Female		9.909 *** (2.358)		9.909 *** (2.360)
Round	-0.123 (0.117)	-0.141 (0.123)		
PR*Round			-0.104 (0.233)	-0.147 (0.243)
PRWL*Round			0.110 (0.246)	0.090 (0.263)
T*Round			-0.529 *** (0.205)	-0.551 ** (0.216)
S*Round			0.072 (0.237)	0.105 (0.247)
Constant	28.35 *** (2.983)	20.22 *** (7.898)	28.12 *** (3.087)	20.30 ** (8.637)
Observations	2190	2025	2190	2025
Participants	146	135	146	135
R^2	0.007	0.035	0.009	0.036
Wald χ^2	5.52	25.47	13.56	33.54
$p > \chi^2$	0.238	0.000	0.060	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3	Model 4
PRWL = 0	$\chi^2(1) = 0.58$ $p = 0.447$	$\chi^2(1) = 1.68$ $p = 0.195$	$\chi^2(1) = 1.24$ $p = 0.266$	$\chi^2(1) = 2.28$ $p = 0.131$
PRWL = T	**$\chi^2(1) = 3.35$** **$p = 0.067$**	**$\chi^2(1) = 3.53$** **$p = 0.060$**	**$\chi^2(1) = 8.12$** **$p = 0.004$**	**$\chi^2(1) = 7.38$** **$p = 0.007$**
T = 0	$\chi^2(1) = 0.86$ $p = 0.353$	$\chi^2(1) = 0.24$ $p = 0.625$	**$\chi^2(1) = 3.97$** **$p = 0.046$**	$\chi^2(1) = 2.28$ $p = 0.131$
S = 0	$\chi^2(1) = 0.29$ $p = 0.593$	$\chi^2(1) = 1.85$ $p = 0.174$	$\chi^2(1) = 0.85$ $p = 0.357$	**$\chi^2(1) = 3.03$** **$p = 0.082$**
T = S	**$\chi^2(1) = 2.65$** **$p = 0.103$**	**$\chi^2(1) = 3.82$** **$p = 0.051$**	**$\chi^2(1) = 7.80$** **$p = 0.005$**	**$\chi^2(1) = 9.45$** **$p = 0.002$**
PR*Round = PRWL*Round			$\chi^2(1) = 0.40$ $p = 0.528$	$\chi^2(1) = 0.44$ $p = 0.510$
PRWL*Round = T*Round			**$\chi^2(1) = 3.98$** **$p = 0.046$**	**$\chi^2(1) = 3.54$** **$p = 0.060$**
PR*Round = T*Round			$\chi^2(1) = 1.88$ $p = 0.171$	$\chi^2(1) = 1.54$ $p = 0.214$
PR*Round = S*Round			$\chi^2(1) = 0.28$ $p = 0.595$	$\chi^2(1) = 0.53$ $p = 0.468$
T*Round = S*Round			**$\chi^2(1) = 3.69$** **$p = 0.055$**	**$\chi^2(1) = 3.99$** **$p = 0.046$**

Bold typeface indicates statistical significance at the 10% level or better.

4.5. Summary

This chapter analysed our main treatment effects and addressed our first two research questions: which pay scheme induces the highest performance from workers; and what effects do relative performance feedback and rank-dependent payoffs have in the performance of tournaments?

First comparing the different pay schemes, we find that fixed salaries perform better than piece rates and tournaments. The pay schemes that pay for performance elicit comparatively low effort from participants, resulting in lower performance relative to salaries. Tournaments and piece rates perform similarly to one another, consistent with the Piece Rate Equivalence property of tournament theory.

The performance ordering of pay schemes is consistent with Cognitive Evaluation Theory. Additional support is found when we compare the T and S treatments more closely. Cognitive Evaluation Theory posits that intrinsic motivation is reduced when players are put in situations which are controlling, reducing their autonomy. This is reflected in the data, whereby players in the T treatment report higher levels of tension compared to those in the S treatment. Cognitive Evaluation Theory also suggests that people's intrinsic motivation improves as they receive favourable feedback regarding their competency, while reduces as they receive unfavourable feedback. When we categorise participants in the T treatment by whether or not they have won more than half of their post-intervention rounds, we found that T losers report both lower interest and competency than T winners. T winners report similar levels of interest, competency and effort as S participants, and in fact, their levels of forecast performance are not statistically different.

Our second research question looks deeper into the decomposition of factors that motivate tournaments to perform. We look at the effects of tournament feedback and payoffs. We found that when relative performance feedback was provided to players who were paid a piece rate on their performance, the performance of these players improved. In other words, the competition brought about by feedback has a strong motivating effect on players. This finding has real implications for the workplace. Given that relative feedback is inexpensive to provide, it would be a cost effective way for firms to boost worker productivity.

To complete our decomposition of tournaments, we ask how the rank-dependent payoffs under tournaments compare to piece rates when feedback is perfectly controlled for. We find that, purely in terms of the incentives at play, piece rates outperform the rank-dependent payoffs inherent in tournaments. This in turn implies that the property of Piece Rate Equivalence relies crucially on relative feedback for it to hold.

5. Results: Learning

The previous chapter took an aggregated view of how different pay schemes and feedback performed relative to one another. These aggregated differences can be thought of as the one-off treatment effect. In this chapter, we complement our previous analyses by studying how the treatment effects change over time. The temporal element is particularly interesting in the context of our experiment. First of all, it is not unreasonable to expect these effects to be dynamic and occur over time. Second, learning is especially interesting given that our experimental task is difficult; so there is much scope for it to occur.

We need to make a distinction between two aspects of learning: the learning process and the results that arise from this. The 'results' aspect refer to the outcome or performance, and whether it improves over time or not.[36] This essentially measures the effectiveness of the learning process, which looks at how people come about acquiring knowledge or skills that helps them to perform. The learning process is multi-dimensional: involving heuristics and rules (Roth & Erev, 1995; Erev & Roth, 1998; Charness & Levin, 2005), feedback (Rick & Weber, 2010) and observation (Merlo & Schotter, 2003; Cardella, 2012), and payoffs (Merlo & Schotter, 1999).

We note that learning in our experiment is different to that in similar studies. The reason is related to the choice of experimental task. Of the experimental studies similar to ours which are run over a number of rounds, the real-effort task that is chosen is simple. For example Kuhnen and Tymula (2012) and Cadsby et al. (2010) use an arithmetic task, while Charness et al. (2014) use a decoding task. These tasks are similar to each other in that they mainly rely on effort in order to do well. Even if performance improves over time, there is nothing to 'learn' per se. Our task is different in that it is cognitively challenging. In order to improve forecasts, participants need to uncover the underlying relationship of cues and the actual value, or at least get as close to it as possible. This requires the formulation of a forecasting rule – which we do not observe as experimenters – and subsequent refinement of this rule through trial, error and reinforcement. The learning process here is clearly quite different to that in a simpler task.

[36] The process of learning is often repetitive and iterative in nature, hence learning is often referred to in a temporal context.

With respect to learning, our task relates more closely to those used to specifically study the processes and mechanics of learning. Underpinning these tasks or games is a definitive way that it should be played. For example, in the Merlo and Schotter (1999, 2003) maximisation problem, payoffs are maximised at the equilibrium effort level, which players should be searching for to maximise their payoffs. In multi-player strategic games (Charness & Levin, 2005; Rick & Weber, 2010; Cardella, 2012), the 'way to play' is often prescribed as a dominant strategy (or at least one that is not dominated) which players should learn to play over time. Our task is similar to these in that the definitive way to do well in the forecasting task is to deduce the underlying formula.

One potential difficulty we encounter when studying learning in our forecasting task is that we do not actually observe the forecasting rule that participants base their forecasts upon. Despite this, we are still able to investigate learning by looking at how forecast errors change over time. Since forecast errors measure how accurate forecasts are in relation to the underlying stock value, it also measures how well a particular forecasting rule performs, provided that forecasts were made from this latent rule. Most of the forthcoming analyses in this chapter will focus on forecast errors as the main variable with respect to learning, since we would expect that if learning occurs, forecasts will become increasingly accurate.

In addition to the accuracy of forecasts, we also look at the consistency of forecast performance. As a participant's forecast rule tends towards the underlying formula, we would expect forecast errors for this participant to become increasingly consistent. We measure consistency of forecast errors with the within-subject standard deviation of forecast errors defined over various three-round periods and track how they change across these periods.

This chapter proceeds in the following manner. We first look for evidence of learning, seeing whether forecast accuracy or consistency improves over time in any of our treatments. We find that learning is only present in the T treatment, and this is supported by a string of robustness checks. Following from this we look into the underlying mechanisms that drive learning in tournaments and we back this up with a new line of evidence from the Tournament-No-Info treatment. The next sections looks at learning from a different perspective by analysing how learning comes about, looking at how feedback is utilised in the process of learning. More specifically, we look for signs of reinforcement learning, by seeing how forecast error feedback is used to improve forecasts in subsequent rounds. The final section looks at whether winning or losing in previous rounds affects forecast errors. The chapter wraps up with a summary and discussion.

5.1. Forecast Accuracy

We start off by seeing how forecast accuracy changes over time across different treatments. To study this, we regress forecast errors against treatment dummies as well as treatment-interacted time trends, building upon the baseline regressions from Tables 4.4 to 4.6. We will mainly refer back to model 4 of the baseline pooled, single and dual cue regressions in Tables 4.4 to 4.6, focusing on the treatment-round interaction terms, which represent linear time trends specific to each treatment. These regression models regressed forecast errors against the treatment dummies (with the PR treatment serving as the reference category), the trait anxiety scores and gender for each participant, as well as treatment-interacted time trends. We replicate these in models 1, 3 and 5 of Table 5.1. To simplify the presentation of results, we present only the treatment-round interaction terms, suppressing the other regressors: treatment dummies, trait anxiety and gender.

The odd-numbered regression models in Table 5.1 show the trends across rounds 6 to 20. However comparison of learning across these rounds is arguably unfair. While the treatment intervention kicks in after round 5 in the PRWL and T treatments, the PR and S treatments continue playing the game in the exact same manner after round 5. This means that, to some extent, we are comparing subjects experienced with their experimental environment to subjects who are only beginning to adapt to the interventions that have just come into play.

To facilitate comparison, we repeat the same regressions but run with observations from rounds 11 to 20, allowing an arbitrary five rounds for participants in the PRWL and T treatments to adapt. Regression models 2, 4 and 6 show the trends across the final ten rounds. Wald tests comparing the estimated trends across treatments are presented at the bottom of the table.

We begin by looking at the time trends across rounds 6 to 20. From Table 5.1, models 1, 3 and 5 show that the estimated trends for the PR, PRWL and S treatments are not statistically different to zero. Accordingly, Wald tests show that there are no cross-treatment differences in the trends estimated for these treatments. That is, there are no learning differences between the PR and PRWL treatments, and between the PR and S treatments.

While there are no distinct forecast error time trends in the PR, PRWL and S treatments, there is a clear downward trend in the T treatment. In model 1, the time trend for the T treatment has a negative coefficient of -0.348, which is highly significant at the 1% level. This means that forecast errors improve on average by 0.348 points every round in the T treatment. Wald tests at the bottom of Table 5.1 reveal that while this trend is in itself not significantly different from the PR trend, there are significant differences between the PRWL, and the S treatments.

The pooled result in model 1 which shows that the T treatment has a significant downward trend across rounds 6 to 20 is corroborated by the analogous dual cue regression in model 5, which shows that the T trend is significantly different to that in the PRWL and S treatments. In model 3 for the single cue task, even with a different set of cue values, the time trend for the T treatment is again negative and significant. However pairwise tests show that the T trend is not statistically different to those in other treatments.

In the pooled regression for rounds 11 to 20 in model 2 of Table 5.1, the T*Round coefficient is negative and is substantially smaller than the other trends. It is again the only treatment to have a statistically significant time trend. The trend on the T treatment across rounds 11 to 20 is more than twice as steep compared to the corresponding trend across rounds 6 to 20 in model 1 (-0.776 vs -0.348). Across rounds 11 to 20 in the single cue task, in model 4, there is evidence show that forecasts improve over time in all treatments. Despite this, the estimated trend is steeper in the T treatment than in other treatments, though is only significantly so compared to the PR treatment (p = 0.031). The T time trend does not appear to be significant across rounds 11 to 20 in the dual cue task, although the trend has again been estimated to indicate better learning.

From Table 5.1, we have found that the S treatment does not show any learning over time. Comparing the trends of the S treatment with the PR treatment, and bearing Result 4.2 in mind which showed that the S treatment outperformed the PR treatment, we can conclude that fixed pay schemes induce better performance from participants compared to piece rates and this effect is not dynamic.

Table 5.1 Random Effects Regressions of Forecast Error Trends

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PR*Round	-0.141 (0.146)	-0.214 (0.226)	-0.158 (0.140)	-0.701 *** (0.135)	-0.127 (0.246)	0.219 (0.397)
PRWL*Round	0.003 (0.153)	-0.349 (0.338)	-0.038 (0.181)	-0.710 ** (0.323)	0.055 (0.261)	0.114 (0.646)
T*Round	-0.348 *** (0.119)	-0.776 *** (0.186)	-0.197 ** (0.096)	-1.216 *** (0.203)	-0.494 ** (0.212)	-0.347 (0.294)
S*Round	0.021 (0.118)	-0.326 (0.292)	-0.116 (0.077)	-0.953 *** (0.195)	0.101 (0.251)	0.483 (0.591)
Observations	4245	2830	2220	1480	2025	1350
Participants	283	283	148	148	135	135
R^2	0.036	0.036	0.009	0.032	0.035	0.035
Wald χ^2	50.50	61.99	15.07	131.32	31.80	29.33
$p > \chi^2$	0.000	0.000	0.089	0.000	0.000	0.001
PR*Round = PRWL*Round	$\chi^2 = 0.47$ p = 0.494	$\chi^2 = 0.11$ p = 0.741	$\chi^2 = 0.28$ p = 0.599	$\chi^2 =0.00$ p = 0.980	$\chi^2 = 0.26$ p = 0.612	$\chi^2 = 0.02$ p = 0.890
PRWL*Round = T*Round	**$\chi^2 = 3.27$** **p = 0.071**	$\chi^2 = 1.22$ p = 0.270	$\chi^2 = 0.59$ p = 0.442	$\chi^2 = 1.76$ p = 0.185	$\chi^2 = 2.66$ p = 0.103	$\chi^2 = 0.42$ p = 0.516
PR*Round = T*Round	$\chi^2 = 1.20$ p = 0.274	**$\chi^2 = 3.68$** **p = 0.055**	$\chi^2 = 0.05$ p = 0.822	**$\chi^2 = 4.46$** **p = 0.035**	$\chi^2 = 1.28$ p = 0.258	$\chi^2 = 1.31$ p = 0.252
PR*Round = S*Round	$\chi^2 = 0.41$ p = 0.521	$\chi^2 = 0.09$ p = 0.761	$\chi^2 = 0.07$ p = 0.794	$\chi^2 = 1.13$ p = 0.287	$\chi^2 = 0.42$ p = 0.516	$\chi^2 = 0.14$ p = 0.711
T*Round = S*Round	**$\chi^2 = 3.77$** **p = 0.052**	$\chi^2 = 1.68$ p = 0.195	$\chi^2 = 0.41$ p = 0.524	$\chi^2 = 0.87$ p = 0.352	**$\chi^2 = 3.28$** **p = 0.070**	$\chi^2 = 1.58$ p = 0.208

Partial results only. Other coefficients that are suppressed include: treatment dummies, trait anxiety, gender and the constant term. Models 1, 3 and 5 are replicated from model 4 of Tables 4.4, 4.5 and 4.6. Regressions are estimated with a Random Effects GLS procedure. Forecast errors in parentheses and are clustered by participant. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald χ^2 tests are presented at the bottom of the table. Bold typeface indicates a Wald test to be significant at the 10% level or better.

Overall it seems that the T treatment is the only one which shows consistent signs of learning. In what follows, we will focus on the T treatment alongside the directly comparable PR and

PRWL treatments. We restrict focus on treatments with performance pay schemes since we do not find any learning in the S treatment. For this reason we will proceed without the S treatment.

We replicate the finding that the T treatment is the only one with significant learning with a series of robustness checks. The first robustness check is in the form of a fixed effects regression. Random effects regressions were previously used to analyse the treatment effects in Chapter 4 because fixed effects regressions were unable to estimate the treatment dummies, which varies across participants but not across time. Since we focus on learning in this chapter, we need not estimate the overall time-invariant treatment effect – in which case fixed effects regressions are appropriate. Fixed effects regressions control for the time-invariant heterogeneity of participants through a de-meaning procedure, which means we can have more confidence in the results since these are not affected by participants' initial ability and various other traits, many of which are unobservable. The fixed effects estimator is statistically consistent.

Table 5.2 presents the linear time trends estimated from a series of fixed effects regressions, with a couple of cross-treatment F tests at the bottom of the table. A comparison of the trends estimated under both random and fixed effects show that the trends for the PR and T treatments tend to be slightly smaller in magnitude under fixed effects estimation, while they are slightly larger for the PRWL treatment. Despite the minor discrepancies in the estimated trends, the basic pattern of results still holds true. The T treatment shows significant learning in every regression model except the last, for rounds 11 to 20 in the dual cue task – this trend was also insignificant under random effects estimation in Table 5.1. Like before, there is also evidence of learning across rounds 11 to 20 in all single cue treatments, and this learning is greatest in the T treatment.

The second robustness check applies the Pesaran and Smith (1995) mean-group estimator. They show that coefficients estimated from 'pooled' regressions, where the t observations of the n participants are combined and common trends and intercepts are imposed, are not consistent when the regression model is dynamic. This problem arises because potential heterogeneity across participants is unaccounted for. The mean-group estimator that Pesaran and Smith proposed overcomes this problem. Instead of pooling the t observations of the n participants and estimating a common trend for them, the mean-group estimator estimates a unique trend for all *i* participants in an OLS regression, then subsequently takes the mean of these trends. This procedure means that individual-specific, but time-invariant variables – such as treatment dummies, trait anxiety and gender – cannot be estimated.

Table 5.2 Fixed Effects Regressions of Forecast Error Trends

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PR*Round	-0.096 (0.129)	-0.212 (0.198)	-0.100 (0.118)	-0.632 *** (0.145)	-0.092 (0.236)	0.239 (0.369)
PRWL*Round	0.010 (0.147)	-0.361 (0.333)	-0.042 (0.177)	-0.797 ** (0.326)	0.072 (0.244)	0.161 (0.609)
T*Round	-0.322 *** (0.109)	-0.726 *** (0.172)	-0.170 * (0.090)	-1.010 *** (0.187)	-0.481 ** (0.201)	-0.333 (0.279)
Observations	3540	2360	1860	1240	1680	1120
Participants	236	236	124	124	112	112
R^2	0.002	0.000	0.000	0.009	0.004	0.003
F	3.07	6.75	1.46	19.81	1.99	0.64
p > F	0.029	0.000	0.228	0.000	0.120	0.590
PR*Round = PRWL*Round	F = 0.29 p = 0.588	F = 0.15 p = 0.701	F = 0.08 p = 0.784	F = 0.21 p = 0.645	F = 0.23 p = 0.631	F = 0.01 p = 0.913
PRWL*Round = T*Round	**F = 3.28** **p = 0.071**	F = 0.95 p = 0.331	F = 0.42 p = 0.517	F = 0.64 p = 0.424	**F = 3.06** **p = 0.083**	F = 0.54 p = 0.462
PR*Round = T*Round	F = 1.78 p = 0.183	**F = 3.84** **p = 0.052**	F = 0.23 p = 0.635	**F = 3.89** **p = 0.051**	F = 1.57 p = 0.212	F = 1.54 p = 0.218

Regressions are estimated with a Fixed Effects procedure. Forecast errors in parentheses and are clustered by participant. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. F tests are presented at the bottom of the table. Bold typeface indicates an F test to be significant at the 10% level or better.

The time trends estimated with the Pesaran Smith mean-group estimator are presented in Table 5.3. Each coefficient in Table 5.3 represents the linear time trend averaged across all participants in each treatment sub-group, estimated from separate regressions. For this reason we do not formally compare these mean group trends across treatments. Instead we will qualitatively compare the estimated trends to those previously estimated.

Table 5.3 Pesaran Smith Mean Group Estimator of Time Trends

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	(1)	(2)	(3)	(4)	(5)	(6)
PR*Round	-0.096 (0.129)	-0.212 (0.199)	-0.100 (0.119)	-0.632 *** (0.146)	-0.092 (0.238)	0.239 (0.372)
PRWL*Round	0.010 (0.147)	-0.361 (0.334)	-0.042 (0.178)	-0.797 ** (0.329)	0.072 (0.246)	0.161 (0.615)
T*Round	-0.322 *** (0.110)	-0.726 *** (0.172)	-0.170 * (0.090)	-1.099 *** (0.189)	-0.481 ** (0.203)	-0.333 (0.281)

The coefficients presented for each cell have been estimated under a separate Pesaran Smith procedure for each (pooled, single or dual cue) treatment and for each time period. This allows us to distinguish different levels of aggregation for which averaging takes place. As such, we cannot formally test differences between the reported trends.

A comparison of the mean-group estimated trends in Table 5.3 with the fixed effects trends in Table 5.2 show that they are virtually identical. This means that the results confirm the basic learning result. The T treatment stands out being the only one which shows learning over time. Although forecast errors improve over time across rounds 11 to 20 in the single cue task, learning is still more pronounced in the T treatment than in the others.

A third robustness check relaxes the implicit assumption that learning occurs in a linear manner, where the same magnitude of improvement occurs in every round. We repeat analyses by regressing forecast errors against the inverse of the Round variable. This is essentially an asymptote converging towards zero. Learning is conceived to occur at a diminishing rate, which we regard as a realistic assumption. Regression results are presented in Table 5.4.

The asymptotic trends in Table 5.4 show a familiar pattern of results. Since the asymptote is naturally downward sloping to begin with, positive coefficients indicate the downward sloping trend, while negative coefficients represent an upward sloping trend. The magnitude of the coefficient relates to the curvature. We see that the asymptotic trend for the T treatment is always positive, although is not statistically significant in regression models 3 and 6. Across rounds 6 to 20, the pooled and dual cue regressions in models 1 and 5 show the T treatment to have a significant downward trend, while the other treatments have trends that are not statistically different to zero. Interestingly, the single cue regression across rounds 6 to 20, model 3, no longer shows the T treatment to have a significant trend. Across rounds 11 to 20, the pattern of significance is similar to what we have seen before. In the pooled regression only the T treatment has a significant trend; all treatments have a significant trend in the single cue regression; and no treatments have significant trends in the dual cue regression.

Table 5.4 Fixed Effects Regressions of Non-Linear Forecast Error Trends

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PR / Round	12.63 (16.22)	48.98 (46.68)	5.667 (14.79)	144.7 *** (32.68)	20.12 (29.75)	-54.06 (87.68)
PRWL / Round	-1.054 (15.35)	107.2 (71.08)	-6.085 (18.66)	192.3 *** (67.68)	4.983 (25.36)	5.067 (132.1)
T / Round	29.32 ** (14.83)	152.8 *** (41.48)	3.368 (12.94)	240.6 *** (38.47)	56.64 ** (26.60)	60.43 (72.18)
Observations	3540	2360	1860	1240	1680	1120
Participants	236	236	124	124	112	112
R^2	0.004	0.002	0.000	0.011	0.008	0.002
F	1.51	5.65	0.11	22.26	1.68	0.36
p > F	0.214	0.001	0.956	0.000	0.176	0.781
PR / Round = PRWL / Round	F = 0.38 p = 0.541	F = 0.47 p = 0.494	F = 0.24 p = 0.623	F = 0.40 p = 0.527	F = 0.15 p = 0.699	F = 0.14 p = 0.710
PRWL / Round = T / Round	F = 2.02 p = 0.156	F = 0.31 p = 0.580	F = 0.17 p = 0.678	F = 0.38 p = 0.536	F = 1.98 p = 0.163	F = 0.14 p = 0.714
PR / Round = T / Round	F = 0.58 p = 0.448	**F = 2.77** **p = 0.098**	F = 0.01 p = 0.907	**F = 3.61** **p = 0.060**	F = 0.84 p = 0.362	F = 1.02 p = 0.316

Regressions are estimated with a Fixed Effects procedure. Forecast errors in parentheses and are clustered by participant. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. F tests are presented at the bottom of the table. Bold typeface indicates an F test to be significant at the 10% level or better.

All in all it seems that the T treatment is the only one which shows consistent signs of learning. The improvement in forecast errors over time however should not be interpreted in isolation. The previous chapter found the T treatment, along with the PR treatment, to be one of the lowest performing treatments. A natural question to ask at this point is whether the learning in the T treatment is adequate to make up for the lower performance vis à vis the PRWL treatment. If learning occurs gradually round by round – as we model it – then we should expect the cumulative improvements in performance to be the greatest in the final round. Using parameters estimated from the regressions in Table 5.1, we can construct the predicted forecast errors in round 20 for each treatment. We calculate the round 20 treatment differences as:

$$Predicted\ Round\ 20\ Forecast\ Error_k = \beta_0 + \beta_k + \beta_{k*Round} \times 20 \qquad (5.1)$$

where β_k and $\beta_{k*Round}$ are the coefficients of the treatment dummy and treatment-round interaction term for treatment k, while β_0 is the constant term. k indexes the PR, PRWL, T treatments, with $\beta_k = 0$ for the PR treatment, representing the reference category for the treatment dummies. These coefficients are taken from the regressions that underpin Table 5.1. Since we are mainly interested in the ceteris paribus treatment differences, our calculation does not include the effect of the other regressors that feature in the regression model, such as gender. These forecast error predictions are presented in Table 5.5.

Panel A shows the predicted forecast errors based on the regressions run over rounds 6 to 20. We see that the PR treatment appears to do poorly overall. In round 20, holding the control variables constant, forecast errors are estimated to be highest in the PR treatment in each of the pooled, single and dual cue regression models. While the T treatment performed poorly to start off with, the faster pace of learning makes it perform better than the PR treatment by the last round. In the single cue task, the predicted forecast errors in round 20 are slightly lower in the T treatment than in the PRWL treatment. On the other hand, in the dual cue task it appears that the greater rate of learning in the T treatment is inadequate to close the initial performance gap with the PRWL treatment. This is also reflected in the pooled predictions.

From Table 5.1 we previously found that the trend on the T treatment was much steeper across rounds 11 to 20 than across rounds 6 to 20. In Panel B of Table 5.5 with the trends estimated across rounds 11 to 20, we see a similar pattern of predictions as in Panel A. The PR treatment performs poorly overall. In the single cue task the T treatment is able to slightly outperform the PRWL treatment with the faster rate of learning, but this is not true in the dual cue task.

Our analyses here is limited somewhat by our experimental design, with only 20 rounds. If the estimated rates of learning in these treatments could be extrapolated to a greater number of rounds, the T treatment would be expected to perform better than each of the other treatments due to the faster rate of learning even when initial performance lagged behind. We already found evidence that by round 20, the T treatment in the single cue task would slightly outperform the PRWL treatment. In the dual cue task, if we held the estimated coefficients constant and hypothetically allow the number of rounds to increase, we would see the T treatment perform better than the PRWL treatment if the number of rounds increased to 27. If this holds true, there are policy implications regarding how workers should be paid to maximise their productivity.

Table 5.5 Predicted Forecast Errors in Round 20

Panel A: Rounds 6 to 20

	Pooled	Single Cue	Dual Cue
PR	11.20	11.37	17.27
PRWL	7.977	11.09	14.03
T	10.42	10.79	17.00

Panel B: Rounds 11 to 20

	Pooled	Single Cue	Dual Cue
PR	10.38	7.485	18.61
PRWL	6.245	6.554	14.53
T	8.281	5.142	17.41

5.2. Forecast Consistency

We have previously analysed how forecast errors trend over time and have found that the T treatment is the only one that showed any consistent signs of learning. Since forecast errors measure the accuracy of forecasts, this tells us that the accuracy of forecasts improves over time in the T treatment. Another aspect of learning relates to consistency. If a participant displays significant learning over time, we would not only expect the accuracy of their forecasts to improve, but also that their forecast errors become increasing consistent. If learning does indeed occur, then we would expect people to settle on a forecast rule as it converges to the underlying formula.

In settling upon a given forecast rule or not making major adjustments to it, forecast errors should become more consistent, irrespective of its accuracy. In conjunction, measures of both accuracy and consistency better reflect the notion of learning.

The analysis of forecast accuracy can help us rule out the possibility that learning, or its absence, is driven by luck. Due to the nature of real-effort tasks, we only observe the forecasts that participants make rather than the underlying heuristics and rules that they apply. As a result we do not know whether the forecasts were random guesses or were the result of a carefully thought through process. Moreover, there is a possibility that random guesses yield small forecast errors out of luck. If we observe the pattern that forecast errors are becoming increasingly consistent over time, we are more likely to attribute this behaviour to learning than luck, since it would be quite unlikely that luck would consistently work in the direction of learning. If we can confirm in the forthcoming analyses that forecast errors are indeed becoming more consistent, we can safely interpret the downward sloping trends as learning.

We measure the consistency of forecasts with the standard deviation of each participant's forecast errors across the dimension of time. Since we want to see how this changes over time, we break up the 15 post-intervention rounds into 5 three-round blocks. The within-subject standard deviation is calculated for each of these round-blocks. This yields five standard deviation observations for each participant. The treatment-averages of these are plotted in Figure 5.1.

Panel A of Figure 5.1 shows the within-subject standard deviations for the single cue treatments. Notably in the first block, rounds 6 to 8, the standard deviations were initially highest in the T treatment, with an average of just below 8 forecast error points. Although there is some variation in these within-subject standard deviations over time, we see that in the final block, rounds 18 to 20, the standard deviations are lowest in the T treatment with a value of 6. By contrast, these end up higher in the PR and PRWL treatments relative to the first block.

For the dual cue task, Panel B, improvements in forecast error consistency is clear in the T treatment. Standard deviations are substantially higher in the T treatment in the first block compared to other treatments, while are lowest in the final block. A clear downward trend can be made out inbetween these first and last points. In comparison, there does not appear to be any significant trend in standard deviations in the other dual cue treatments.

We formalise these findings by estimating a linear time trend in a random effects generalised least squares regression. Within-subject standard deviation is regressed against a time trend, as well as the trait anxiety score and gender of the participant. These regressions are presented in Table 5.6. Model 1 shows the trend for single and dual cue treatments pooled together, while models 2 and 3 present these separately. The trends are denoted by the 'Block' variable interacted by treatment.

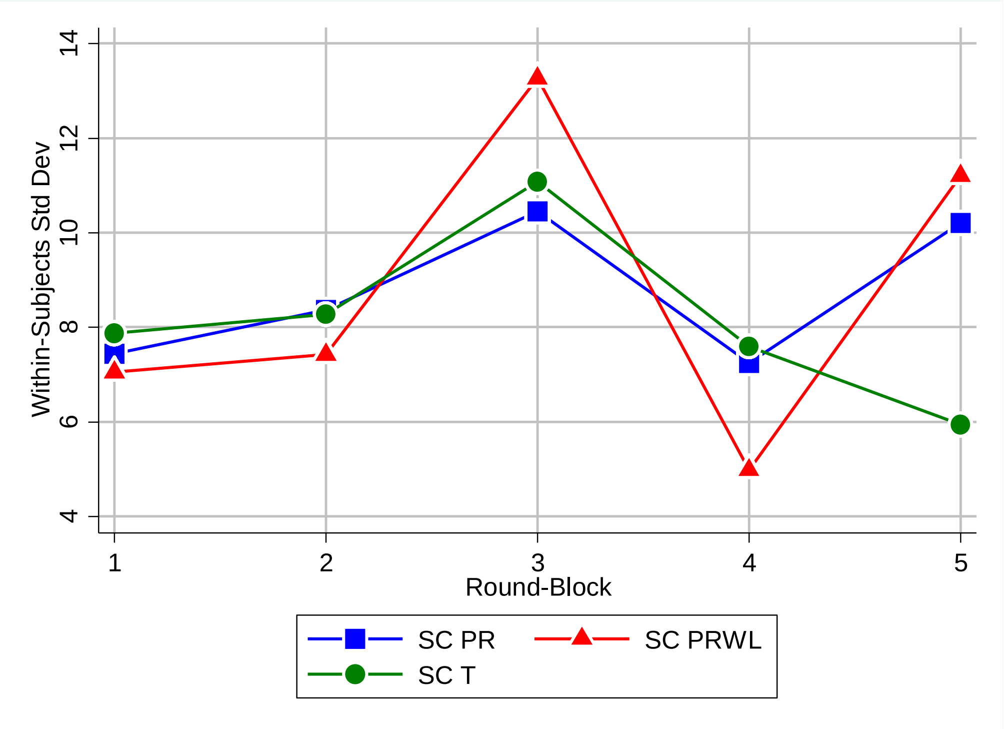
The estimated within-subject standard deviation trends are insignificant in the PR and PRWL treatments across the single and dual cue tasks, and also when these observations are pooled. The T treatment stands out. In regression model 1, we see that forecast error standard deviations reduce on average by 1.21 points across each three-round block – significant at the 5% level. Wald tests show that consistency improves at a greater rate in the T treatment than in the PR and PRWL treatments. A similar pattern of learning is observed in the dual cue treatments in model 3. In the single cue task, we do not observe improved consistency in any treatment – although we do see that the trend in the T treatment is negative while the trends for other treatments is positive.

Our finding that forecast errors become increasingly consistent over time in the T treatment supplement the previous finding that forecasts improve over time in the T treatment. This strengthens our claim that there is substantial learning in the T treatment, while being absent in the other treatments.

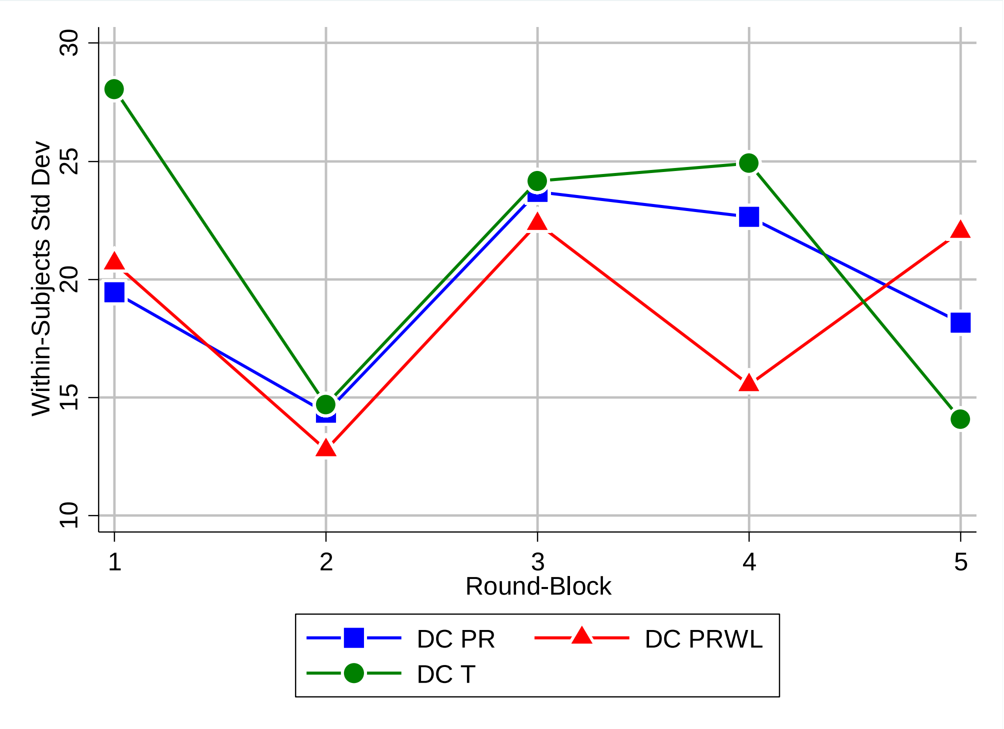
Given that there appears to be no differences in how forecast error consistency changes over time across the PR and PRWL treatments, the improving consistency in the T treatment over time is attributable to the rank-dependent payoffs in tournaments as opposed to relative feedback. In this regard, it is similar to what was established in the previous result that rank payoffs drive improvements in forecast errors over time.

Figure 5.1 Within-Subject Standard Deviation of Forecast Errors across Time

Panel A: Single Cue Task



Panel B: Dual Cue Task



Round-Block 1 for rounds 6-8, block 2 for rounds 9-11, block 3 for rounds 12-14, block 4 for rounds 15-17, and block 5 for rounds 18-20.

Table 5.6 Regressions of Within-Subject Standard Deviation Time Trends

Dep Var: Within Subject Std Dev	Pooled	Single Cue	Dual Cue
	Model 1	Model 2	Model 3
PR	(base)	(base)	(base)
PRWL	-2.621 (2.430)	-1.398 (2.338)	-1.853 (3.851)
T	4.908 * (2.796)	2.148 (2.221)	7.929 * (4.533)
PR*Block	0.535 (0.536)	0.267 (0.509)	0.773 (0.911)
PRWL*Block	0.572 (0.578)	0.627 (0.808)	0.501 (0.829)
T*Block	-1.211 ** (0.491)	-0.566 (0.367)	-1.838 ** (0.893)
Trait Anxiety	0.084 (0.104)	-0.065 (0.101)	0.098 (0.160)
Female	5.296 *** (1.416)	1.352 (1.452)	6.895 *** (1.985)
Constant	7.447 (4.682)	10.30 ** (4.570)	10.407 (7.453)
Observations	1060	540	520
Participants	212	108	104
R^2	0.034	0.009	0.042
Wald χ^2	23.24	4.14	18.64
$p > \chi^2$	0.002	0.764	0.009
PR*Block = PRWL*Block	$\chi^2(1) = 0.00$ p = 0.963	$\chi^2(1) = 0.14$ p = 0.706	$\chi^2(1) = 0.05$ p = 0.825
PRWL*Block = T*Block	**$\chi^2(1) = 5.53$** **p = 0.019**	$\chi^2(1) = 1.81$ p = 0.179	**$\chi^2(1) = 3.68$** **p = 0.055**
PR*Block = T*Block	**$\chi^2(1) = 5.77$** **p = 0.016**	$\chi^2(1) = 1.76$ p = 0.185	**$\chi^2(1) =4.19$** **p = 0.041**

Block denotes the time trend. Regressions are estimated with a Random Effects GLS procedure. Standard errors are in parentheses and are clustered by participants. Wald chi-squared tests of trends across treatments are presented at the bottom of the table, with bold typeface indicating statistical significance at the 10% level or better.

5.3. What Drives Tournament Learning?

Why is learning more pronounced in the T treatment than in other treatments? The answer lies in the rank-dependent reward structure, with the winner earning $1 while loser earns nothing. The Wald tests in Table 5.1 support this. In comparing the T and S trends, we are looking at how the trend differs when people are paid according to relative performance to when they are not paid for performance at all. If incentives are important and motivation crowding out plays a minor role, then we would expect better learning under tournaments. There is evidence in models 1 and 5 to show that the T trend is negative and steeper than the S treatment, showing significantly better learning. This is the first indication that it is the incentives in tournaments that foster learning.

The source of learning can be narrowed down further, by comparing PR and T treatments: both incentivise performance, but according to absolute and relative performance respectively. In models 2 and 4 of Table 5.1, there is evidence that the T treatment exhibits better learning than the PR treatment. This shows that learning comes about from relative incentives rather than from incentives based on absolute performance.

More specifically, we can compare the PRWL and T treatments. Both feature relative feedback but differ by pay scheme: piece rates versus the rank dependent payoffs. Here we distinguish between tournament feedback and payoffs. Again, the Wald tests show that the T treatment learns better than the PRWL treatment, though this time in models 1 and 5. This suggests that the rank-dependent payoffs inherent in tournaments is the source of the learning that we observe.

While the rank-dependent payoffs inherent in tournaments motivate learning in the T treatment, it is interesting to note that relative feedback has no impact on learning. Comparing the time trends of the PR and PRWL treatments, where pay is the same but the PRWL treatment includes feedback on relative performance, we see that the learning is identical in both treatments. This suggests that the winning/losing feedback itself does not contribute to learning in the T treatment. Similar inter-treatment comparisons of forecast consistency presented in Table 5.6 confirm this point.

The learning in the T treatment is attributable to the rank-order payoffs inherent in tournaments. The winner of the tournament receives a high prize while the loser receives a low prize. In our experiment the prize structure accentuates this effect, with players receiving $1.00 for winning or $0.00 for losing in the tournament round. Players improve their forecast errors in order to improve their chances of receiving the winning prize and/or reduce their chances of receiving the losing prize. While our experiments do not distinguish between the objectives of striving to win or avoiding the loss, a recent paper by Dutcher et al. (2015) show that the avoid-being-last objective has a greater effect than the strive-to-be-first objective in terms of eliciting effort. They find that when both objectives are present, effort is higher than each of the two separately. We believe the prize structure of our tournament is most reminiscent to their combined case.

We summarise our findings with respect to learning below:

Result 5.1. *Learning occurs only under tournaments. Both forecast accuracy and consistency improves over time. This learning is attributable to the rank-dependent payoffs inherent in tournaments.*

5.4. Tournament-No-Info Treatment

The proposition that the rank incentives of tournaments promote learning can independently be verified by comparing the Tournament-No-Info (TNI) treatment with the Tournament (T) treatment. At the end of the five piece rate rounds, the TNI treatment randomly matches a participant with another, and the winner with the smallest forecast errors receive the winning prize of $1 while the loser receives nothing. It features the same tournament incentives as the T treatment. The only difference is that in the TNI treatment, feedback pertaining to winning/losing – as well as round earnings, since it depends on winning and losing – is not shown inbetween rounds. This means that TNI participants do not know whether they have won or lost in the previous round. Since the T and TNI treatments feature the same rank-dependent incentives but differ in terms of relative feedback provision, the design is parallel to that of the PR and PRWL treatments. Here the T and TNI treatment comparison provides us an independent test of how relative feedback and incentives affect learning.

A priori, if we observe no differences in the rate of learning across these treatments, then we can definitively rule out the effect relative feedback has on learning. Furthermore – since we have already ascertained that there is learning in the T treatment – if we find that the TNI treatment has an identical rate of learning, then we can affirm that the learning is driven by the rank-payoffs.

We present the estimated time trends for the dual cue T and TNI treatments in Table 5.7.[37] Models 1 and 2 estimate a random effects specification with the TNI treatment dummy (with the T treatment serving as the reference category), trait anxiety, gender and linear time trends for both treatments. Regression models 3 and 4 run fixed effects regressions of the time trends, while purging the time-invariant variables. The regressions in models 1 and 3 are run over all post-intervention rounds, 6 to 20, while models 2 and 4 repeat the same over rounds 11 to 20.

In model 1, across rounds 6 to 20, the linear trend for the T treatment has a slope of -0.494 and is significant with a p-value of 0.02. This is identical in magnitude to the dual cue T trend that was estimated in model 5 of Table 5.1. In the TNI treatment, we observe a significant downward trend of -0.447, with a p-value of 0.043. This is not statistically different to that for the T treatment. It means that learning occurs even in the absence of relative performance feedback, lending more weight to the notion that the tournament incentives drive learning.

Unlike in previous regressions where we argued that comparing time trends over rounds 6 to 20 is potentially unfair due to interventions kicking in in some treatments but not others, we believe it is not inappropriate to study the T and TNI treatments over these rounds, since both treatments face similar interventions from round 6. Nevertheless the trends over rounds 11 to 20 are also presented. In model 2, although the trends for both treatments retain their negative sign, the T*Round coefficient is now insignificant. This is in line with model 6 of Table 5.1, where the T treatment did not have a significant trend line. On the other hand, the trend line for the TNI treatment of -1.226 is statistically significant. This suggests that the provision of relative feedback in the T treatment actually impedes learning, given that learning is significant in the TNI treatment but not the T treatment. Despite this, a Wald test cannot reject the null

[37] We find similar results for analogous regressions for the single cue task. However we do not report these since we have only one session of the single cue TNI treatment, limiting the number of observations. Since results are qualitatively similar, we will focus on the dual cue T and TNI treatments.

hypothesis that the estimated trend lines for the T and TNI treatment have the same slope (p = 0.153).

Table 5.7 Time Trends in Dual Cue Tournament and Tournament-No-Info Treatments

Dep Var: Forecast Errors	Random Effects		Fixed Effects	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4
T	(base)	(base)		
TNI	-4.659 (5.959)	10.78 (12.22)		
Trait Anxiety	0.412 (0.287)	0.379 (0.333)		
Female	9.070 ** (4.420)	10.27 ** (4.604)		
T*Round	-0.494 ** (0.213)	-0.347 (0.295)	-0.481 ** (0.201)	-0.333 (0.279)
TNI*Round	-0.447 ** (0.220)	-1.226 ** (0.540)	-0.495 ** (0.205)	-1.311 ** (0.506)
Constant	13.57 (13.38)	11.85 (16.26)		
Observations	1065	710	1140	760
Participants	71	71	76	76
R^2	0.027	0.028	0.004	0.002
Wald χ^2	16.06	10.08		
$p > \chi^2$	0.007	0.073		
F			5.78	4.07
p > F			0.005	0.021

T*Round = TNI*Round	$\chi^2(1) = 0.02$
	p = 0.877
		$\chi^2(1) = 2.04$
		p = 0.153
			F = 0.00
			p = 0.959
				F = 2.86
				p = 0.095

Regressions are run with observations from the dual cue T and TNI treatments. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. A hypothesis test is presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

The fixed effects regressions in models 3 and 4 show similar results. The rate of learning across both treatments is identical across rounds 6 to 20. When the trend lines are drawn over rounds 11 to 20, the T trend is negative but insignificant, while the TNI trend indicates significant improvements to forecast errors round by round. The pairwise F-test shows the trend is significantly different at the margin, with a p-value of 0.095. The corresponding pairwise test in model 2 did not show significance (p = 0.153).

To reiterate, for the T and TNI treatments we see no difference in learning between these treatments across rounds 6 to 20. Since the only difference between these treatments is in the provision of relative feedback, we can conclude that it has no effect on learning over these rounds. Across rounds 11 to 20, the regressions provide ambiguous results. There is marginal evidence to suggest that learning is actually better in the TNI treatment in the absence of relative feedback than in the T treatment with feedback provided. Since the evidence is marginal and differs from the unambiguous findings across rounds 6 to 20, and also since the data here is limited to only the dual cue task, we will err on the side of conservatism to adopt the interpretation of this evidence that there is no learning difference between these treatments across the latter rounds.[38]

Having found that forecast errors in the T and TNI treatments improve at indistinguishable rates, we proceed by comparing changes in forecast consistency across time in these two treatments. We again look at the within-subject standard deviation of forecast errors over three-round blocks over rounds 6 to 20. Figure 5.2 shows the within-subject standard deviations averaged across the T and TNI treatments for each three-round block. Like before, higher forecast consistency is represented by smaller within-subject standard deviations.

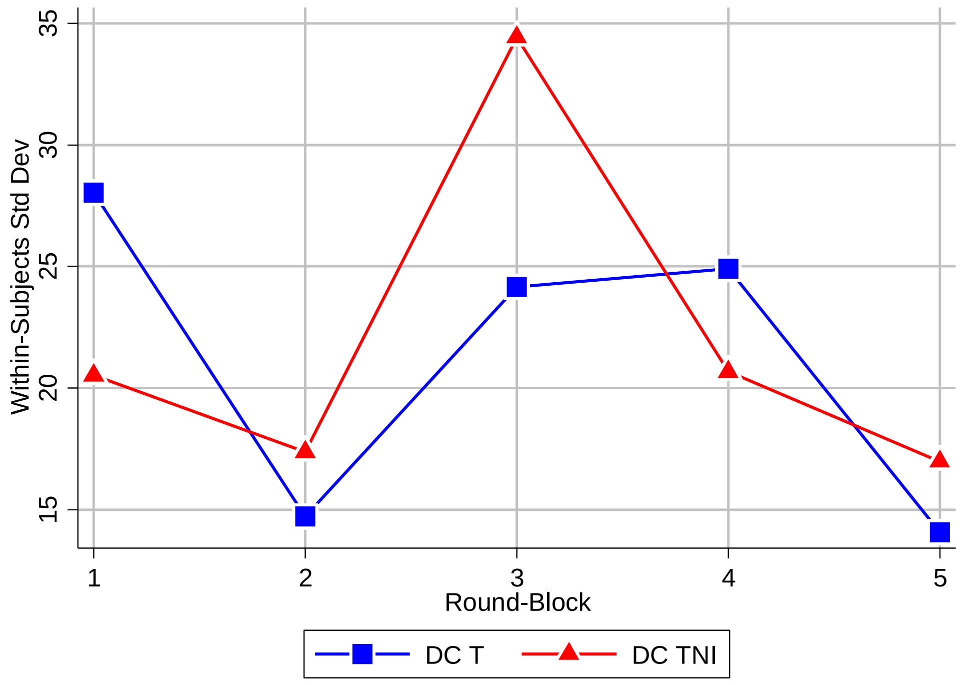
We see that forecast errors are more consistent in the TNI treatment than the T treatment in the first block of post-intervention rounds, rounds 6 to 8. In the last block, rounds 18 to 20, forecast errors in both treatments are lower than their levels in the first block. We observe that, compared to the first block, forecast consistency is much lower in the T treatment while is only

[38] If we adopted the alternative interpretation of the marginal evidence, that the TNI treatment learns better than the T treatment, we are suggesting that relative feedback impedes learning. Since there are close parallels to Merlo and Schotter (1999), it could be explained by their explanation of myopia. The provision of relative feedback in the T treatment diverts participants' attention away from the absolute feedback. Relative feedback does not assist one to learn of the underlying relationship per se, while absolute feedback can be useful via reinforcement. To shift attention towards relative feedback induces a myopic view of learning.

slightly so in the TNI treatment. Consequently, the downward trend in consistency is more apparent in the T treatment. There is a higher degree of fluctuation in forecast consistency in the TNI treatment, with a sharp peak in the third block, then sharply dropping over the fourth and fifth blocks.

We formalise analyses with regression analysis. Table 5.8 presents regressions of the within-subjects standard deviation of forecast errors for T and TNI treatments. The regressors include the TNI treatment dummy, trait anxiety scores and gender of each participant, as well as a linear trend for T and TNI treatments, denoted as T*Block and TNI*Block respectively. The first regression model is run across all round blocks, over rounds 6 to 20. We see that while the intercept is slightly lower in the TNI treatment compared to the T treatment, although not significantly so, consistency improves at a much faster rate in the T treatment. The linear trend for the T treatment has a slope of -1.838 and is significant at the 5% level. On the other hand, the TNI trend is negative but mild and insignificant. A Wald test comparing the trends, however, do not reject the null that the trends are identical ($\chi^2(1)$ = 2.16, p = 0.142).

Figure 5.2 Within-Subject Standard Deviation of Forecast Errors in T and TNI Treatments



Regression model 2 in Table 5.8 draws the trend lines from the second block. In other words, we are dropping 71 observations pertaining to the first block and estimating the trends starting from a different base. This allows us to test the sensitivity of these trends to their starting points, noting that consistency was initially much lower in the T treatment than the TNI treatment. In model 2, we again see that consistency improves in both T and TNI treatments over time, but now at a faster rate in the TNI treatment. Despite these negative trends, neither are statistically significant at conventional levels. Wald tests again do not suggest any differences in these trends ($\chi^2(1) = 0.60$, p = 0.437).

Finally, the third regression model in Table 5.8 replicates the previous regressions starting from the third block of rounds. While we have lost 40% of observations compared to model 1, drawing trends over three periods instead of five, model 3 is particularly interesting because it begins with the distinct peak in the TNI treatment – shown again in Figure 5.2. Accordingly, we observe the obvious downward sloping trend in the TNI treatment, with significance at a level better than 1%. Interestingly we also see significant improvements in consistency for the T treatment, also highly significant. As in the previous two regression models, hypothesis tests do not indicate any differences in the estimated trends for T and TNI treatments ($\chi^2(1) = 0.73$, p = 0.394).

The estimated trends of forecast consistency in Table 5.8 seem to be highly sensitive to the starting point. It is likely to be attributable to the small number of temporal periods for which the trends are drawn. Nevertheless two recurring observations can be made out from these regressions. The first is that across each regression model, both T and TNI treatments show a downward trend, whether it is significant or not. The second observation is that these downward sloping trends are no different across T and TNI treatments in each regression model, even though they vary considerably in magnitude in some cases. Based on these two overarching observations, we conclude that forecasts become increasingly consistent in both T and TNI treatments, and that there are no differences in this rate.

We have found evidence that there is leaning in both T and TNI treatments, where both the accuracy of forecasts and its consistency improves over time. More interesting is that the rates of learning are indistinguishable across these two treatments. Since the design differences between these treatments lie in the suppressed winning/losing feedback in the TNI treatment, the finding that both have similar rates of learning suggest that this feedback does not drive learning in the T treatment. Rather, since learning occurs in the TNI treatment when this feedback is absent, we can pinpoint learning down to the common rank-based incentives that feature in both treatments.

The findings here from the T and TNI treatments reinforce our previous explanation that it is the rank-based incentives that drives learning in tournaments. The extremity of the winner-take-all rewards provides the impetus for players to perform since the differences in reward between winning and losing are stark. The rank-payoffs reward the winner well, while the zero losing prize acts as a deterrence to losing. As a result, those in the T and TNI treatments have the motivation to improve their forecast errors, irrespective of whether they know whether they have won or lost in prior play.

Table 5.8 Regressions of Within-Standard Deviation Time Trends in T and TNI Treatments

Dep Var: Within-Subject Std Dev	Blocks 1-5	Blocks 2-5	Blocks 3-5
	Rounds 6-20	Rounds 9-20	Rounds 12-20
	Model 1	Model 2	Model 3
T	(base)	(base)	(base)
TNI	-4.858 (4.608)	6.014 (7.585)	13.541 (16.112)
Trait Anxiety	0.287 (0.221)	0.255 (0.261)	0.299 (0.290)
Female	4.561 (3.524)	5.019 (4.088)	5.955 (4.371)
T*Block	-1.838 ** (0.896)	-0.199 (0.871)	-5.289 *** (1.703)
TNI*Block	-0.179 (0.689)	-1.291 (1.103)	-8.134 *** (2.866)
Constant	11.72 (10.76)	6.211 (12.01)	25.70 * (15.22)
Observations	355	284	213
Participants	71	71	71
R^2	0.027	0.024	0.084
Wald χ^2	7.61	3.20	19.32
$p > \chi^2$	0.179	0.668	0.002
T*Block = TNI*Block	$\chi^2(1) = 2.16$ p = 0.142	$\chi^2(1) = 0.60$ p = 0.437	$\chi^2(1) = 0.73$ p = 0.394

Regressions are run with observations from the dual cue T and TNI treatments. Regressions are estimated with Random Effects GLS. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald chi-squared tests are presented at the bottom of the table.

5.5. Reinforcement Learning

We have previously shown that there is significant learning in the T treatment, and this is attributable to the rank-based payoffs that feature in tournaments. These findings have come about by looking at linear time trends and comparing them across treatments. An alternative approach to analyse learning is to look at how people respond to feedback from previous rounds. We now begin to look deeper into the processes associated with learning.

In a cognitively challenging task such as ours, we expect that feedback plays an important role in facilitating learning. In this section, we focus on absolute feedback on forecast errors and how it is utilised to improve forecasts in subsequent rounds. In the following section we will look at the role relative feedback has on learning.

In studying how absolute feedback affects performance, we look at how participants process the forecast error feedback that they receive and whether this improves forecasts in latter rounds. This is the essence behind reinforcement learning. Reinforcement learning is founded upon the premise that if a decision receives 'good' feedback, then the same decision will be more likely played again in the future. Feedback therefore reinforces the decision. Likewise, bad decisions will less likely be played again.

Theoretical models of reinforcement learning are presented in Roth and Erev (1995) and Erev and Roth (1998).[39] They model monetary payoffs as the reinforcing event, where higher payoffs in the previous period will result in a higher propensity to replay the decision which had led to the increased earnings in the first place. In these models, the choice of payoffs as the reinforcing feedback is appropriate in the context of strategic games, where payoffs serve as the sole measure of whether a strategy played is good or bad. Payoffs are less appropriate in the forecasting task that we use, especially when payoffs vary with different pay schemes. Instead we will consider forecast errors as the reinforcing feedback, whereby forecast errors measure performance.

In the early stages of the game, when people begin to make forecasts by trial and error, the forecast error feedback that they receive at the end of the round is very valuable. Since forecast

[39] For related work, see Feltovich (2000), Erev, Bereby-Meyer, and Roth (1999) and Bereby-Meyer and Roth (2006).

errors reflect how accurate forecasts are, they also indicate how good the 'trial' is, and in turn the performance of the forecast rule that is used. If forecast errors are low, then people may choose to stick with the same forecast rule for future rounds, or may only make minor refinements. On the other hand, if forecast errors are high, then there is reason for people to trial another forecasting rule, with the process continuing until they find a satisfactory rule.[40] Throughout this process of reinforcement, we would expect favourable feedback to improve future outcomes. In the context of the forecasting game, lower forecast errors in the past would be expected to improve future forecasts.

Since we do not actually observe the forecasting rules that participants formulate to base their forecasts upon, we are not able to directly apply the reinforcement learning models of Roth and Erev to the forecasting task. While we are not able to fully exploit the theory, we can nevertheless test for signs of reinforcement learning in each of the treatments. In particular, we can see which treatment makes the most of prior forecast error feedback, and how it translates to improvements in forecasts. Given that we have previously found that the T treatment exhibits significant learning over time, we would also expect participants in the T treatment to show better reinforcement learning.

According to reinforcement theory, greatest weight is placed on the most recent piece of feedback. We begin by looking at how the forecast errors in round $t - 1$ affect forecast errors in round t in each treatment. Panel A of Table 5.9 presents a regression of forecast errors against its first lag, interacted by treatment.[41] Other regressors include the treatment dummies, trait anxiety and gender. The regression is run with pooled, single and dual cue data over rounds 6 to 20. A series of Wald chi-squared tests that relate to the lagged forecast errors across treatments are presented at the bottom of the table.

[40] There is an inherent trade-off between 'exploring' and 'exploiting', to use the words of Merlo and Schotter (1999). If a participant is satisfied that the forecasting rule is good enough, then they would stop searching for a better rule and start to exploit the rule which they have formulated.

[41] Levin, Lin, and Chu (2002) panel unit root tests run for each single and dual cue treatment rejects the presence of a unit root in the data (p = 0.000 in all tests). This suggests stationarity of the panels.

Table 5.9 Reinforcement Learning: Effect of First Forecast Error Lag

Dep Var: Forecast Error	Pooled	Single Cue	Dual Cue
	Model 1	Model 2	Model 3
PR	(base)	(base)	(base)
PRWL	-0.694 (2.052)	0.385 (1.918)	0.573 (3.013)
T	-0.213 (2.202)	-0.739 (1.952)	1.136 (3.463)
Trait Anxiety	0.112 (0.103)	-0.095 (0.094)	0.187 (0.186)
Female	6.045 *** (1.333)	1.056 (1.198)	9.528 *** (2.053)
L1 PR*Error	0.293 *** (0.062)	0.164 *** (0.059)	0.238 *** (0.071)
L1 PRWL*Error	0.150 *** (0.053)	0.046 (0.035)	0.063 (0.059)
L1 T*Error	0.319 *** (0.065)	0.214 ** (0.090)	0.236 *** (0.080)
Constant	6.076 (4.535)	12.25 *** (4.098)	7.901 ** (8.096)
Observations	3180	1620	1560
Participants	212	108	104
R^2	0.111	0.028	0.090
Wald χ^2	85.10	18.56	46.28
$p > \chi^2$	0.000	0.010	0.000
L1 PR*Error = L1 PRWL*Error	**$\chi^2(1) = 3.07$ p = 0.080**	**$\chi^2(1) =2.95$ p = 0.086**	**$\chi^2(1) = 3.52$ p = 0.061**
L1 PRWL*Error = L1 T*Error	**$\chi^2(1) = 4.09$ p = 0.043**	**$\chi^2(1) = 2.98$ p = 0.084**	**$\chi^2(1) = 2.99$ p = 0.084**
L1 PR*Error = L1 T*Error	$\chi^2(1) = 0.08$ p = 0.772	$\chi^2(1) = 0.21$ p = 0.649	$\chi^2(1) = 0.00$ p = 0.983

Regressions are estimated with Random Effects GLS. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

We focus our attention on the coefficients of the lagged forecast error for each treatment. The first observation from the regressions is that – significant or not – every coefficient on the lagged forecast errors, across treatments and across different regression models, has a positive sign. The direction of the estimated coefficients is consistent with reinforcement learning. Higher forecast errors in the previous round are associated with higher forecast errors in the present round. Conversely, feedback about better forecasts in the previous round drive further improvements in forecasts in the present round.

There is evidence to show that reinforcement learning occurs in every treatment, though to varying degrees. In model 1, the pooled regression, the lagged forecast error terms are significantly different to zero in all treatments. The coefficient for the PRWL treatment is smaller in magnitude than in the PR and T treatments; these differences are statistically significant at conventional levels. In fact when the regressions are replicated with the single and dual cue tasks separately, the lagged forecast error for the PRWL treatment is no longer significant in either regression. The coefficients on the PR and T lagged error terms continue to be strongly significant.

In terms of the degree of learning, both the PR and T treatments show greater reinforcement than the PRWL treatment, since they have larger coefficients. The Wald chi-squared tests at the bottom of Table 5.9 show that while there are no differences in reinforcement learning between the PR and T treatments, the learning in these treatments is significantly better than the PRWL treatment in all three regression models.

Given that the PR and PRWL treatments differ only in terms of the relative feedback provided in the PRWL treatment, the difference in reinforcement learning that we observe between these treatments is attributable to the relative feedback. We see that the process of reinforcement is slower in the PRWL treatment compared to the PR treatment. This suggests that the provision of relative feedback impedes reinforcement learning. A possible explanation is that the provision of relative feedback distracts the participant from focusing on historical forecast errors. Since relative feedback relates to benchmarking, it provides no reinforcement value per se.[42] An increased focus on relative feedback reduces the focus on previous forecast errors, reducing the effectiveness of reinforcement learning. We will revisit this point in the following section.

We also observe that there is greater reinforcement in the T treatment than in the PRWL treatment. Both feature relative feedback, but incentives differ. Although relative feedback crowds out reinforcement learning, the rank-based feedback seem to shift emphasis back to forecast errors. Since there is more to gain (lose) from performing well (poorly), there are greater incentives in place to encourage T participants to perform. With greater incentives to perform, they will make better use of the feedback that will assist them to do so.

We further investigate reinforcement learning by looking at the effect of deeper forecast error lags. As time progresses, the stock of feedback accumulates. If the PR and T treatments exhibit better reinforcement than other treatments, we expect to see them utilise more information that they have at their disposal to assist them with their forecasts. Table 5.10 shows how five forecast error lags affect present round forecasts in different treatments. These regressions pool single and dual cue treatments in order to increase the sample size we have to work with. The different regression models represent each of the PR, PRWL and T treatments.

The random effects regressions that are run in Table 5.10 are done so in two stages for each treatment. In the first stage, forecast errors are regressed only against the controls of trait anxiety and gender. Residuals from the first stage regression are calculated, then put through a second stage regression against five forecast error lags. [43] The first lag is labelled as L1, the second lag L2 and so on. The reason why regressions are run in two stages is so that in the second stage regressions, the coefficient of determination (R^2) measures the fit of the regression models that is attributable solely to the five lagged forecast error values, excluding the contribution by the first stage variables. This way, the R^2 is an aggregated measure of how well these five lags jointly

[42] Although relative feedback provides no information over and above forecast error feedback that is useful for reinforcement, it may spur learning through other channels. We cover this point in more detail in the following section.

[43] The choice of the number of lags is an arbitrary decision. Since we want to look at how much prior information was utilised in the learning process, we want to include as many lags as feasible. However, the trade-off is a reduced number of observations available for analyses. We believe five lags forms a nice balance.

influence the choice of forecast errors. The coefficients on these lagged forecast error terms are quantitatively similar when the lags, trait anxiety and gender are included in a standard single-stage regression (results suppressed).

Table 5.10 Reinforcement Learning: Effect of Five Forecast Error Lags

Dep Var: Forecast Error	PR	PRWL	T
	Model 1	Model 2	Model 3
First Stage Regression			
Trait Anxiety	0.435 (0.347)	-0.037 (0.180)	0.115 (0.252)
Female	8.706 ** (3.940)	6.465 *** (2.263)	10.31 *** (3.783)
Constant	-2.336 (13.846)	13.80 * (7.426)	10.10 (10.74)
R^2	0.034	0.020	0.029
Wald χ^2	5.04	8.27	7.44
$p > \chi^2$	0.080	0.016	0.024
Second Stage Regression			
L1 Forecast Error	0.129 (0.099)	0.072 * (0.039)	0.157 *** (0.050)
L2 Forecast Error	0.037 (0.037)	0.038 (0.045)	0.045 (0.028)
L3 Forecast Error	0.157 *** (0.049)	0.090 *** (0.027)	0.132 *** (0.030)
L4 Forecast Error	0.121 ** (0.054)	0.212 *** (0.030)	0.197 *** (0.036)
L5 Forecast Error	0.125 *** (0.047)	0.073 ** (0.030)	0.117 *** (0.031)
Constant	-12.19 *** (1.271)	-8.186 *** (1.267)	-14.71 *** (1.357)
Observations	1020	1095	1065
Participants	68	73	71
R^2	0.186	0.102	0.207
Wald χ^2	503.7	95.51	119.9
$p > \chi^2$	0.000	0.000	0.000

The first stage regression regresses forecast errors against trait anxiety and gender. Residuals are calculated and run in the second stage regression against five forecast error lags. Both regression stages are estimated with Random Effects GLS across rounds 6 to 20. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively.

We focus on the second stage regression output which shows the coefficients on the five lags of forecast errors. The first thing to point out is that the coefficient on the first lag is substantially smaller than those presented earlier in Table 5.9. Given there is a positive correlation between forecast errors and their lagged values, the coefficients reported in Table 5.9 have incorporated the effect of these deeper lags. Of the first period lags, the one for the T treatment takes on the largest value as before. We also see that the first forecast error lag in the PR treatment is still large, but is now insignificant.

It is interesting to see that the second forecast error lag is insignificant in all of the treatments at conventional levels, though is only marginally so in the T treatment (p = 0.112), especially when deeper lags have a significant effect on forecasts. The third and fourth lagged errors influence current round forecasts in all treatments. The fifth lag has a smaller effect, though is still significant in all treatments.

We use the R^2 in the second stage regression as an aggregated measure of how effective participants in each treatment utilise the stock of feedback at their disposal to assist them with the task. The second stage R^2 is highest in the T treatment, with a value of 0.207. This is followed by the fit of 0.186 for the PR treatment. These R^2 coefficients for the PR and T treatments are much higher than that for the PRWL treatment, at 0.102. This tells us that reinforcement learning is more effective in the PR and T treatments. Result 5.2 summarises this:

Result 5.2. *Reinforcement learning occurs under tournaments. Players under tournaments are more likely to utilise past performance feedback to improve future performance than in other treatments.*

The fact that the T treatment shows better reinforcement learning than other treatments corroborates with various strands of evidence that relates to learning in the T treatment. Better reinforcement would lead to forecasts that are more accurate and consistent over time. Again this would be attributable to the nature of rank-dependent payoffs in tournaments as opposed to the relative feedback, given that the reinforcement learning is greater in the PR treatment than in the PRWL treatment when relative feedback is provided.

While greater reinforcement learning in the T treatment explains our previous results that the T treatment exhibits better learning, we note that the better reinforcement in the PR treatment does not seem to have much impact on learning. In the PR treatment, the trends indicate that forecast errors decrease – although it is not statistically significant in most regression models. Despite this, there is no evidence that forecasts become more consistent in the PR treatment.

5.6. Effect of Winning

Most of the previous emphasis in this chapter has been on whether relative feedback has any effect on performance. More specifically by comparing the PR and PRWL treatments, we are looking at the effect that the *provision* of relative feedback has on performance and learning. In this section we pose a closely related research question: what effect does the *content* of relative feedback have on performance? In other words, how does the receipt of winning feedback influence future performance relative to those who receive losing feedback?

This is similar to the distinction made by Blanes i Vidal and Nossol (2011) of anticipation and revelation effects of feedback. In the period when the provision of feedback had been announced, but before the feedback revealed, would people increase their performance in order to be ranked favourably and accordingly receive favourable feedback? This is the ex ante anticipation effect. The ex post revelation effect relates to how participants respond to past feedback. In this section we will focus on the revelation effect to see whether winning improves future performance. Answering this requires us to restrict ourselves to the PRWL and T treatments, since they are the only treatments that provide feedback on relative performance.

One avenue which relative feedback may motivate performance is through the competency evaluation aspect of Cognitive Evaluation Theory (Deci & Ryan, 1985). People are intrinsically motivated to perform when they receive favourable feedback which suggests that they are adept, while their intrinsic motivation falls when the feedback is unfavourable. The binary nature of the relative feedback that we provide facilitates studying the effect of this feedback, since there is no ambiguity in assessing whether the feedback is favourable or not. If this aspect of Cognitive Evaluation Theory is borne out in the data, we would expect participants who receive the context-loaded feedback of "win" to perform better in subsequent rounds relative to those who receive the feedback of "lose".

The random rematching of participants also facilitates studying the effects of winning versus losing. When participants play with different people in each round, we avoid situations where a partner dominates the other for the entire duration of play. In these cases, we may expect the losing partner to simply give up due to the unfavourable matchup. With random rematching, participants who have lost a round do not necessarily expect to consistently lose in successive rounds. Given this, we expect most participants to experience a mix of both wins and losses throughout the game, with few people at either extreme who wins or loses consistently.

Figure 5.3 shows the distribution of the proportion of games that participants have won over rounds 6 to 20, by treatment. The horizontal axes show the proportion of post-intervention rounds won, with a bin width of ten percentage points. The vertical axes show the proportion of participants in each treatment who have attained the particular win rate. The histograms show that in each treatment, the peak lies in the centre with 60-70% of participants winning 40-70% of their games. Since most people experience both wins and losses, the effect of winning over losing – if the effect exists – should be more reliable under random rematching than we would expect under a fixed matching protocol.

There are two primary links that need to be distinguished. The first is how winning or losing affects competency and the effort to perform. This link is well supported by the psychology literature on intrinsic motivation. The second is how this affects performance.

To study the first link, we employ the competency variable which was elicited from the post-experiment questionnaire. See Section 3.4 for details. A higher competency score represents higher self-reported competency. We look at the correlation of the self-reported competency score with the proportion of the 15 post-intervention rounds for which each participant has won. Table 5.11 presents an ordinary least squares regression of competency against treatment dummies, and the percentage of rounds won interacted by treatment. The second regression model also includes trait anxiety and gender as controls.

In Table 5.11 we observe self-reported competency scores to be significantly lower in the T treatment than in the PRWL treatment. This is true for both single and dual cue treatments. These differences widen as we introduce controls in model 2. There are no gender differences in how competent participants feel. Higher trait anxiety suppresses feelings of competency.

The proportion of games won has a significant impact on how people self-report their competency. As expected, the more games that a participant has won, the higher they will rate their own competency. Regression model 1 shows that in the single cue task, the average PRWL participant rates himself 0.131 points more competent if his winning record improves by a single percentage point, compared to 0.261 in the T treatment. However an F-test shows that this difference is not statistically significant with a p-value of 0.165; the significance changes to 0.018 when trait anxiety and gender are included as controls. The dual cue task also shows that the T treatment reports higher competency than the PRWL treatment given the same level of improvement, and the difference is significant (F(1,141) = 6.43; p = 0.012).

Figure 5.3 Histogram of Participants' Win Rates, by Treatment

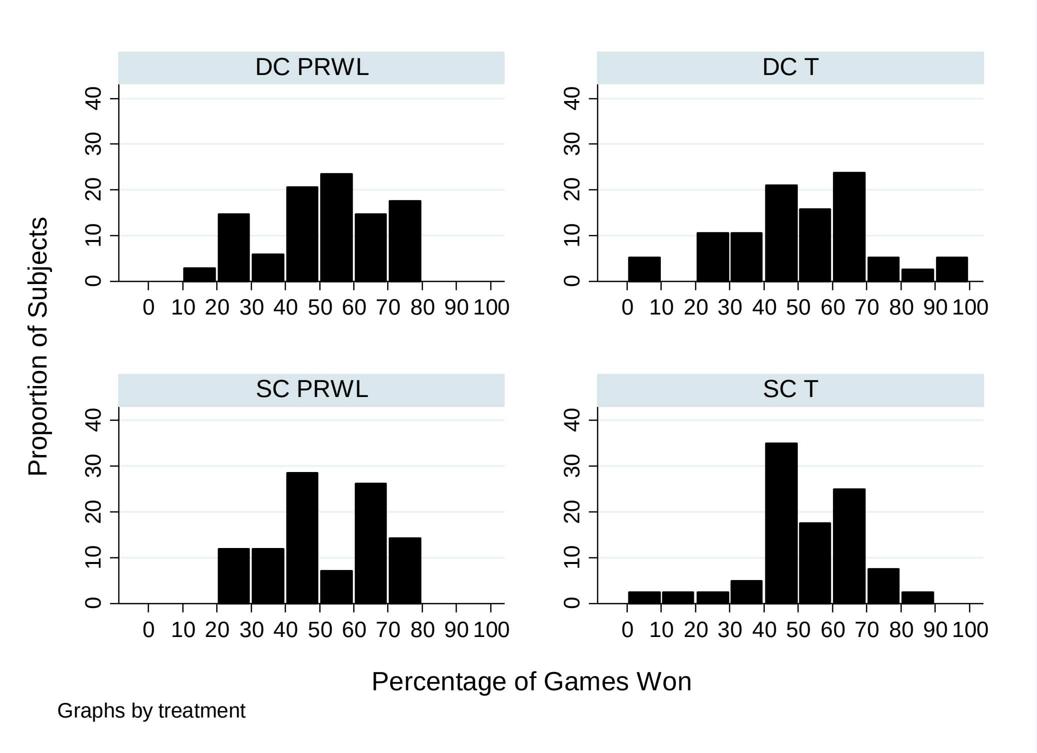


Table 5.11 Regression of Competency Scores and Win Record

Dep Var: Competency	Model 1	Model 2
SC PRWL	(base)	(base)
SC T	-8.995 * (5.223)	-13.44 *** (4.914)
DC PRWL	-2.253 (3.637)	-2.114 (3.764)
DC T	-12.25 *** (3.700)	-14.02 *** (3.896)
SC PRWL * Win Percent	0.131 ** (0.054)	0.112 ** (0.057)
SC T * Win Percent	0.261 *** (0.076)	0.325 *** (0.067)
DC PRWL * Win Percent	0.177 *** (0.043)	0.158 *** (0.046)
DC T * Win Percent	0.324 *** (0.040)	0.350 *** (0.046)
Trait Anxiety		-0.225 *** (0.072)
Female		-0.031 (1.038)
Constant	21.14 *** (3.091)	31.57 *** (4.494)
Observations/Participants	149	139
R^2	0.402	0.471
SC T = 0	**F = 2.97** **p = 0.087**	**F = 7.48** **p = 0.007**
DC PRWL = DC T	**F = 12.80** **p = 0.001**	**F = 16.33** **p = 0.000**

SC PRWL * Win Percent =
SC T * Win Percent
	F = 1.95
	p = 0.165
		F = 5.79
		p = 0.018

DC PRWL * Win Percent = DC T * Win Percent	**F = 6.43** **p = 0.012**	**F = 9.45** **p = 0.003**

Regressions are estimated with OLS. Robust standard errors are in parentheses. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. F tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Despite reporting lower levels of competency overall, the self-reported competency scores of T participants are much more responsive to their winning record than PRWL participants. The responsiveness of competency scores to the record of winning for the single and dual cue T treatments are almost twice the magnitude than for the corresponding PRWL treatments. This is best explained by the fact that in the T treatment, the proportion of wins can be interpreted as not only the number of wins, but also the number of instances for which participants have been paid the winning prize. In other words, the rank-dependent payoffs seem to have an additional effect on competency over and on top of the frequency of winning, increasing its saliency.

We have established the first link that winning improves players' self-reported competency scores. This means that the winning/losing feedback is not simply ignored by participants. When winners perceive themselves as being more competent than losers, how does the greater perception of competency affect forecast errors? It is not easy to answer this question directly. The only measure of competency that we have at our disposal was elicited at the conclusion of the final round. It is quite plausible that participants who have made better forecasts in the experiment report higher competency, for which we introduce simultaneity in our regressions if we allow competency to have an effect on forecast errors à la Cognitive Evaluation Theory. This endogeneity does not allow us to look at the second link directly, especially when we do not have instrumental variables at our disposal to disentangle the effects. Although we cannot study this second link directly, we can approximate this second link by studying how winning in a particular round affects forecast errors in the future.

We study the immediate motivating effect of winning versus losing by utilising a dummy variable denoted as "Win" which takes the value of 1 if the participant has won or the value of 0 if the participant has lost. It therefore looks at the marginal effect of winning over losing. This dummy identifies instances when someone has won at the observation level, rather than identify individual participants according to their ability. Since practically all participants have experienced both wins and losses, this allows us to get a direct and reliable indication of how winning affects performance relative to losing, irrespective of their ability.

Table 5.12 Effect of Winning on Forecast Errors

Dep Var:	Pooled	Single Cue	Dual Cue	Pooled	Single Cue	Dual Cue
Forecast Errors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Piece Rate Win Lose	(base)	(base)	(base)	(base)	(base)	(base)
Tournament	5.051 ** (2.475)	0.654 (1.446)	7.220 * (3.720)	5.003 * (2.669)	-0.563 (2.004)	9.795 ** (4.337)
Lagged Win	0.468 (0.989)	-1.881 * (0.987)	-0.096 (1.964)			
Lagged PRWL*Win				0.369 (1.346)	-2.983 ** (1.331)	2.259 (2.627)
Lagged T*Win				0.468 (1.462)	-0.555 (1.445)	-3.013 (3.029)
Trait Anxiety	0.022 (0.162)	-0.155 (0.136)	-0.029 (0.245)	0.022 (0.162)	-0.163 (0.137)	-0.015 (0.246)
Female	8.449 *** (2.224)	1.367 (1.611)	11.39 *** (3.293)	8.445 *** (2.223)	1.380 (1.619)	11.38 *** (3.258)
Constant	10.21 (6.873)	16.63 *** (5.458)	18.12 * (10.56)	10.27 (6.937)	17.50 *** (5.627)	16.39 (10.84)
Observations *	2003	1064	939	2003	1064	939
Participants	144	76	68	144	76	68
R^2	0.033	0.014	0.041	0.033	0.015	0.045
Wald χ^2	16.66	6.51	15.76	17.43	8.41	16.11
$p > \chi^2$	0.022	0.164	0.003	0.004	0.135	0.007
L.PRWL*Win = L.T*Win				$\chi^2(1) = 0.00$ $p = 0.960$	$\chi^2(1) = 1.51$ $p = 0.219$	$\chi^2(1) = 1.69$ $p = 0.193$

* Data pertaining to one dual cue PRWL participant has been removed. Due to random matching, individual observations for which players are matched with the removed participant are also removed for purposes of analysing the effect of winning. Regressions are estimated with Random Effects GLS. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Regressions are run with forecast errors against this lagged variable, allowing us to study the immediate motivating effect of winning over losing. Table 5.12 presents these regressions for the pooled, single and dual cue data. Models 1 to 3 are regressed with the lagged win variable, while models 4 to 6 split this effect by treatment.

In model 1 of Table 5.12 when single and dual cue treatments are pooled, there is no evidence to show that winning the previous round affects current round forecasts. The lagged win variable is again insignificant in the dual cue regression of model 3. However in the single cue task, we see that immediately following a win, forecast errors are estimated to improve by 1.881 points relative to losers, with a p-value of 0.057.

We look deeper into this by interacting the lagged win variable by treatment. In the analogous single cue regression in model 5, we see that this is driven by the PRWL treatment. In the single cue PRWL treatment, a previous round win is associated with an improvement of almost 3 forecast error points in the round that follows. A previous round win has no significant effect for the single cue T treatment. In the dual cue task, the lagged PRWL*Win and T*Win variables are have different signs, with a win improving performance in the T treatment while reducing performance in the PRWL treatment. However, both coefficients are not significantly different from zero, and a pairwise Wald test cannot reject the null that these effects are identical (p = 0.193). The aggregated results in the pooled regression of model 4 also shows that previous round wins have no effect on forecast errors.

Although there is some evidence to suggest that winning is associated with improved forecast errors in the single cue PRWL treatment, this does not appear to be a broader result. Winning has no effect on PRWL participants in the dual cue task, nor for participants in the T treatment. On the balance of the evidence, winning does not have an effect on forecast errors.

A possible explanation for this is that winning in one particular round does not necessarily equate to higher competency evaluation. In other words, one observation of winning might not be salient enough in itself to induce greater feelings of competency, and as a result has no effect on forecasts. This is especially true given that most participants experience a mix of both wins and losses.

To examine this possibility, we repeat the regressions in Table 5.12 but instead use a measure of each participant's entire history of winning as a regressor. We define a participant's win record to be a rolling measure of the percentage of prior post-intervention rounds for which they have won. For example, if in round 9 a participant has won two of the three prior rounds, then his win record is 66.67%. The win record is updated every round as the wins or losses accumulate.

If the participant loses in round 9, then his win record at the beginning of round 10 would reflect two wins and two losses, so his win record will fall to 50%. This rolling metric accounts for the entire history of winning and losing for each participant at each point in time. By comparison, the lagged win variable incorporates information about winning and losing only from the previous round. Table 5.13 shows the effect of this win record variable on forecast errors in a series of regressions.

Table 5.13 Effect of Win Record on Forecast Errors

Dep Var: Forecast Errors	Pooled	Single Cue	Dual Cue	Pooled	Single Cue	Dual Cue
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Piece Rate Win Lose	(base)	(base)	(base)	(base)	(base)	(base)
Tournament	5.010 ** (2.482)	0.669 (1.431)	7.185 ** (3.607)	3.617 (4.334)	-0.436 (2.922)	8.329 (7.201)
Win Record	0.009 (0.034)	-0.036 (0.023)	-0.039 (0.061)			
PRWL*Win Record				-0.005 (0.040)	-0.046 (0.031)	-0.029 (0.068)
T*Win Record				0.023 (0.055)	-0.023 (0.034)	-0.052 (0.098)
Trait Anxiety	0.017 (0.163)	-0.148 (0.138)	-0.042 (0.237)	0.012 (0.165)	-0.153 (0.139)	-0.039 (0.241)
Female	8.520 *** (2.208)	1.066 (1.627)	11.07 *** (3.116)	8.563 *** (2.190)	1.128 (1.633)	11.03 *** (3.067)
Constant	10.21 (6.833)	17.28 *** (5.337)	20.81 ** (10.27)	11.12 (7.289)	17.97 *** (5.711)	20.21 * (11.59)
Observations *	2015	1064	951	2015	1064	951
Participants	144	76	68	144	76	68
R^2	0.032	0.016	0.050	0.031	0.016	0.051
Wald χ^2	16.76	5.40	16.96	18.36	5.40	18.50
$p > \chi^2$	0.002	0.248	0.002	0.003	0.370	0.002
PRWL*Win Rec = T*Win Rec				$\chi^2 = 0.17$ p = 0.678	$\chi^2 = 0.24$ p = 0.627	$\chi^2 = 0.04$ p = 0.847

* Data pertaining to one dual cue PRWL participant has been removed. Due to random matching, individual observations for which players are matched with the removed participant are also removed for purposes of analysing the effect of winning. Regressions are estimated with Random Effects GLS. Standard errors are in parentheses and are clustered by participants. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Regression models 1 to 3 for the pooled, single and dual cue series shows that the rolling win record does not have any effect whatsoever on forecast errors. This is also true in regression models 4 to 6 where this variable is interacted by treatment. The history of winning does not affect forecast errors for both PRWL and T treatments. The evidence corroborates with the previous finding with the lagged win variable, that by and large, winning has no effect on forecast errors. In other words, participants in both PRWL and T treatments are forward-looking and do not worry about what has happened in the past. We state this finding as Result 5.3:

Result 5.3. *Relative performance feedback has no influence on learning. A players' knowledge of whether they have won or lost does not impact their future performance.*

Comparing the PRWL treatment to the T treatment in models 4 to 6 of Tables 5.12 and 5.13, the historical effects of both winning in the prior round and of the rolling win record are no different across treatments. When interpreting this, we should bear in mind that the difference between the treatments lie in the payoffs – piece rates versus the rank-dependent payoffs – while are matched in terms of the feedback provided to participants. Following from the different payoffs, the implications of winning are different across these treatments. In the PRWL treatment, winning has no impact on payoffs per se. On the other hand, winning in the T treatment is associated with receiving the higher monetary prize. In this sense, the winning in the T treatment can also be interpreted as an instance for which a participant receives the higher prize rather than the smaller one.

Bearing this in mind, the interpretation of the effect across treatments can be any combination of two different explanations. The first interpretation is that people respond to past feedback in an identical manner under both piece rates and rank-dependent payoffs. This follows from the treatment differences. The second interpretation of the result is that the additional saliency of winning in the T treatment, the knowledge of receiving the winning prize, has absolutely no effect on future performance. Given the very nature of tournaments and that the prizes are tied to winning or losing, these two different effects cannot be disentangled.

It should be noted that while the winning/losing feedback has no effect on performance, it is not ignored altogether. The fact that participants in the PRWL and T treatments condition their self-reported competency on their win record suggests that winning makes people feel better about their ability, although this does not translate into better performance in the forecasting task.

Despite Cognitive Evaluation Theory suggesting that winning should motivate people to perform vis à vis losing, in a sense it is also unsurprising that it has no effect on performance. Unlike feedback on forecast errors which can be used to verify whether a particular forecasting rule is working or not, feedback on winning or losing does not assist with the process of learning. Relative feedback merely facilitates benchmarking. This is similar to the point made by Kluger and DeNisi (1996) in their meta-analysis of the effect of feedback, where they conclude that feedback only improves performance if it provides information about how to go about performing the task at hand.

The finding that historical feedback on winning and losing has no effect on performance lends support to a bigger picture. We have previously attributed the learning in the T treatment to the rank-dependent payoffs intrinsic to tournaments. Relative feedback does not appear to play any role in participants' learning. The null finding presented here that the event of winning does not affect future forecast performance adds another line of support to this claim.

5.7. Summary and Discussion

This chapter focused on the temporal dimension of learning across treatments. Of the treatments, the T treatment stands out in terms of their superior learning. Forecast accuracy improves at a significantly faster rate than any other treatment. Not only are forecasts becoming increasingly accurate in the T treatment, we find evidence that these forecast errors are becoming increasingly consistent too. Both the improved accuracy and consistency of forecasts in the T treatment constitute strong evidence for learning. This is exactly what we would expect when participants' own forecast rules converge to the underlying relationship of cues to the actual stock price. In no other treatment do we see consistent signs of learning. The learning in the T treatment is robust to various specifications and estimation methods.

Why does the T treatment learn significantly better than other treatments? The answer lies in the rank-dependent payoffs characteristic of tournaments. This is supported by two independent strands of evidence. The first line of support comes from pairwise treatment comparisons of the time trends. Comparing the learning of the T and S treatments, we see that the T treatment learns significantly better – this is attributable to the performance pay of tournaments. Refining this further, by comparing the T and PR treatments, we still see the T treatment to show better learning. Both treatments pay based on performance, but according to rank-payoffs and piece rates respectively. Given that the T treatment shows better learning, the tournament schemes motivate learning more than piece rates.

We narrow this down further by comparing the PRWL and T treatments. Both treatments provide identical feedback, but differ only by the pay scheme. Again the T treatment shows better learning, ruling out the effect relative feedback has on learning. Learning is no different between the PR and PRWL treatments, suggesting once again that relative feedback has no influence on learning patterns. Ruling out the effect relative feedback has on learning, the significant learning in tournaments is therefore attributable to its rank payoffs.

The second strand of evidence to support the claim that rank-dependent payoffs motivate learning in the T treatment is unequivocal. By comparing the T and TNI treatments we can directly infer the effect relative performance feedback has on learning. Both these treatments are based on tournaments, but the TNI treatment withholds any feedback that relates to winning or losing, including earnings feedback. We see that the rate of learning is identical in both treatments. This means that the relative feedback has no effect on learning, and in turn it is the rank payoffs that drive learning. The winner-take-all nature of the rank incentives means that people are much more motivated to win, for losing will yield no return.

6. Results: Ability of Players

The previous results have focused on overall treatment effects and learning across treatments, where analyses had mainly been conducted at the aggregate level. While the aggregate results provide a broad overview of the data, it may not necessarily be a faithful reflection of it when there is a large degree of heterogeneity in participants' ability and traits.

Different effects for people of different ability may exist, which would be masked by the aggregate results. Heterogeneity of effects could potentially explain why many of the treatment dummy coefficients reported earlier in the regressions in Chapter 4, although large in magnitude, were insignificant and had very large standard errors.

In this chapter, we disaggregate analyses by players' ability. Most of the analyses presented here rely on the categorisation of two groups of participants, each with similar ability. We refer to these groups as 'high' and 'low' performers. Details relating to how these groups are defined are presented below in Section 6.1. Having defined these two ability categories, we conduct analyses along two dimensions. The first dimension of analyses looks at how participants perform across treatments within each category of high and low performers. This allows us to distinguish the effect of treatment interventions for people of different ability. We can, in turn, compare these findings to the aggregate results reported earlier to see its composition. Along this dimension of analyses we will also look at learning across treatments for both high and low performers.

The second dimension of analyses directly compares the performance of high and low performers in each treatment. In comparing high and low performers, we focus on performance dynamics – how high performers learn in relation to low performers. This allows us to infer whether the gap in performance between them narrows or widens over time. This chapter is primarily structured along the two dimensions of analyses. However, before we proceed with analyses, we start by outlining how high and low performers are categorised.

6.1. Defining Ability

Since this chapter focuses on ability, we ought to discuss how we identify and measure people's ability. We measure a player's ability by how they performed across the first five rounds of play.

Since the first five rounds of play are identical across treatments, they are not influenced by the treatment interventions that come into play after round 5. Furthermore, since these are initial rounds, they are less likely to be influenced by learning. For these reasons, the performance over rounds 1 to 5 serve as a good proxy for each participant's underlying ability.

More specifically, we define ability to be the median forecast error of subjects across the first five rounds of play. It is more appropriate to use the median over the mean to define ability, since the median is robust to outliers which may arise from bad trials in these early rounds. We will simply refer to this median statistic as ability from here onwards. Earlier in Chapter 4, the average ability and its distribution were presented in Table 4.1 and Figure 4.1 respectively. From these, which will not be reproduced here, we can see that there are no differences in participants' ability across treatments in both the single cue (χ^2 = 1.60, p = 0.659, n = 166) and the dual cue task (χ^2 = 2.04, p = 0.564, n = 146), based on Kruskal Wallis tests.

With this measure of ability, we in turn define two groups of participants: high and low performers. We distinguish high performers from low performers according to a median performance split. Participants whose ability is higher than or equal to the median forecast error across all participants in each of the single and dual cue tasks are considered to be high performers, and the others are low performers. Since higher ability is represented by smaller forecast error values, if the within-subject median forecast error of a participant is lower than or equal to the median threshold, then we classify him to be a high performer. Low performers have a median forecast error higher than this median threshold. The median threshold for each version of the forecasting game is the median forecast error across both dimensions of participants and time over the first five rounds of play. Accordingly, the median threshold is defined at the task level and does not vary by treatment. These forecast error thresholds are 8 in the single cue task and 21 in the dual cue task.

Table 6.1 shows how high and low performers are split across treatments. We see that in most treatments, the proportions of high and low performers within each treatment are similar, with a slightly larger proportion of high performers than low performers. A series of two-sided binomial tests run for each treatment shows that in all but one treatment, there are no differences between the proportions of high and low performers. In most treatments, high and low performers are well balanced. However in the single cue S treatment, there is a significantly higher proportion of high performers than low performers at the margin (64% vs 36%), with a p-value of 0.088.

Table 6.1 Distribution of High and Low Performers by Median Split, by Treatment

	High Performers	Low Performers	Total	Binomial Test
Single Cue				
PR	23 (55%)	19 (45%)	42 (100%)	p = 0.644
PRWL	23 (55%)	19 (45%)	42 (100%)	p = 0.644
T	24 (60%)	16 (40%)	40 (100%)	p = 0.268
S	27 (64%)	15 (36%)	42 (100%)	**p = 0.088**
Dual Cue				
PR	21 (54%)	18 (46%)	39 (100%)	p = 0.749
PRWL	19 (54%)	16 (46%)	35 (100%)	p = 0.736
T	18 (47%)	20 (53%)	38 (100%)	p = 0.871
S	18 (53%)	16 (47%)	34 (100%)	p = 0.864

Number of high/low performing participants, by treatment. Proportions relative to treatment size are in parentheses. The last column shows two-sided binomial tests of whether there are any differences in the proportions of high and low performers in each treatment; bold typeface indicates significance at the 10% level or better.

The fact that there is a treatment with different properties highlights the merits of this categorisation. In the aggregate results, the single cue S treatment might have only outperformed other treatments because of the higher proportion of high performing participants. With the categorisation of high and low performers, we can isolate this effect by seeing how people of similar ability perform across treatments. In this regard, the different proportions of high and low performers in a particular treatment is not relevant. Rather, it is more important to check that a) there are no systematic differences in ability across treatments within each ability category, and b) that there is sufficient distinction in ability levels for high and low performers for each treatment.

Table 6.2 Average Forecast Errors for High and Low Performers by Treatment

Single Cue Task	High Performers	Low Performers	High Perf = Low Perf
PR (n = 42)	7.09 [6] (8.59)	21.39 [13] (19.50)	**$\lvert z \rvert = 5.54$** **p = 0.000**
PRWL (n = 42)	8.64 [6] (14.85)	18.63 [13] (17.87)	**$\lvert z \rvert = 5.55$** **p = 0.000**
T (n = 40)	8.36 [5.5] (14.08)	20.39 [13] (25.53)	**$\lvert z \rvert = 5.33$** **p = 0.000**
S (n = 42)	9.00 [5] (18.96)	21.21 [14] (21.57)	**$\lvert z \rvert = 5.35$** **p = 0.000**
PR = PRWL = T = S	$\chi^2(3) = 0.94$ p = 0.817	$\chi^2(3) = 0.47$ p = 0.925	
Dual Cue Task	High Performers	Low Performers	High Perf = Low Perf
PR (n = 39)	24.21 [11] (34.60)	47.72 [29.5] (53.18)	**$\lvert z \rvert = 5.34$** **p = 0.000**
PRWL (n = 35)	18.24 [10] (22.49)	42.75 [30.5] (39.91)	**$\lvert z \rvert = 5.04$** **p = 0.000**
T (n = 38)	23.67 [14.5] (27.65)	41.48 [31.5] (32.48)	**$\lvert z \rvert = 5.27$** **p = 0.000**
S (n = 34)	19.16 [11] (25.58)	40.58 [30.5] (33.44)	**$\lvert z \rvert = 4.98$** **p = 0.000**
PR = PRWL = T = S	**$\chi^2(3) = 5.14$** **p = 0.162**	**$\chi^2(3) = 1.91$** **p = 0.591**	

Mean forecast errors of high and low performers by treatment and version of task. Median values are in square brackets, representing average ability. Standard deviations are in parentheses. Averages taken across subjects and rounds, in rounds 1 to 5. A series of Kruskal Wallis test with subject median forecast errors (ability) in the first five rounds is shown in the below the descriptive statistics for each treatment. Wilcoxon Rank Sum tests of the difference in subject median forecast errors in the first five rounds between high and low performers in each treatment are presented in the last column. Bold typeface indicates significance at the 10% level or better.

Table 6.2 displays descriptive statistics for high and low performers' forecast errors in the first five rounds by treatment. There we see that forecast errors are remarkably similar across treatments within each category of ability. A series of Kruskal Wallis tests of each participant's ability (median forecast errors) across treatments, displayed below the descriptive statistics for each treatment, shows that there are no significant differences in ability across treatments for each ability category.

From the descriptive statistics presented in Table 6.2, it is also clear that the ability of high performers is better than that for low performers in every treatment for both single and dual cue tasks. These are supported by a series of ranksum tests of participant ability (median forecast error of participants) between high and low performers, conducted for each treatment. These tests are presented in the last column of Table 6.2 and show stark differences in ability, with p-values for each treatment to be 0.000.

6.2. Ability and Treatment Effects

6.2.1. Hypotheses

In this section, we compare participants' forecast errors across treatments within the categories of high and low performers. Before proceeding with formal analyses, we discuss the various *a priori* effects that we would expect to occur. In some instances, these effects differ by player ability.

The first cluster of effects are brought about by the pay schemes of piece rates, tournaments and salaries if players were motivated solely by the respective monetary payoffs. We first focus on the effects for high performers. Under piece rates, players are paid according to their individual performance in each round. High performers would therefore be expected to exert a high level of effort under a piece rate scheme. According to the Piece Rate Equivalence property of tournaments, we would also expect tournaments to induce a similar level of effort from high performers.[44] Under fixed salaries, players are not rewarded for the effort that is exerted, so it follows that high performers would exert minimal effort. In other words, if participants are motivated solely by money, high performers would perform better in performance pay schemes – piece rates and tournaments – than under the performance invariant pay scheme of salaries.

[44] The property of Piece Rate Equivalence did not consider agents of different ability. By a simple extension, we would expect the property to continue to hold in particular circumstances even when ability is defined. Under a two player symmetric tournament where each player has an ability of α, under Piece Rate Equivalence, we would expect equilibrium effort levels to be identical to an α ability player facing a piece rate.

We would, therefore, expect forecast errors to be higher in the S treatment than in the PR and T treatments, with no difference in forecast errors between the latter treatments.

For low performers, we would expect similar effects if money was their sole consideration. Low performers under piece rates and tournaments would exert the same effort. This effort would be considered to be 'high', although lower than that of high performers due to the higher marginal cost of effort. Under fixed salaries, low performers would again exert minimal effort. For low performers who are motivated only by monetary payoffs, as with high performers, we would expect better performance under piece rates and tournaments than under salaries.

The second cluster of effects come about from the degree of control that is associated with each pay scheme. According to Cognitive Evaluation Theory, the intrinsic motivation of people falls as they are exposed to situations that are considered to be controlling, reducing their perceived autonomy. In turn, lower intrinsic motivation leads to lower performance. Performance pay schemes are considered to be highly controlling as they require people to perform well in order to receive large payoffs. The performance requirement is more stringent under tournaments than piece rates, since to achieve higher earnings under tournaments, players need not only to improve their performance, but to improve it such that it is higher than their partner's performance. Tournaments are therefore more controlling than piece rates are. Salaries are least controlling, since they are invariant to performance. As such, there is no extrinsic pressure for them to perform. If we rank the treatments in terms of their degree of control, we observe that the S treatment is least controlling, followed by the PR, and with the T treatment being the most controlling. Accordingly, if this control effect is prominent, then we would expect best performance in the S treatment, second in the PR treatment and worst in the T treatment. The effect of control does not distinguish between people of different ability, so we would expect the same effects for both high and low performers.

The third effect focuses on relative feedback and people's desire to win. Comparing the PR and PRWL treatments, although both pay according to piece rates, we would expect both high and low performers in the PRWL treatment to perform better than their counterparts in the PR treatment. Assuming that people have preferences for winning, we would expect both high and low performers in the PRWL treatment to exert greater effort in order to improve (reduce) their chances of winning (losing).

The fourth effect that we might encounter relates to the competency evaluation aspect of relative feedback. Relative performance feedback allows players to assess their performance against that of their random partner. According to Cognitive Evaluation Theory, a person would be more (less) intrinsically motivated if they receive (un)favourable feedback, which suggests that they are (not) competent at the task that they are undertaking. High performers in the PRWL treatment, whom we expect to receive winning feedback more often than losing feedback, would be more motivated to perform than high performers in the PR treatment upon receiving this favourable feedback. On the other hand, intrinsic motivation falls when relative feedback is unfavourable, for which we would expect low performers in the PRWL treatment to be less motivated than low performers in the PR treatment, and as a result have lower performance.

These various effects, for which some play out in different directions, make it difficult to make a priori predictions about how high and low ability players perform in each treatment. We proceed with analysis of how high and low ability players perform across treatments without listing formal hypotheses.

6.2.2. Results

Tables 6.3 and 6.4 present random effects estimates of two regression specifications for each of the pooled, single and dual cue data series for a total of 6 regression models in each of the tables for high and low performers respectively. In the first regression specification, forecast errors are regressed against treatment dummies (with PR serving as the reference category), participants' trait anxiety and gender as controls, as well as an aggregate linear trend. The second specification replaces the aggregate time trend with individual trends interacted by treatment, but is otherwise identical. These two regression equations are repeated for each of the pooled, single and dual cue series – yielding six regressions in each table. These six regressions are in turn repeated separately for high and low performers in Tables 6.3 and 6.4 respectively. Panel B of each table presents Wald tests of cross-treatment comparisons.

Table 6.3 Regressions of Treatment Effects for High Performers

Panel A: Regression Results

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PR	(base)	(base)	(base)	(base)	(base)	(base)
PRWL	-0.964 (2.104)	-3.401 (3.409)	-0.077 (1.141)	1.513 (1.820)	-0.569 (2.631)	-5.771 (4.929)
T	1.341 (2.841)	1.424 (4.104)	-0.009 (1.139)	0.881 (2.110)	4.745 (3.104)	5.774 (5.622)
S	-0.416 (2.038)	-3.198 (3.513)	0.698 (1.135)	2.323 (2.072)	-0.309 (3.104)	-6.198 (5.867)
Trait Anxiety	-0.092 (0.110)	-0.092 (0.110)	0.072 (0.051)	0.072 (0.051)	-0.382 ** (0.186)	-0.382 ** (0.186)
Female	5.837 *** (1.719)	5.837 *** (1.720)	0.180 (0.807)	0.180 (0.808)	9.851 *** (2.553)	9.851 *** (2.556)
Round	-0.240 *** (0.060)		-0.050 (0.047)		-0.464 *** (0.113)	
PR*Round		-0.346 *** (0.124)		0.038 (0.100)		-0.652 *** (0.183)
PRWL*Round		-0.158 (0.107)		-0.085 (0.064)		-0.252 (0.230)
T*Round		-0.352 ** (0.138)		-0.030 (0.113)		-0.731 *** (0.240)
S*Round		-0.132 (0.105)		-0.088 (0.092)		-0.199 (0.226)
Constant	17.85 *** (4.685)	19.23 *** (5.138)	4.917 ** (2.203)	3.784 (2.541)	36.41 *** (7.654)	38.86 *** (8.008)
Observations	2355	2355	1275	1275	1080	1080
Participants	157	157	85	85	72	72
R^2	0.027	0.028	0.004	0.004	0.054	0.056
Wald χ^2	31.14	32.69	5.55	7.16	38.07	49.30
$p > \chi^2$	0.000	0.000	0.475	0.621	0.000	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Hypothesis Tests

	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PRWL = 0	χ^2 = 0.21 p = 0.647	χ^2 = 1.00 p = 0.318	χ^2 = 0.00 p = 0.946	χ^2 = 0.69 p = 0.406	χ^2 = 0.05 p = 0.829	χ^2 = 1.37 p = 0.242
PRWL = T	χ^2 = 0.76 p = 0.382	χ^2 = 1.86 p = 0.173	χ^2 = 0.01 p = 0.939	χ^2 = 0.10 p = 0.747	χ^2 = 1.57 p = 0.211	**χ^2 = 4.43** **p = 0.035**
T = 0	χ^2 = 0.22 p = 0.637	χ^2 = 0.12 p = 0.729	χ^2 = 0.00 p = 0.994	χ^2 = 0.17 p = 0.676	χ^2 = 1.26 p = 0.262	χ^2 = 1.05 p = 0.304
S = 0	χ^2 = 0.04 p = 0.838	χ^2 = 0.83 p = 0.363	χ^2 = 0.38 p = 0.539	χ^2 = 1.26 p = 0.262	χ^2 = 0.01 p = 0.921	χ^2 = 1.12 p = 0.291
T = S	χ^2 = 0.46 p = 0.498	χ^2 = 1.62 p = 0.203	χ^2 = 0.40 p = 0.528	χ^2 = 0.44 p = 0.507	χ^2 = 1.20 p = 0.273	**χ^2 = 3.59** **p = 0.058**
PR*Round = PRWL*Round		χ^2 = 1.30 p = 0.254		χ^2 = 1.06 p = 0.304		χ^2 = 1.85 p = 0.174
PRWL*Round = T*Round		χ^2 = 1.23 p = 0.267		χ^2 = 0.18 p = 0.670		χ^2 = 2.08 p = 0.149
PR*Round = T*Round		χ^2 = 0.00 p = 0.973		χ^2 = 0.20 p = 0.656		χ^2 = 0.07 p = 0.793
PR*Round = S*Round		χ^2 = 1.73 p = 0.188		χ^2 = 0.84 p = 0.359		χ^2 = 2.43 p = 0.119
T*Round = S*Round		χ^2 = 1.62 p = 0.203		χ^2 = 0.16 p = 0.691		χ^2 = 2.62 p = 0.106

Bold typeface indicates statistical significance at the 10% level or better.

Focusing first on the overall treatment effects amongst high performers, represented by the treatment dummies in Table 6.3, we see that the treatment interventions have little effect for high performers. Amongst the treatment dummies – PRWL, T and S – none are statistically significant from zero in each of models 1 to 6, suggesting that these treatments do not differ from the PR treatment which serves as the reference category. Using the Wald chi-squared tests shown in Panel B to make pairwise treatment comparisons, we see that there are no differences across treatments in models 1 to 5 across the various treatments. In regression model 6, the dual cue task which allows for treatment-specific learning, we find evidence that the T treatment performs worse than both the PRWL and S treatments (p = 0.035 and p = 0.058 respectively).

The absence of cross-treatment differences in performance suggests two things. First, pay schemes have little influence over the performance of high ability players. By comparing the performance pay schemes in a pairwise manner, we see that piece rates and tournaments perform similarly – shown by the insignificant T dummy in every regression in Table 6.3. In addition, we find that high performers perform similarly in both PR and S treatments. The forecast errors between the T and S treatments are no different in the pooled and single cue regressions. In the dual cue task in model 5, we again find no statistical differences between the T and S treatments, but we do indeed observe that the T treatment performs worse than the S treatment once different rates of learning are accounted for in the dual cue task (p = 0.058).

The second thing that the results show is that high ability players are unresponsive to the provision of relative feedback. This is shown by the insignificant coefficient on the PRWL dummy, with the PR treatment serving as the reference category. The finding that both PR and PRWL treatments perform similarly suggests that feedback on relative performance does not affect high performers. This could be interpreted in various ways. The first interpretation is that capable players simply do not care about relative feedback because the feedback is redundant in affirming these high performers of their own ability. An alternative interpretation is that the relative feedback works to motivate these high performers, but the higher motivation does not translate to significant improvements to forecasts, since it is increasingly difficult to improve upon already accurate forecasts. The latter interpretation bears some credibility when the PRWL coefficients in Table 6.3 are compared across the single and dual cue tasks. Since the dual cue task is the more difficult one with higher forecast errors, there is greater scope to improve forecasts vis à vis the single cue task. Although none of the coefficients are statistically significant, we see that the coefficients on the PRWL dummy in the dual cue task with and without treatment-round interactions (-0.569 and -5.771 in models 5 and 6 respectively) are both negative and are larger in magnitude than the corresponding estimates for the single cue task (-0.077 and 1.513 in models 3 and 4).

We now turn our attention to learning for high performers. In model 1 of Table 6.3, the round coefficient is negative and statistically significant when single and dual cue tasks are pooled. This suggests that high performers, overall, exhibit learning. Regression models 3 and 5 replicate the regression but only with single and dual cue data respectively. The linear trends in models 3 and 5 show that the overall learning by high performers is driven by participants in the dual cue task, with a larger downward trend than in the single cue task, where the trend is not significant.

In the even-numbered regression models of Table 6.3, we disaggregate the time trends by treatment. While the estimated trends are negative for each treatment in each regression model, they are only statistically significant in the pooled and dual cue PR and T treatments. We elaborate on learning in Section 6.3.

These findings for high performers differ from the aggregate results, Results 4.1 to 4.5. The aggregate results from Chapter 4 found that relative feedback improved performance, the PRWL treatment outperformed the T treatment, and that the S treatment performed better than both the PR and T treatments. The only finding that persists with high performers is the Piece Rate Equivalence result: tournaments perform no differently to piece rates. Since most of the results for high performers do not resemble the aggregate findings, we can infer that the aggregate results are driven by low performers.

Table 6.4 presents the analogous regressions for low performers. The findings are similar to the aggregate results. First by comparing the different pay schemes, we see that the T dummy is insignificant in every regression model, suggesting that the T treatment performs similarly to the reference PR treatment. It is interesting to note that in each of the pooled, single and dual cue regressions – despite being insignificant – the T dummy is negative in the odd-numbered regression models where all treatments share a common time trend, but the same coefficient becomes positive with a much larger magnitude when treatment-specific trends are included in the even-numbered regression models. We will touch on this shortly when we discuss learning.

Amongst low performers, the S treatment performs particularly well. The S treatment dummy is negative and statistically significant in the pooled regression models, showing forecast errors for low performers to be approximately 8 points lower in the S treatment than in the PR treatment. In models 3 and 5, the S dummy is insignificant at conventional levels, but only marginally so at the 10% level, with p-values of 0.104 and 0.132 respectively. Wald tests presented in Panel B of Table 6.4 shows that the S treatment also performs better than the T treatment.

Table 6.4 Regressions of Treatment Effects for Low Performers

Panel A: Regression Results

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Piece Rate	(base)	(base)	(base)	(base)	(base)	(base)
Piece Rate Win Lose	-7.532 * (4.186)	-9.112 (5.713)	-2.232 (3.075)	-7.122 (6.590)	-8.284 (6.102)	-7.248 (9.255)
Tournament	-0.101 (4.962)	5.498 (5.309)	-0.199 (3.430)	0.665 (5.554)	-0.920 (7.438)	9.626 (7.803)
Salary	-7.971 * (4.217)	-8.755 * (4.629)	-4.650 (2.862)	-7.038 (4.883)	-9.663 (6.407)	-8.843 (7.343)
Trait Anxiety	0.206 (0.180)	0.206 (0.180)	-0.179 (0.159)	-0.179 (0.159)	0.249 (0.264)	0.249 (0.264)
Female	8.271 *** (2.611)	8.271 *** (2.613)	3.223 (2.218)	3.223 (2.222)	3.887 (3.912)	3.887 (3.918)
Round	0.016 (0.131)		-0.221 (0.144)		0.253 (0.221)	
PR*Round		0.088 (0.271)		-0.354 (0.254)		0.530 (0.457)
PRWL*Round		0.210 (0.318)		0.023 (0.407)		0.450 (0.504)
T*Round		-0.343 * (0.200)		-0.420 *** (0.164)		-0.282 (0.335)
S*Round		0.148 (0.251)		0.170 (0.140)		0.467 (0.470)
Constant	12.95 * (7.695)	13.89 * (7.741)	23.13 *** (7.854)	24.85 *** (8.589)	21.04 * (12.06)	17.45 (12.87)
Observations	1890	1890	945	945	945	945
Participants	126	126	63	63	63	63
R^2	0.034	0.035	0.022	0.024	0.018	0.020
Wald χ^2	13.92	19.93	11.02	17.36	5.27	9.65
$p > \chi^2$	0.031	0.018	0.088	0.043	0.510	0.380

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Hypothesis Tests

	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PRWL = 0	**χ^2 = 3.24** **p = 0.072**	χ^2 = 2.54 p = 0.111	χ^2 = 0.53 p = 0.468	χ^2 = 1.17 p = 0.280	χ^2 = 1.84 p = 0.175	χ^2 = 0.61 p = 0.434
PRWL = T	**χ^2 = 3.52** **p = 0.061**	**χ^2 = 5.75** **p = 0.017**	χ^2 = 0.59 p = 0.441	χ^2 = 1.81 p = 0.178	χ^2 = 1.95 p = 0.162	**χ^2 =2.93** **p = 0.087**
T = 0	χ^2 = 0.00 p = 0.984	χ^2 = 1.07 p = 0.300	χ^2 = 0.00 p = 0.954	χ^2 = 0.01 p = 0.905	χ^2 = 0.02 p = 0.902	χ^2 = 1.52 p = 0.217
S = 0	**χ^2 = 3.57** **p = 0.059**	**χ^2 = 3.58** **p = 0.059**	χ^2 = 2.64 p = 0.104	χ^2 = 2.08 p = 0.150	χ^2 = 2.27 p = 0.132	χ^2 = 1.45 p = 0.229
T = S	**χ^2 = 3.94** **p = 0.047**	**χ^2 = 8.16** **p = 0.004**	**χ^2 = 3.47** **p = 0.062**	**χ^2 = 4.06** **p = 0.044**	χ^2 = 2.50 p = 0.114	**χ^2 = 5.55** **p = 0.019**
PR*Round = PRWL*Round		χ^2 = 0.08 p = 0.771		χ^2 = 0.61 p = 0.434		χ^2 = 0.01 p = 0.907
PRWL*Round = T*Round		χ^2 = 2.16 p = 0.142		χ^2 = 1.02 p = 0.314		χ^2 = 1.46 p = 0.227
PR*Round = T*Round		χ^2 = 1.63 p = 0.201		χ^2 = 0.05 p = 0.826		χ^2 = 2.05 p = 0.152
PR*Round = S*Round		χ^2 = 0.03 p = 0.871		χ^2 = 0.40 p = 0.527		χ^2 = 0.01 p = 0.923
T*Round = S*Round		χ^2 = 2.34 p = 0.126		χ^2 = 1.35 p = 0.246		χ^2 = 1.68 p = 0.195

Bold typeface indicates statistical significance at the 10% level or better.

It is interesting to see low performing S players outperform their low performing counterparts in the PR and T treatments, who are paid for their performance. If participants are motivated solely by money, then we would expect PR and T subjects to outperform S players, who would be expected to shirk. This is not the case. The finding that S players perform better can be explained by the higher degree of autonomy players have under salaries compared to piece rates and tournaments, which are considered to be more controlling. Cognitive Evaluation Theory predicts S participants to have higher intrinsic motivation than those in PR and T treatments – which in turn leads to higher performance. The fact that low performers have smaller forecast errors in the S treatment compared to the PR and T treatments suggests that this intrinsic motivation effect overpowers the extrinsic motivation effect brought about by monetary incentives.

We have found that the choice of pay scheme impacts the performance of low performers. The next question is whether low performers are affected by the provision of relative performance feedback. Since PR serves as the reference treatment in the regressions, the coefficient on the PRWL dummy variable identifies the effect of feedback provision. From Table 6.4, we can see that the PRWL dummy has a large negative coefficient in most regression models. In the pooled regressions, it is statistically significant in model 1, and marginally misses out on 10% significance in model 2 with a p-value of 0.111. Broken down by the individual tasks, the PRWL dummy is no longer significant in regression models 3 to 6, although the coefficient continues to be negative and in the case for models 4 to 6, the coefficients are similar in magnitude to the estimates from the pooled regressions. The single and dual cue regressions of models 3 to 6 in Table 6.4 are likely to be concealing the underlying effect which the pooled regressions show, since the number of observations in these regressions are spread thin.

With evidence to show that the PRWL treatment performs better than the PR treatment, this means that low performers perform better when relative feedback is present. We previously also found that low performers in the PR and T treatments had similar forecast errors. These two findings jointly suggest that the result of Piece Rate Equivalence of tournaments for low performers is masking compositional differences, where a component effect includes relative feedback motivating performance. Having controlled for relative feedback, do low performers continue to perform similarly in piece rates and tournaments? In other words for low performers, how does the PRWL treatment perform relative to the T treatment? Wald tests in Panel B of Table 6.4 show forecast errors of low performers to be smaller in the PRWL treatment than in the T treatment. Like the aggregate findings reported earlier in Chapter 4, the primary reason why tournaments perform like piece rates is due to the competitive nature of tournaments. When this element of competition is introduced to piece rates, Piece Rate Equivalence no longer holds, with piece rate incentives being superior to rank-dependent rewards.

The findings presented for the low performers in Table 6.4 reflect the aggregate results reported earlier in Chapter 4, before distinction was made between high and low ability participants. Since high performers do not respond to treatment interventions, it means that the aggregated results are driven by low performers. Result 6.1 summarises the main finding:

Result 6.1. *Treatment effects are driven by low performers. High performers have similar performance across treatments. On the other hand for low performers, performance is higher in the PRWL and S treatments than in the PR and T treatments.*

In terms of learning for low performers, from Table 6.4, we do not observe any learning for low performers overall. The common trend for each of the pooled, single and dual cue treatments are statistically insignificant in regression models 1, 3 and 5 respectively. This contrasts with the common downward trend we found for high performers in Table 6.3. When we allow for different trends by treatment, we only observe learning for low performers in the T treatment.

We elaborate on the issue of learning in a number of ways in the following sections. We disaggregate the common trend by treatment, while also expanding analyses of learning to the round 11 to 20 time horizon. In later sections, rather than comparing the trends across treatments, we compare the trends across high and low performers for each treatment – touching upon the notions of bifurcation and catching up.

6.3. Learning

Having explored how treatment interventions affect the performance of high and low performers overall, we now pay closer attention to learning. Within each group of high and low performers, how do the treatment interventions influence learning? The aggregated learning results presented earlier in Chapter 5 showed us that learning only occurred in the T treatment – none of the other treatments showed any signs of learning. We now look at how the earlier reported learning results differ by high and low performers.

Tables 6.5 and 6.6 present time trends for estimated high and low performers respectively. Similar to how results were presented earlier in Chapter 5, we only report on the slopes for each treatment's trend line while suppressing the other coefficients in a larger regression specification. These other regressors include the treatment dummies, as well as the trait anxiety and gender of each participant. The odd-numbered regression models in Tables 6.5 and 6.6 are identical to the even-numbered regressions in Tables 6.3 and 6.4, replicated to facilitate comparison with other regression models. These regressions are run over rounds 6 to 20, the post-intervention rounds. We carry over an earlier argument made about allowing an arbitrary five rounds for participants to familiarise themselves with the treatment interventions in the PRWL and T treatments. For this reason we also look at learning across rounds 11 to 20. The even-numbered regression models in Tables 6.5 and 6.6 are run over rounds 11 to 20.

Beginning with high performers in Table 6.5, in model 1 we see that there is a significant downward trend for the PR and T treatments across rounds 6 to 20. These two trends are similar in magnitude and are not statistically different from one another (p = 0.973). Interestingly, while these trend coefficients are statistically different from zero and the estimated trends for PRWL and S treatments are not, pairwise Wald tests do not show the PR or T trends to be different to the PRWL or S trends. Looking deeper, models 3 and 5 provide some disaggregation by replicating the regression specification of model 1 with single and dual cue data. From these, it appears that this learning is driven by dual cue participants, where we observe a similar pattern of learning. Dual cue PR and T participants exhibit significant learning, while the trends for PRWL and S are insignificant. On the other hand, in regression model 3, there is no learning for any of the single cue treatments across rounds 6 to 20.

If we track the learning of high performers across a different time period, rounds 11 to 20, the pattern of learning across treatments differ substantially from the round 6-20 horizon. In regression model 2 of Table 6.5, we look at learning across rounds 11 to 20 when single and dual cue tasks are pooled together. Learning occurs in each of the four treatments, with statistically significant time trends for each treatment. A series of Wald tests show that learning does not vary by treatment. Looking at regression model 4, it is clear that this pattern of universal learning in every treatment is attributable to the single cue task, where we again observe significant rates of learning for high performers in each of the treatments. Amongst these trends, the trend for the PRWL treatment is largest in magnitude, and is statistically different to the PR trend (p = 0.055). Across rounds 11 to 20 in the dual cue task, although the estimated trend coefficients are negative in each of the treatments, they are only statistically significant in the T treatment.

Table 6.5 Regressions of Learning for High Performers

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
PR*Round	-0.346 *** (0.124)	-0.381 ** (0.184)	0.038 (0.100)	-0.602 *** (0.136)	-0.652 *** (0.183)	-0.204 (0.310)
PRWL*Round	-0.158 (0.107)	-0.644 ** (0.308)	-0.085 (0.634)	-1.055 *** (0.193)	-0.252 (0.230)	-0.120 (0.640)
T*Round	-0.352 ** (0.138)	-0.761 *** (0.171)	-0.030 (0.113)	-0.755 *** (0.194)	-0.731 *** (0.240)	-0.768 *** (0.298)
S*Round	-0.132 (0.105)	-0.539 * (0.276)	-0.088 (0.092)	-0.812 *** (0.238)	-0.199 (0.226)	-0.121 (0.586)
Observations	2355	1570	1275	850	1080	720
Participants	157	157	85	85	72	72
R^2	0.028	0.031	0.004	0.052	0.056	0.059
Wald χ^2	32.69	46.81	7.16	90.33	49.30	27.29
$p > \chi^2$	0.000	0.000	0.621	0.000	0.000	0.001
PR*Round = PRWL*Round	$\chi^2 = 1.30$ p = 0.254	$\chi^2 = 0.54$ p =0.462	$\chi^2 = 1.06$ p = 0.304	**$\chi^2 = 3.67$** **p = 0.055**	$\chi^2 = 1.85$ p = 0.174	$\chi^2 = 0.01$ p = 0.906
PRWL*Round = T*Round	$\chi^2 = 1.23$ p = 0.267	$\chi^2 =0.11$ p = 0.741	$\chi^2 = 0.18$ p = 0.670	$\chi^2 = 1.19$ p = 0.274	$\chi^2 = 2.08$ p = 0.149	$\chi^2 = 0.84$ p = 0.359
PR*Round = T*Round	$\chi^2 = 0.00$ p = 0.973	$\chi^2 = 2.28$ p = 0.131	$\chi^2 = 0.20$ p = 0.656	$\chi^2 = 0.42$ p = 0.519	$\chi^2 = 0.07$ p = 0.793	$\chi^2 = 1.72$ p = 0.190
PR*Round = S*Round	$\chi^2 = 1.73$ p = 0.188	$\chi^2 = 0.23$ p = 0.634	$\chi^2 =0.84$ p = 0.359	$\chi^2 = 0.58$ p = 0.445	$\chi^2 = 2.43$ p = 0.119	$\chi^2 = 0.02$ p = 0.901
T*Round = S*Round	$\chi^2 = 1.62$ p = 0.203	$\chi^2 = 0.47$ p = 0.494	$\chi^2 = 0.16$ p = 0.691	$\chi^2 = 0.03$ p = 0.854	$\chi^2 = 2.62$ p = 0.106	$\chi^2 = 0.97$ p = 0.326

Partial results only. Other coefficients that are suppressed include: treatment dummies, trait anxiety, gender and the constant term. Models 1, 3 and 5 are replicated from models 2, 4 and 6 of Table 6.3. Regressions are estimated with a Random Effects GLS procedure. Forecast errors in parentheses and are clustered by participant. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald χ^2 tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Table 6.6 Regressions of Learning for Low Performers

Dep Var: Forecast Errors	Pooled		Single Cue		Dual Cue	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
PR*Round	0.088 (0.271)	-0.027 (0.432)	-0.354 (0.254)	-0.800 *** (0.232)	0.530 (0.457)	0.747 (0.792)
PRWL*Round	0.210 (0.318)	0.030 (0.661)	0.023 (0.407)	-0.269 (0.685)	0.450 (0.504)	0.415 (1.235)
T*Round	-0.343 * (0.199)	-0.791 ** (0.344)	-0.420 *** (0.164)	-1.830 *** (0.340)	-0.282 (0.335)	0.029 (0.478)
S*Round	0.148 (0.251)	-0.000 (0.605)	-0.170 (0.140)	-1.217 *** (0.334)	0.467 (0.470)	1.216 (1.075)
Observations	1890	1260	945	630	945	630
Participants	126	126	63	63	63	63
R^2	0.035	0.033	0.024	0.043	0.020	0.017
Wald χ^2	19.93	25.03	17.36	84.58	9.65	6.15
$p > \chi^2$	0.018	0.003	0.043	0.000	0.380	0.725
PR*Round = PRWL*Round	χ^2 = 0.08 p = 0.771	χ^2 = 0.01 p = 0.943	χ^2 = 0.61 p = 0.434	χ^2 = 0.54 p = 0.463	χ^2 = 0.01 p = 0.907	χ^2 = 0.05 p = 0.821
PRWL*Round = T*Round	χ^2 = 2.16 p = 0.142	χ^2 = 1.22 p = 0.270	χ^2 = 1.02 p = 0.314	**χ^2 = 4.17** **p = 0.041**	χ^2 = 1.46 p = 0.227	χ^2 = 0.08 p = 0.771
PR*Round = T*Round	χ^2 = 1.63 p = 0.201	χ^2 = 1.92 p = 0.166	χ^2 = 0.05 p = 0.826	**χ^2 = 6.25** **p = 0.012**	χ^2 = 2.05 p = 0.152	χ^2 = 0.60 p = 0.438
PR*Round = S*Round	χ^2 = 0.03 p = 0.871	χ^2 = 0.00 p = 0.972	χ^2 = 0.40 p = 0.527	χ^2 = 1.05 p = 0.306	χ^2 = 0.01 p = 0.923	χ^2 = 0.12 p = 0.725
T*Round = S*Round	χ^2 = 2.34 p = 0.126	χ^2 = 1.29 p = 0.255	χ^2 = 1.35 p = 0.246	χ^2 = 1.65 p = 0.199	χ^2 = 1.68 p = 0.195	χ^2 = 1.02 p = 0.313

Partial results only. Other coefficients that are suppressed include: treatment dummies, trait anxiety, gender and the constant term. Models 1, 3 and 5 are replicated from models 2, 4 and 6 of Table 6.4. Regressions are estimated with a Random Effects GLS procedure. Forecast errors in parentheses and are clustered by participant. *, **, *** represent significance at the 10%, 5% and 1% levels respectively. Wald χ^2 tests are presented at the bottom of the table. Bold typeface indicates a Wald test to be significant at the 10% level or better.

Comparing the patterns of learning for high performers across the different time horizons, we see that they are clearly different. The pooled regressions of models 1 and 2 of Table 6.5 show that while learning occurs only in the PR and T treatments over rounds 6 to 20, there is evidence for learning in all treatments over rounds 11 to 20. The discrepancies in these trends can be attributed back to the single cue task, where the differences in learning across each of these time horizons is stark. On the other hand, as we can see in models 5 and 6, learning in the dual cue task across rounds 11 to 20 are more reminiscent of the learning across rounds 6 to 20.

We now turn our attention to learning for low performers. In the pooled regressions for low performers in model 1 of Table 6.6, learning occurs in the T treatment but not in other treatments. Forecast errors for low performers in the T treatment improve on average by 0.343 points every round. It is interesting to note that this rate of learning for low performing T participants is similar to that of high performers (-0.352 in model 1 of Table 6.5). We see a similar pattern of learning in the single cue task in model 3, where the T treatment again stands out by being the only treatment that exhibits learning. Although there is no statistical evidence to suggest that learning occurs in any of the treatments in regression model 5 for low performers in the dual cue task across rounds 6 to 20, a closer look at the trend coefficients show that the trend is downward sloping in the T treatment, while is upward sloping and with a larger magnitude in each of the other treatments. Despite being insignificant, this reinforces the basic pattern of learning in each of the pooled and single cue regressions in models 1 and 3.

Across rounds 11 to 20, the overall pattern of learning for low performers is similar to that across rounds 6 to 20. In regression model 2 of Table 6.6, we see once again that learning occurs only in the T treatment. In fact, learning in the T treatment across rounds 11 to 20 occurs at more than twice the rate of that across rounds 6 to 20 (-0.791 in model 2 vs -0.343 in model 1). Looking at the composition of this learning, regression models 4 and 6 look at the trends associated with single and dual cue treatments separately. In the single cue task, model 4 shows that there is significant learning in the PR, T and S treatments for low performers across rounds 11 to 20. Amongst these significant trends, the trend for the T treatment is the steepest and with the highest level of significance. In other words, even though there is learning in most treatments, the rate of learning is highest in the T treatment. Learning occurs at a significantly faster rate in the T treatment compared to both the PR treatment (p = 0.012) and PRWL treatment (p = 0.041), but is not statistically different to the S treatment (p = 0.199). In the dual cue task, regression model 6 reveals that there is no learning for low performers in any of the four treatments.

To summarise, there are consistent signs of learning for high performers in the PR and T treatments, and also for low performers in the T treatment. This is stated as Result 6.2:

Result 6.2. *There is learning for high performers in the PR and T treatments. Learning is also observed for low performers in the T treatment.*

We focus on learning amongst both high and low performers in the T treatment. For both high and low performers, there is learning in the T treatment across both round 6 to 20 and 11 to 20 time periods. The fact that there is learning for both high and low performers further suggest that the rank-dependent payoffs of tournaments underpin this learning. The rank incentives reward players for winning with a high $1 prize while punishes players for losing with zero monetary earnings every round. These incentives seem to work symmetrically for both high and low performers. Insomuch high performers strive to win by improving their forecast errors, low performers also have the incentive to improve their performance to avoid losing.[45]

Similar to how we attributed learning in tournaments to the rank dependent payoffs earlier in Chapter 5, we attribute learning here for both high and low performers in the T treatment to these rank payoffs. Consistent with what we have found earlier in Chapter 5 on learning, where there was no learning associated with the PRWL treatment, we find here that neither high nor low performers improve their forecasts over time. The absence of learning in the PRWL treatment is not masked by compositional differences by high and low performers. Comparing the linear trends for PR and PRWL treatments for both high and low performers, we see that, by and large, there is no difference in learning between these two treatments. This suggests that winning/losing feedback has no influence on learning over time. Rather, its motivating effect is one off and remains uniform over time.

[45] See Dutcher et al. (2015) for more about the motives of striving to win and avoiding the loss.

6.4. High and Low Performers within Treatment

The previous sections of this chapter have studied treatment differences within the categories of high and low performers separately, first in terms of the overall effect and then in terms of learning. We now look at the data from another angle, analysing the differences in performance between high and low performers within treatments.

Recall that high performers were defined to have smaller forecast errors than low performers across the first five rounds of play (see Section 6.1 for more details). This however does not necessarily mean that high performers will continue to outperform low performers across the post-intervention rounds. Performance differences across post-intervention rounds could be attributable either to asymmetric effects of the treatment interventions on high or low performers, or due to differences in learning across treatments. The former was addressed in Section 6.2 while the latter in Section 6.3. We now analyse differences in learning between high and low performers. As such the focus on this section will be on learning, as opposed to the overall performance differences between high and low performers.

In looking at how high and low performers learn relative to one another, we touch upon the concepts of bifurcation and catching up. These terms refer to how the performance spread between high and low performers change over time. Bifurcation occurs when the performance gap between high and low performers widens over time. This may arise if high performers improve their performance and/or low performers drop out, or if the rate of learning is greater for high performers than for low performers.[46] On the other hand, catching up refers to instances when the performance gap diminishes.

In analysing bifurcation and catching up, we confine our focus to the PRWL and T treatments, since these two treatments feature the element of competition. It is the knowledge of relative performance or ability that drives bifurcation or catching up behaviours. This is in line with the

[46] The term bifurcation was first coined by Müller and Schotter (2010) to describe the bimodal effort chosen by players, depending on their ability – high ability players tend to exert effort higher than equilibrium, while low ability players choose effort of zero. We adopt a slightly different definition of bifurcation whereby we focus on the *performance spread* between high and low performers *across time*.

literature, where these concepts are studied exclusively in the context of contests and tournaments. We now review the relevant literature before proceeding with analyses.

6.4.1. Literature Review

Both bifurcation and catching up effects feature prominently in the literature. These effects first appeared in Schotter and Weigelt (1992) in their study of uneven tournaments. In a Bull et al. (1987) tournament game, fixed pairs select numerical effort numbers which are costly to players. A high cost subject was matched with a low cost partner for the entire duration of 20 rounds, where these high and low cost players represent low and high ability players respectively. Both players know the effort costs of the other, so they also know whether or not they are disadvantaged. They found that a large proportion of disadvantaged participants chose effort values close to zero, while their partners chose higher-than-equilibrium effort values. Schotter and Weigelt (1992) attributed the dropout behaviour of low ability players to two factors: their partners choosing high effort values, and also to unfavourable draws of luck when random numbers were augmented to the effort values. If the low ability player feels he cannot defeat the high ability player, or that it will be too costly in terms of effort to do so, he will opt to 'drop out' to minimise the effort he exerts. Bifurcation therefore appears to be conditioned behaviour based on how the game plays out. The authors do not offer any explanation as to why high ability players exert higher than equilibrium effort.

Bifurcation behaviour also appears in Müller and Schotter (2010). In their experiment, fixed groups of four participants played in a contest for 50 rounds. Similar to a tournament, players selected costly effort numbers which were ranked within the group and prizes awarded accordingly. Unlike Schotter and Weigelt (1992) where ability was determined by a fixed cost parameter, Müller and Schotter (2010) allowed the ability of a player to change every round, with a random draw of a continuously distributed cost parameter every round. On average, participants' effort levels are consistent with the theoretical predictions. However, bifurcation shows up at the individual level when the data is disaggregated. Effort levels are stepped: effort levels are high when ability is high, but falls sharply to zero when ability falls below a certain threshold.

On the other hand, the catching up effect is observed in a number of different studies. While Schotter and Weigelt (1992) observed bifurcation, they also found evidence of catching up, with low ability players exerting higher-than-optimal effort while high ability players exert effort lower than equilibrium. Subsequent research has shown that this catching up behaviour is conditioned by feedback on relative performance, with players coming to realise that they are falling behind their peers (Kuhnen & Tymula, 2012; Ludwig & Lünser, 2012; Charness et al., 2014; Fu, Ke, & Tan, 2015; Eriksson et al., 2009). Kuhnen and Tymula (2012) suggests that this effect is anchored to expectations. Players who received a lower-than-expected rank improves their rank in subsequent rounds, while those with a higher-than-expected rank subsequently receive lower ranks.

The catching up effect is closely associated with peer effects, where people want to 'keep up with the Joneses'. Falk and Ichino (2006) show that students who were asked to stuff envelopes individually had better performance when they were asked to do so in the presence of a partner in the same room, compared to when they were working alone in the room. The idea is that when the 'peer' is introduced, people are likely to behave so that they do not compare unfavourably with him. An example of this is based on the postcode lottery in the Netherlands. Kuhn et al. (2011) find that when a household wins the lottery and receives a new BMW car as a prize, households in the surrounding neighbourhood are more likely than households in non-winning neighbourhoods to purchase a new car in the following six months. Another example is that of peer salaries. Card et al. (2012) contacted employees from three University of California campuses and notified them of a publicly accessible web portal which allowed them to compare the salaries of all employees at the university – allowing them to benchmark their salaries against their colleagues. The authors followed up with a survey. The respondents who were notified of such comparison and were paid less than their peer group reported lower job satisfaction and greater intentions to find a new job, compared to the control group who were not given details of this salary comparison website. See also peer effects in Feltovich and Ejebu's (2014) life-cycle savings experiment.

The bifurcation and catching up effects are conditioned upon players' perceived chances of winning. In most of the studies cited above, players were provided information which allowed them to assess their chances of winning. For example, in Schotter and Weigelt (1992) players know whether or not they are impaired by an unfavourable cost parameter. In other studies, the provision of relative feedback allow players to assess their chances of winning based on prior rounds of play. The ex-ante knowledge or ex-post feedback seems to be driving the bifurcation and catching up effects. For example, if a player finds out that they are disadvantaged, they are less likely to win and as a result may give up by exerting minimal effort.

Fershtman and Gneezy (2011) support the notion that behaviour is conditioned upon feedback. In their study of quitting in tournaments, male high school students were asked to run in two 60 metre races. In the first, students ran individually with neither competition nor rewards. In the second race, students ran the race while competing with another student with a similar first run time, either directly or indirectly. In the direct race, the pair ran on the same track side by side, so that both could observe the pace and progress of the other. In contrast, students competing indirectly ran individually on the track by themselves. They won if they had a faster running time than their partner, who also ran individually. Running individually in the indirect race does not allow players to gauge the speed and progress of their partners. The winner of each of the direct or indirect races received either a high or low reward, or none at all; the loser received nothing.

Two findings from their paper are relevant here. First, quitting behaviour is mainly observed amongst those who were offered large tournament prizes, with the incidence of quitting in tournaments much lower with low rewards or with no rewards at all. Second and more interestingly, quitting was observed only amongst students who ran in the direct race, where students could observe the progress of their partner. Quitting did not occur in any of the indirect races with no, small or large tournament rewards. This suggests that feedback is necessary to induce dropping out.

While the effects of bifurcation and catching up are conditioned upon feedback and knowledge of the competitor's ability, it is natural to think that repeated interactions are also necessary to bring about these effects. For a low ability player, working harder may only be an optimal strategy for a player if they know that they will be matched with the same partner in the following round. If this low ability player is matched with a higher ability partner in the following round, then working harder is not likely to be effective in securing a win. Similarly, low performers may not feel the need to drop out if there is a possibility they will be matched with someone of lower ability in subsequent rounds. Random rematching allows us to reduce the effect of such conditioning.

Since bifurcation and catching up are expected to be less likely to happen under the protocol of random rematching, if either of these effects are indeed observed, then we can infer that these effects are especially salient. With the example of catching up, if it occurs even under random rematching, then that means that low performers are motivated to improve their performance relative to others, even under the knowledge that they will face a different partner and their improvement in performance might be inadequate to win. If this occurs under random rematching, then we can infer that catching up arises from a player's innate desire to converge, rather than from a rational assessment of whether or not they are likely to win.

Random rematching allows us to study the effects of bifurcation and catching up without players forming predisposed beliefs from prior play that they are bound to win (lose). As such, we would expect a lower incidence of drop out behaviour from low performers under random rematching than under fixed matching. In this regard, we would more likely observe catching up than bifurcation.

While most studies cited here employ a fixed matching protocol, two papers provide some insight into the effects associated with the different matching protocols: fixed matching and random rematching. Müller and Schotter (2010) include a treatment where groups were rematched each round. They find bifurcation persists even with random rematching. On the other hand, Ludwig and Lünser (2012) find evidence for catching up when relative feedback is provided in two-stage tournaments, with rematching after each two-stage round. Both papers provide evidence for the aforementioned effects even with random rematching.

A caveat must be made here with regard to the experimental design of these two papers. In both Müller and Schotter (2010) and Ludwig and Lünser (2012), the rematching across each round was restricted to a smaller fixed pool of participants within the session. In Müller and Schotter (2010), each session was split into pools of 8 participants and 2 groups of four were constructed within these fixed pools for each of 50 rounds. Similarly Ludwig and Lünser (2012) matches 3 pairs within fixed pools of 6 players for each of 30 rounds. Given the limited rematching and the relatively long time horizon in each of these studies, players will inevitably encounter players whom they have played with before, partially undermining the very purpose of rematching. The rematching procedures of Müller and Schotter (2010) and Ludwig and Lünser (2012) are therefore inadequate to address the question of whether bifurcation or catching up effects rely on repeated encounters brought about by a fixed matching protocol.

Our experiment allow us to answer this question. Participants in our study are randomly rematched with another participant after each round. To this end, we will proceed by studying the patterns of learning for high and low performers and see how they learn relative to each other.

6.4.2. Results

We analyse the effects of bifurcation and catching up in three different ways. First, we study the dynamics of how forecast errors change for high and low performers in the PRWL and T treatments. Second, we calculate the between-subjects standard deviation of forecast errors in each round and see how it changes over time. This measures the dispersion of forecast errors across subjects; if bifurcation (catching up) is present, then such dispersion should increase (decrease) over time. Third, we examine the margin of winning between winners and losers and see how it changes over time. We discuss each of these in separate parts below.

Forecast Error Trends for High and Low Performers

We begin analysis of bifurcation and catching up by observing how forecast errors for high and low performers change over time. Tables 6.7 and 6.8 present regressions of participants' forecast errors against a dummy variable which indicates whether a participant is classified as a high performer or not (low performers as the reference category), the trait anxiety and gender of the participant, as well as time trends for both high and low performers. Table 6.7 shows the regression estimates run over the single cue PRWL and T treatments, while Table 6.8 repeats the same regressions over the dual cue treatments. For each regression table, models 1 and 2 are run for the PRWL treatment over rounds 6 to 20 and rounds 11 to 20 respectively, while models 3 and 4 are run for the T treatment over the same periods. At the bottom of the tables are Wald tests of the difference in the estimated trends between high and low performers. Despite the different ways in which the regressions are run, we note that the linear time trends estimated in Tables 6.7 and 6.8 are identical to those reported earlier in Tables 6.5 and 6.6.

Before discussing the time trends, we first discuss the coefficient on the dummy variable for high performers. In most of the regression models across Tables 6.7 and 6.8, this dummy is insignificant, suggesting that forecast errors for both high and low performers are similar in these treatments. This is the case for the single cue PRWL treatment as well as the dual cue PRWL and T treatments. High performers do indeed outperform low performers in the single cue T treatment in models 3 and 4 of Table 6.7.

The finding that high performers do not actually outperform low performers in the post-intervention rounds in most treatments is particularly interesting because they have been defined to have better performance than low performers in the pre-intervention rounds (see Section 6.1). This suggests disproportionate improvement in forecasts by low performers relative to high performers. In fact, if we compare the coefficients on the treatment dummies across Tables 6.3 and 6.4 for high and low performers, we see that – despite many of the coefficients being insignificant – the coefficients for low performers tend to be negative and are larger in magnitude than the analogous coefficients for high performers. The disproportionate effect that treatment interventions have on low performers can, at least in part, explain this finding.

We now return to comparing the rates of learning between high and low performers. In models 1 and 2 of Table 6.7, the single cue PRWL treatment, we see that there is no learning amongst both high and low performers across rounds 6 to 20, where the estimated trends for high and low performers are close to zero in model 1. A Wald test of these differences confirm that they are not different to one another (p = 0.795). Across rounds 11 to 20, model 2 is slightly different for the single cue PRWL treatment. While we continue to see no learning for low performers, we now observe significant learning for high performers. Despite this, a Wald test suggests that the trends are identical (p = 0.272).

Table 6.7 Regressions of Forecast Error Trends between High and Low Performers, by Single Cue Treatments

Dep Var: Forecast Errors	Single Cue PRWL		Single Cue T	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4
Low Performers	(base)	(base)	(base)	(base)
High Performers	-4.058 (4.923)	7.104 (10.02)	-12.13 *** (3.471)	-23.63 *** (7.800)
Trait Anxiety	0.155 (0.107)	0.184 (0.147)	-0.361 ** (0.156)	-0.325 * (0.186)
Female	-1.265 (1.433)	-0.632 (1.732)	0.807 (2.055)	0.211 (2.321)
Low Performers * Round	0.023 (0.409)	-0.269 (0.687)	-0.420 ** (0.165)	-1.830 *** (0.343)
High Performers * Round	-0.085 (0.064)	-1.055 *** (0.194)	-0.030 (0.114)	-0.755 *** (0.197)
Constant	6.427 (8.172)	9.872 (14.44)	34.12 *** (8.133)	56.48 *** (12.75)
Observations	615	410	525	350
Participants	41	41	35	35
R^2	0.028	0.035	0.096	0.130
Wald χ^2	22.11	68.33	16.85	50.02
$p > \chi^2$	0.001	0.000	0.005	0.000
Low Perf*Round = High Perf*Round	χ^2 = 0.07 p = 0.795	χ^2 = 1.21 p = 0.272	**χ^2 = 3.79** **p = 0.052**	**χ^2 = 7.41** **p = 0.007**

Regressions are estimated with Random Effects GLS. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

In the single cue T treatment in model 3 of Table 6.7, we observe that low performers significantly improve their forecast errors over time across rounds 6 to 20, while high performers show no sign of learning. Low performers improve forecast errors on average by 0.42 points every round, significant at the 5% level. Since there is significant learning for low performers but not for high performers, it follows that the performance gap between the two groups diminishes over time, whereby low performers catch up. This is supported by a Wald test of the differences in the rates of learning between high and low performers (p = 0.052).

Table 6.8 Regressions of Forecast Error Trends between High and Low Performers, by Dual Cue Treatments

Dep Var: Forecast Errors	Dual Cue PRWL		Dual Cue T	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4
Low Performers	(base)	(base)	(base)	(base)
High Performers	0.674 (8.438)	-1.776 (22.19)	-5.284 (7.707)	0.487 (11.23)
Trait Anxiety	-0.017 (0.255)	-0.198 (0.264)	-0.109 (0.329)	-0.107 (0.345)
Female	7.514 *** (2.892)	8.443 *** (2.962)	9.882 * (5.129)	11.33 ** (5.192)
Low Performers * Round	0.450 (0.508)	0.415 (1.245)	-0.282 (0.337)	0.029 (0.481)
High Performers * Round	-0.252 (0.233)	-0.120 (0.646)	-0.731 *** (0.241)	-0.768 ** (0.300)
Constant	18.61 (10.53)	26.40 (19.20)	38.19 ** (16.16)	32.07 * (18.43)
Observations	480	320	540	360
Participants	32	32	36	36
R^2	0.063	0.077	0.049	0.058
Wald χ^2	33.57	40.01	24.59	22.63
p > χ^2	0.000	0.000	0.000	0.000
Low Perf*Round = High Perf*Round	χ^2 = 1.58 p = 0.209	χ^2 = 0.15 p = 0.703	χ^2 = 1.18 p = 0.278	χ^2 = 1.97 p = 0.160

Regressions are estimated with Random Effects GLS. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

Across rounds 11 to 20, regression model 4 also shows that the forecast errors of low performers catches up to those of high performers over time in the single cue T treatment. Here both high and low performers show significant learning over time, but the rate of learning is much higher for low performers than for high performers. Low performers improve their forecasts on average by 1.83 points each round, while the average improvement is only 0.76 points for high performers. These rates of learning between high and low performers are significantly different from each other (p = 0.007). Since low performers exhibit a faster rate of improvement than high performers, it follows that the performance gap between high and low performers diminishes. While the pattern of learning is different across rounds 6 to 20 as it is across rounds 11 to 20, both of these patterns support the notion of catching up in the single cue T treatment.

We now look at the rates of learning of high and low performers in the dual cue PRWL and T treatments in Table 6.8. For the dual cue PRWL treatment in models 1 and 2, we see that the rates of learning for both high and low performers are statistically insignificant. Although insignificant, we do see that the time trend for low performers is positive while it is negative for high performers. Overall, we do not find the performance gap between high and low performers to change over time in the dual cue PRWL treatment.

In the dual cue T treatment, there is significant learning for high performers but not for low performers across rounds 6 to 20. This suggests that bifurcation occurs, where the rate of learning is greater amongst high performers than for low performers. The difference in learning, however, does not appear to be statistically significant according to a Wald test (p = 0.278). Across rounds 11 to 20 in the dual cue T treatment in model 4, we continue to see significant learning for high performers while no learning for low performers.

Overall, there are no consistent signs of learning in the PRWL treatment in both single and dual cue tasks for both high and low performers. Accordingly, the performance gap between high and low performers remains stagnant across the post-intervention rounds. It appears that feedback on relative performance itself has no effect on learning for both high and low performers. The disutility associated with losing does not seem to encourage low performers to improve their performance over time, nor does it seem to provide the impetus for them to reduce it.[47] This is not to say that relative feedback has no effect whatsoever, since we previously found that the introduction of relative feedback lowered the forecast errors of low performers while having no effect on high performers. However, this effect appears to be one-off. A possible interpretation

[47] This could potentially be explained by the piece rates inherent in the PRWL treatment. Since players are motivated by money, amongst other sources of motivation, they would not want to reduce their performance for it would lead to low monetary earnings.

is that relative performance feedback is not salient enough to continuously motivate performance over time, hence we find neither bifurcation nor catching up.

In tournaments, where payoffs are tied to winning and losing, we observe both bifurcation and catching up. Compared to the PRWL treatment, it appears the additional saliency brought about by rank payoffs are important. In the single cue task, learning occurs at a faster rate amongst low performers compared to high performers. As a result, forecast errors between high and low performers diminishes over time. In the dual cue task the opposite occurs. Bifurcation takes place, where we observe learning for high performers but not for low performers.

The different effects according to task difficulty in the T treatment can be explained by the following. While low performers strive to win, it is much easier for low performers to do so by trying harder and improving their forecasts in the single cue task rather than in the dual cue task. The results support this, with learning amongst low performers in the single cue task but not for low performers in the dual cue task. For high performers, it appears that there is less scope for them to further improve in the single cue task compared to the dual cue task. The pattern of learning for high performers across tasks lends support to this.

Between-Subject Forecast Error Variability

We have touched upon the bifurcation and catching up effects by looking at how high and low performers learn relative to each other in the PRWL and T treatments. In this section, we study bifurcation and catching up at an aggregated level with the between-subject standard of forecast errors, without having to identify groups of high and low performers.

The between-subject standard deviation of forecast errors is defined in the spirit of Falk and Ichino (2006), where they study peer effects. Since participants only ever play against other participants in the same session, we calculate this variable at the session level, so it yields a single observation per round for each session of each treatment. It necessary means that we will be working with few observations, so analyses stemming from this will have low statistical power. Nevertheless, it gives us a better understanding of bifurcation and catching up. If the between-subjects standard deviation falls, there is lower dispersion of forecast errors between participants – indicating the convergence of forecast errors.

The between-subjects forecast error standard deviation observations are regressed in a basic ordinary least squares model against a linear time trend, over rounds 6 to 20 and over rounds 11 to 20. The regressions are run separately for each session. The estimated trends are presented in Table 6.9, with the constant term in each regression suppressed for presentation purposes. Each cell shows the estimated time trends in each session of the respective treatment, run separately in different regressions.

In Table 6.9 across rounds 6 to 20, the between-subjects standard deviations do not show any significant trend in any of the PRWL and T sessions in either single or dual cue tasks. In the single cue PRWL treatment, one session has a positive trend while the other has a negative trend. Both sessions of the dual cue PRWL treatment have a trend that slopes upwards. By comparison, all sessions of the T treatment – in both single cue and dual cue tasks – are fitted with a negative trend across rounds 6 to 20. Although insignificant, this points towards catching up in the T treatment, where between-subject variability diminishes.

It is perhaps unsurprising that the estimated trends are not statistically significant given the small number of observations used to estimate them: 15 observations over rounds 6 to 20 for each session. However, if a trend is significant, then it means that the effect is especially salient.

Over rounds 11 to 20, with even fewer observations to work with, we see some evidence of conformity. With the trend lines starting from a different base, we now see the between-subjects standard deviation falling significantly over time in both sessions of the single cue T treatment. In the single cue T treatment, forecast errors of participants in each session converge to that of other participants, with between-subjects standard deviation reducing by 1.166 and 1.608 forecast error points over each round in the respective sessions. The trends for both sessions of the dual cue T treatment also have a negative sign, but are insignificant.

Table 6.9 Time Trend of Between-Subject Standard Deviation of Forecast Errors

Dep Var: Between Subjects Std Dev of Forecast Errors	Rounds 6-20		Rounds 11-20	
	Session 1	Session 2	Session 1	Session 2
SC PRWL	1.054 (1.153)	-0.223 (0.387)	1.854 (2.388)	-1.033 (0.957)
SC T	-0.372 (0.462)	-0.033 (0.353)	-1.166 ** (0.462)	-1.608 * (0.730)
DC PRWL	0.244 (0.686)	0.281 (0.587)	-0.209 (1.549)	0.736 (1.079)
DC T	-0.540 (1.112)	-0.401 (0.400)	-0.338 (1.699)	-0.894 (0.786)
Observations	15	15	10	10

Between subjects standard deviation of forecast errors in each session regressed with a time trend. Each coefficient is estimated under a separate regression. The constant term in each regression is suppressed. OLS regression with robust standard errors.

Margin of Winning

The third analytical method which we use to study bifurcation and catching up is the margin of winning in the PRWL and T treatments – the treatments that feature competition. We look at how the margin of winning changes across time in these treatments. It should be noted that participants only learn of whether they win or lose, rather than the margin by which they win or lose. Despite this, the winning margin nevertheless serves as a latent measure of performance spread. If forecast errors converge, then we would expect the margin of winning to shrink. If this does indeed happen, we not only take this as stronger evidence to support catching up, but also evidence to support a stronger notion of it, given that the margin of winning is not actually observed by players themselves.

We calculate the winning margin as the absolute difference between the forecast errors of winners and losers in each matched pair in the PRWL and T treatments. Since winning is zero sum, the margin of winning is equivalent to the margin which the loser loses by. There is a unique observation for each randomly rematched pair for each of the 15 post-intervention rounds in the PRWL and T treatments.

Table 6.10 Margin of Winning in PRWL and T Treatments

Dep Var: Margin of Winning	Pooled		Single Cue		Dual Cue	
	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20	Rds 6-20	Rds 11-20
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
PRWL	(base)	(base)	(base)	(base)	(base)	(base)
T	9.855 * (5.145)	20.30 * (12.09)	3.507 (4.318)	11.11 (12.11)	16.67 * (8.98)	34.89 (21.27)
PRWL*Round	0.215 (0.251)	0.047 (0.601)	0.064 (0.316)	-0.485 (0.673)	0.454 (0.391)	0.812 (1.02)
T*Round	-0.207 (0.283)	-1.006 ** (0.503)	-0.142 (0.162)	-1.155 *** (0.406)	-0.231 (0.527)	-0.987 (0.912)
Constant	13.14 *** (3.130)	16.03 * (8.990)	9.087 *** (3.507)	18.32 * (9.549)	17.24 *** (5.300)	11.55 (15.96)
Observations	1155	770	615	410	540	360
Pairs	77	77	41	41	36	36
R^2	0.008	0.011	0.001	0.015	0.016	0.016
Wald χ^2	6.31	6.80	1.18	12.19	9.25	6.92
$p > \chi^2$	0.097	0.078	0.758	0.007	0.026	0.075
PRWL*Round = T*Round	χ^2 (1)= 1.24 p = 0.265	χ^2 (1)= 1.80 p = 0.179	χ^2 (1)= 0.34 p = 0.561	χ^2 (1)= 0.73 p = 0.394	χ^2 (1)= 1.09 p = 0.297	χ^2 (1)= 1.73 p = 0.188

The margin of winning is the abs difference between forecast errors of winners and forecast errors of their matched partners. Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared test are presented at the bottom of the table.

With the PRWL treatment serving as the reference category, Table 6.10 regresses the winning margin against a dummy for the T treatment, and time trends represented by PRWL*Round and T*Round representing how the margin of winning changes over time in each of these treatments. A positive trend indicates bifurcation where the margin widens, while a negative trend indicates catching up. Regressions are run over rounds 6 to 20 and also across rounds 11 to 20 for the pooled, single and dual cue data respectively.

In every regression model in Table 6.10, the margin of winning in the T treatment is wider than in the PRWL treatment. This is shown by the significant T dummy in models 1, 2 and 5.

In each of the regression models, the PRWL*Round coefficients are insignificant, suggesting that the winning margin in the PRWL treatment does not change over time. The coefficient on the T*Round interaction is negative in every regression model and is statistically significant in the pooled and single cue regressions across rounds 11 to 20, as shown in models 2 and 4. It shows that the winning margin reduces by approximately a single forecast error point every round in the T treatment, showing convergence of forecast errors between winners and losers. Wald tests, however, do not show any differences in how the margin of winning changes between treatments.

In summary, there is evidence that shows that forecast errors of participants converge to one another in the T treatment in the single cue task. Under each of the three methods to study bifurcation and catching-up – trends for high and low performers, forecast error variation, and the margin of winning – the performance gap between high and low performers diminishes over time in the single cue T treatment. We state this as Result 6.3 below:

Result 6.3. *The performance of low performers catches up to high performers in the single cue T treatment.*

In the dual cue T treatment, while the analyses of within-subjects standard deviation and winning margins also point in the same direction as in the single cue task towards catching up, there is no statistical evidence to support this. This differs slightly from what we found earlier looking at the rates of learning of high and low performers, where we found evidence of bifurcation in the dual cue T treatment. The difference in findings for the dual cue T treatment is entirely attributable to the different methodology involved in analysis. However, the fact that catching up in the single cue treatment is re-affirmed under different analytical methodologies points towards its robustness.

There are no signs of bifurcation nor catching up in the PRWL treatment. Taken together, it suggests that these effects comes about not through the provision of relative feedback per se, but rather from the rank-based payoffs that feature in tournaments.

6.5. Summary and Discussion

In this chapter, we made the distinction between high and low performers and conducted analyses along two main dimensions. We first compared, separately, the forecast errors of high and low performers across different treatments. High performers performed similarly to one another in each of the treatments, suggesting that treatment interventions have little effect on people of higher ability. Low performers, on the other hand, performed significantly better in the PRWL and S treatments compared to the PR and T treatments, with no differences in performance between the two. The findings for low performers mirror what we have found earlier in aggregate in Chapter 4, where we found that participants in the PRWL and S treatments had lower forecast errors than those in the PR and T treatments. In other words, the overall results are driven by low performers.

In terms of learning, earlier in Chapter 5 we found that learning is present only in the T treatment. We repeat similar analyses looking at how forecast errors change over time for high and low performers. There is evidence of learning in the T treatment for both high and low performers across rounds 6 to 20, as well as across rounds 11 to 20. We also observe learning amongst high performers in the PR treatment that was not present at the aggregate level when we analysed learning in Chapter 5.

While the first dimension of analyses focused on cross-treatment comparison within high and low performers, the second dimension took a different perspective and directly compared learning between high and low performers within treatments. In particular, we placed emphasis on performance spread by looking for signs of bifurcation and catching up. Bifurcation refers to the performance spread widening over time, while catching up refers to the spread narrowing.

In looking at performance spread, we do so in three ways: first, by comparing the forecast error trends for high and low performers; second, by seeing how between-subjects standard deviation of forecast errors change over time; and third, by seeing whether the margin of winning diminishes or widens over time.

First, by comparing the time trends of high and low performers, we find no learning for both high and low performers in the PRWL treatment. This suggests that the performance gap between high and low performers does not change over time. In the T treatment, both bifurcation and catching up effects are observed. Catching up occurs in the single cue T treatment, while bifurcation occurs in the dual cue T treatment.

Second, by tracking how between-subjects standard deviations for each round in each session changes over time, we observe catching up in both sessions of the T treatment in the single cue task. While the forecast error spread also seems to reduce over time in each of the dual cue T sessions, the trends are not statistically significant. We also do not find evidence of any changes in spread in any of the PRWL sessions.

The third and final analytical method looks at the difference in forecast errors between winning participants and their losing partners and how these change over time. We again find that the margin of winning reduces in the single cue T treatment. In the dual cue T treatment, and also in both single and dual cue PRWL treatments, the margin of winning does not change over time.

Each of these three distinct methods of analysing bifurcation and catching up effects point towards catching up in the single cue T treatment. There are mixed findings for the dual cue T treatment, depending on the analytical method. Neither bifurcation nor catching up was observed in the single and dual cue PRWL treatments.

The catching up effect that we observe in the single cue T treatment is particularly strong. First, because it persists even though random rematching discourages it. Second, we know the effect is particularly strong because participants in our study do not actually observe the performance of their partners – only of winning and losing – so low performers would not know, based on their experience derived from play, how much extra effort to exert to adequately improve their chances of winning. This additional source of uncertainty would be expected to discourage extra effort to be exerted by low performers. Yet we observe that the margin of winning diminishes in the single cue T treatment, even though such margin is unobservable by players. Third, since our task is cognitively challenging – even in the single cue task – higher effort does not necessarily improve forecast accuracy. So the fact that we do indeed observe better performance by low performers indicates a particularly large outlay of effort by them.

7. Results: Gender

Previously in our regressions, we controlled for participants' gender and found that it had a large influence on our results, where female participants did not perform as well as their male counterparts. From the baseline regressions in Table 4.4, the estimated coefficients on the gender dummy were larger in magnitude than the coefficients on the treatment dummies, suggesting that the gender of a participant has a larger impact on their performance than the treatment interventions. Here we look closer at the effect gender has on our results.

In this chapter, we break down gender differences in performance and try to pinpoint the reason why they occur. Is it due to inherent gender differences in participants' ability, or due to gender differences in treatment effects, which we had not previously allowed for? We distinguish these two effects by looking at whether performance differences exist in the first five rounds of play, and whether or not they carry over to the post-intervention rounds. For example, if pre-intervention gender differences do not exist in a particular treatment, but arise post-intervention, then these post-intervention gender differences must be due to differences in how treatment interventions affect participants of different gender.

7.1. Pre-Intervention Gender Differences in Performance

We first look for gender differences in performance across the first five rounds of play, the pre-intervention rounds. This allows us to ascertain whether or not there are any systematic differences in participants' ability that is associated with gender. If these differences in ability are present, we can mitigate its influence on our results by controlling for it.

By looking at overall gender differences in each of the single and dual cue tasks over the pre-intervention rounds, rounds 1 to 5, we see that women do not perform as well as men do. Table 7.1 presents basic regressions of forecast errors in the first five rounds of play against a single dummy variable representing women (men as the reference category) for each of the pooled, single and dual cue tasks. Here we do not distinguish between the different treatments. From these regressions, forecast errors are larger for women than for men in each of the pooled and single cue regressions, where the female gender dummy is highly significant at the 1% level. For the dual cue task in regression model 3, the gender dummy continues to be large and positive but is not statistically significant, with a p-value of 0.113.

Table 7.1 Pre-Intervention Gender Differences in Forecast Errors

Dep Var: Forecast Errors	Pooled	Single Cue	Dual Cue
	Model 1	Model 2	Model 3
Male	(base)	(base)	(base)
Female	8.045 *** (2.042)	4.757 *** (1.706)	5.044 (3.185)
Constant	17.95 *** (1.307)	11.24 *** (1.092)	28.79 *** (2.220)
Observations	1465	765	700
Participants	293	153	140
R^2	0.018	0.016	0.004
Wald χ^2	15.51	7.77	2.51
$p > \chi^2$	0.000	0.005	0.113

Regressions are estimated with Random Effects GLS over rounds 1 to 5. Standard errors are in parentheses and are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

It is not surprising that women do not perform as well as men. It is well documented that women underperform men in complex problem solving tasks, while the differences are much less apparent under simple arithmetic (Hyde et al., 1990). Women would therefore be expected to underperform men in our forecasting task, which is cognitively challenging.

Narrowing down by treatment, we observe that pre-intervention gender differences are not present in every treatment. Table 7.2 looks at gender differences in each of the treatments, where forecast errors across the first five rounds of play are regressed against gender-treatment interaction terms. With two categories representing gender (male and female) and four categories for treatment (PR, PRWL, T and S), the full interaction is represented by seven interaction terms plus the reference category (male PR).

Table 7.2 Regressions of Pre-Intervention Gender Differences in Forecast Errors, by Treatment

Dep Var: Forecast Errors	Pooled	Single Cue	Dual Cue
	Model 1	Model 2	Model 3
Male PR	(base)	(base)	(base)
Female PR	8.846 (5.397)	-0.494 (4.084)	25.39 * (7.931)
Male PRWL	-4.391 (3.334)	-3.685 (3.213)	-0.428 (5.635)
Female PRWL	1.686 (3.494)	0.906 (3.446)	2.826 (5.232)
Male T	-1.380 (4.113)	-2.142 (4.144)	6.301 (6.840)
Female T	4.284 (3.302)	-0.916 (3.490)	6.580 (4.075)
Male S	-6.826 *** (3.165)	-6.710 ** (2.902)	1.483 (5.331)
Female S	5.025 (3.293)	6.413 (4.251)	2.233 (4.321)
Constant	21.15 *** (2.313)	14.69 *** (2.782)	27.24 *** (3.029)
Observations	1465	765	700
Participants	293	153	140
R^2	0.025	0.034	0.018
Wald χ^2	25.10	32.99	6.51
$p > \chi^2$	0.001	0.000	0.482
Female PR = 0	$\chi^2(1) = 2.69$ p = 0.101	$\chi^2(1) = 0.01$ p = 0.904	**$\chi^2(1) = 3.76$** **p = 0.052**
Male PRWL = Female PRWL	**$\chi^2(1) = 2.93$** **p = 0.087**	**$\chi^2(1) = 3.14$** **p = 0.077**	$\chi^2(1) = 0.26$ p = 0.610
Male T = Female T	$\chi^2(1) = 1.87$ p = 0.171	$\chi^2(1) = 0.11$ p = 0.742	$\chi^2(1) = 0.00$ p = 0.967
Male S = Female S	**$\chi^2(1) = 13.83$** **p = 0.000**	**$\chi^2(1) = 15.62$** **p = 0.000**	$\chi^2(1) = 0.02$ p = 0.889

Regressions are estimated with Random Effects GLS over rounds 1 to 5. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

In Table 7.2, we see that the coefficients on the female interactions for each treatment are consistently larger than those for the corresponding male interaction terms, suggesting that women tend to perform worse across the first five rounds. However, these gender differences in performance are not always statistically significant. At the bottom of the regression table, we present a series of Wald chi-squared tests of gender differences for each treatment. In the pooled regression in model 1, women perform significantly worse than men in the PRWL and S treatments – and marginally misses significance in the PR treatment, with a p-value of 0.101. In the single cue task in model 2, we again observe pre-intervention gender differences in performance across the first five rounds of play in the PRWL and S treatments. In the dual cue task, pre-intervention gender differences only occur in the PR treatment.

It should be noted that we do not observe significant gender differences in the T treatment over the first five rounds of play in each of the pooled, single and dual cue regressions. We will come back to this finding shortly.

The observation that gender differences in performance occur in some treatments but not others can only be attributed to sampling variation. This is because treatments – at least the PR, PRWL and T treatments – are identical in the first five rounds of play. As such there is no other plausible explanation.

7.2. Post-Intervention Gender Differences in Performance

Having examined pre-intervention gender differences in performance, we now proceed to analysing gender performance differences across rounds 6 to 20, the post-intervention rounds. Table 7.3 present regressions of post-intervention gender differences when the single and dual cue treatments are pooled together. We opt to pool the data to improve statistical power, especially when we are disaggregating analyses at the gender level. From the regression table, model 1 regresses forecast errors against the seven gender-treatment interaction terms (with male PR serving as the reference category) and a linear time trend denoted by 'Round'. Regression model 2 builds on the previous regression by additionally controlling for participants' ability. Ability, as was defined earlier in Chapter 6, is the median forecast error for each participant across the first five rounds. As with forecast errors, higher values of this ability variable indicates lower ability. Regression model 3 interacts ability by gender. Panel B of Table 7.3 consists of three series of hypothesis tests. The first series tests for gender differences within each treatment. The second and third series of tests looks for treatment differences amongst male and female participants respectively.

In model 1 of the regressions in Table 7.3, we see that female participants perform worse than male participants across the post-intervention rounds in every treatment. The first set of Wald tests in Panel B of the table show that these gender differences are highly significant in each of the four treatments, with gender differences significant at the 5% level in the PR treatment, and at the 1% level in each of the other treatments.

In regression model 2 of Table 7.3, we additionally control for the ability of participants. Since gender differences in ability are present in some treatments, we control for the ability and see whether post-intervention gender differences persist or not. Given that there are post-intervention gender differences in all treatments, if gender differences are no longer significant when ability is controlled for in particular treatments, then the post-intervention gender differences in these treatments are merely artefacts from earlier rounds. On the other hand, if gender differences persist, then they cannot be explained by gender differences in initial ability. In this case, gender differences would be associated with gender differences in treatment effects, where the treatment interventions have different effects on participants of different gender.

In model 2 of the regressions, we no longer find significant gender differences in the PR, PRWL and S treatments after ability has been controlled for – with ability being highly significant. Given that gender differences in ability have been observed in these treatments, the finding that post-intervention differences dissipate when ability is controlled for suggests that post-intervention gender differences in these treatments are carried over from the pre-intervention rounds.

Table 7.3 Regressions of Post-Intervention Gender Differences in Forecast Errors

Panel A: Regression Results

Dep Var: Forecast Errors	Model 1	Model 2	Model 3
Male PR	(base)	(base)	(base)
Female PR	7.820 ** (3.856)	2.678 (2.474)	1.198 (2.688)
Male PRWL	-2.884 (2.056)	-0.689 (1.687)	-0.881 (1.626)
Female PRWL	3.509 (2.617)	2.096 (2.093)	0.775 (2.315)
Male T	-1.373 (2.544)	-1.382 (1.744)	-1.382 (1.709)
Female T	10.10 *** (3.656)	7.095 ** (3.214)	5.706 * (3.199)
Male S	-2.480 (2.398)	0.199 (1.759)	-0.036 (1.741)
Female S	4.425 * (2.400)	3.299 (2.104)	1.990 (2.361)
Round	-0.126 * (0.066)	-0.126 * (0.066)	-0.126 * (0.066)
Ability		0.624 *** (0.055)	
Male Ability			0.569 *** (0.067)
Female Ability			0.651 *** (0.068)
Constant	16.78 *** (1.910)	7.117 *** (1.774)	7.966 *** (1.597)
Observations	4395	4395	4395
Participants	293	293	293
R^2	0.033	0.154	0.155
Wald χ^2	46.80	193.8	245.1
$p > \chi^2$	0.000	0.000	0.000

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively.

Panel B: Wald Chi-Squared Tests of Hypotheses

	Model 1	Model 2	Model 3
Males vs Females			
Female PR = 0	**$\chi^2(1) = 4.11$** **p = 0.043**	$\chi^2(1) = 1.17$ p = 0.279	$\chi^2(1) = 0.20$ p = 0.656
Male PRWL = Female PRWL	**$\chi^2(1) = 7.78$** **p = 0.005**	$\chi^2(1) = 2.37$ p = 0.124	$\chi^2(1) = 0.56$ p = 0.454
Male T = Female T	**$\chi^2(1) = 9.44$** **p = 0.002**	**$\chi^2(1) = 7.72$** **p = 0.005**	**$\chi^2(1) = 5.07$** **p = 0.024**
Male S = Female S	**$\chi^2(1) = 8.38$** **p = 0.004**	$\chi^2(1) = 2.65$ p = 0.104	$\chi^2(1) = 0.77$ p = 0.379
Male Ability = Female Ability			$\chi^2(1) = 0.73$ p = 0.393
Treatment Effects: Male			
Male PRWL = 0	$\chi^2(1) = 1.97$ p = 0.161	$\chi^2(1) = 0.17$ p = 0.683	$\chi^2(1) = 0.29$ p = 0.588
Male PRWL = Male T	$\chi^2(1) = 0.47$ p = 0.494	$\chi^2(1) = 0.25$ p = 0.614	$\chi^2(1) = 0.14$ p = 0.705
Male T = 0	$\chi^2(1) = 0.29$ p = 0.589	$\chi^2(1) = 0.63$ p = 0.428	$\chi^2(1) = 0.65$ p = 0.419
Male S = 0	$\chi^2(1) = 1.07$ p = 0.301	$\chi^2(1) = 0.01$ p = 0.910	$\chi^2(1) = 0.00$ p = 0.983
Male T = Male S	$\chi^2(1) = 0.19$ p = 0.662	$\chi^2(1) = 1.17$ p = 0.280	$\chi^2(1) = 0.84$ p = 0.360
Treatment Effects: Female			
Female PR = Female PRWL	$\chi^2(1) = 1.17$ p = 0.280	$\chi^2(1) = 0.05$ p = 0.815	$\chi^2(1) = 0.03$ p = 0.866
Female PRWL = Female T	**$\chi^2(1) = 3.04$** **p = 0.082**	$\chi^2(1) = 2.36$ p = 0.124	$\chi^2(1) = 2.32$ p = 0.128
Female PR = Female T	$\chi^2(1) = 0.23$ p = 0.629	$\chi^2(1) = 1.56$ p = 0.211	$\chi^2(1) = 1.62$ p = 0.203
Female PR = Female S	$\chi^2(1) = 0.78$ p = 0.378	$\chi^2(1) = 0.06$ p = 0.803	$\chi^2(1) = 0.10$ p = 0.751
Female T = Female S	$\chi^2(1) = 2.44$ p = 0.119	$\chi^2(1) = 1.36$ p = 0.243	$\chi^2(1) = 1.31$ p = 0.252

Bold typeface indicates statistical significance at the 10% level or better.

In the T treatment, we continue to see men performing significantly better than women. From Table 7.2, there were no significant gender differences in the pre-intervention rounds. As such, we continue to observe post-intervention gender differences in performance irrespective of whether ability has been controlled for or not.

Given that we had not found any gender differences in performance in the first five rounds of play in the T treatment, it is interesting that we find such differences in the post-intervention rounds. This is explained by gender differences in how players respond to the tournament pay scheme introduced after round 5. Women perform significantly worse under tournaments than men, and is not attributed to gender differences in forecasting ability. At a later point, we investigate whether gender differences in competitiveness could be driving this result.

Regression model 3 of the regressions in Table 7.3 show identical results to those in model 2. Model 3 is similar to model 2, but instead interacts players' ability by gender. This regression specification allows ability to affect the performance of male and female participants in different ways, even if the ability of male and female participants is identical. We find that both male and female ability interactions are highly significant, but are not different from one another. Accordingly, the pattern of results do not differ much between regression models 2 and 3.

Our main finding is summarised as Result 7.1 below:

Result 7.1. *Post-intervention gender differences in performance occur in all treatments. Other than for the T treatment, these gender differences can be explained by gender differences in ability.*

It is interesting to point out a general pattern in the results that we find. In treatments where there were initial gender differences in ability, and where post-intervention gender differences exist, these post-intervention differences are explained solely by gender differences in ability. In addition to this, in the T treatment, we observe post-intervention gender differences even though they were not present pre-intervention. These differences are associated with gender differences in how men and women respond to the treatment interventions, namely interventions associated with competition.

7.3. Gender Differences in Treatment Effects

In the previous section, we looked at gender differences in performance in each of the treatments and found that post-intervention gender differences in some treatments were associated with gender differences in how participants responded to these treatment interventions. We now proceed by comparing forecast errors across treatments, for men and women separately. We again refer back to the estimated gender-treatment interacted dummies from Table 7.3, but now we compare dummies across treatments for each gender. The appropriate hypothesis tests are presented in Panel B of the tables.

From the regressions in Table 7.3, there are no differences in the performance of male participants across treatments in each of the regression models. This suggests that men are unaffected by the treatment interventions that take place.

Earlier we found that women in the T treatment performed worse than their male counterparts. From model 1 of Table 7.3, we observe that women in the T treatment tend to perform worse than women in other treatments. A Wald test shows that female T participants perform significantly worse than female PRWL participants, with a p-value of 0.082. This level of significance, however, drops once players' ability is controlled for in models 2 and 3, with the p-value of 0.124 in model 2.

7.4. Why Women Perform Worse in the T Treatment

We have found that women perform significantly worse than men in the T treatment across the post-intervention rounds. These gender differences are not attributable to initial differences in players' ability, but rather to gender differences that relate to the treatment intervention. In this section of the chapter, we focus on the competition aspect of the T treatment and investigate possible reasons why women perform worse than men. In particular, we investigate whether gender differences in competitiveness are driving gender differences in performance.

A reason why women do not perform as well as men in the T treatment is because women are typically less competitive than men. In an experiment where participants are given the choice to be remunerated by piece rates or by a tournament-based scheme, Niederle and Vesterlund (2007) found that 73% of men selected tournaments, while only 35% of women did so. In our experiment, although participants do not face the decision of tournament entry, if women do indeed shy away from competition, they may do so by exerting less effort when they are exogenously assigned to a tournament – as in our T treatment. As men embrace competition, we would expect them to exert higher effort than women.

While tournament entry decisions do not feature in our experiment, we proxy participants' competitiveness by their trait anxiety, which we had elicited from the pre-task questionnaire. Trait anxiety measures how prone participants are to stress and situations that make them anxious. According to Segal and Weinberg (1984), trait anxiety is correlated with competitiveness such that lower levels of competitiveness are represented by higher levels of trait anxiety. They also find that women report higher levels of trait anxiety than men do.

In analysing the effect that competitiveness has on people's performance in the T treatment, we once again draw on the ex-post distinction of 'winners' and 'losers' that we first introduced in Section 4.3.1. Winners are defined to be participants who have won more than half of the post-intervention rounds against their random partner – in other words, those who have won eight or more of the fifteen post-intervention rounds. On the other hand, losers have lost more rounds than they have won.

Table 7.4 Classification of Winners and Losers by Gender

	Winners		Losers	
	Male	Female	Male	Female
Pooled T	20	18	12	25
SC T	12	8	9	10
DC T	8	10	3	15

Count of 'winners' and 'losers' by gender in the T treatment. Winners are participants who have won 8 or more of the 15 post-intervention rounds, while losers have won 7 or less. The sum of all male and female winners and losers roughly sum to the number of participants in each treatment. There are discrepancies because a small number of participants did not provide gender information in the post-task questionnaire.

Table 7.4 shows how male and female participants in the T treatment are classified as winners and losers. Consistent with the gender differences that we found earlier in the chapter, we find clear differences in how winners and losers are classified. There tend to be a smaller incidence of female winners than male winners, and a greater incidence of female losers than male losers. A series of Probit regressions, which we do not present, show that women are less likely than men to be winners.

We investigate the effect of competitiveness on winners' and losers' performance in Table 7.5. The table presents three regression specifications for both winners and losers. In the first specification, we regress forecast errors against the gender dummy (male as the reference category). The second specification additionally controls for trait anxiety, allowing us to assess the effect of competitiveness. The third specification regresses against gender as well as the gender interactions of trait anxiety, allowing trait anxiety to have different effects according to participants' gender. We run these regressions separately for winners and losers, so regression models 1 to 3 in Table 7.5 are for winners in the T treatment, while models 4 to 6 are for losers. As we did before, we pool the single and dual cue observations for each treatment to improve statistical power.

Table 7.5 Regressions of Winners' and Losers' Forecast Errors against Trait Anxiety

Dep Var: Forecast Errors	Winners			Losers		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Male	(base)	(base)	(base)	(base)	(base)	(base)
Female	4.506 (3.205)	2.549 (3.057)	-36.13 * (20.62)	14.45 ** (6.212)	13.15 ** (6.035)	9.655 (31.71)
Trait Anxiety		0.399 * (0.223)			0.348 (0.439)	
Male × Tr Anx			-0.281 (0.428)			0.299 (0.334)
Female × Tr Anx			0.618 *** (0.230)			0.387 (0.756)
Constant	11.62 *** (1.818)	-4.348 (9.604)	24.48 (18.61)	17.38 *** (3.872)	4.625 (18.20)	6.525 (15.13)
Observations	570	525	525	555	540	540
Participants	38	35	35	37	36	36
R^2	0.013	0.026	0.042	0.035	0.034	0.034
Wald χ^2	1.98	3.43	7.70	5.41	4.81	5.04
$p > \chi^2$	0.160	0.180	0.053	0.020	0.090	0.169
Male × Tr Anx = Female × Tr Anx			**χ^2 = 3.42** **p = 0.065**			χ^2 = 0.01 p = 0.915

Regressions are estimated with Random Effects GLS over rounds 6 to 20. Standard errors in parentheses are clustered at the participant level. *, ** and *** represents the 10%, 5% and 1% level of significance respectively. Wald chi-squared tests are presented at the bottom of the table, with bold typeface indicating significance at the 10% level or better.

In model 1 of Table 7.5, we find that female winners tend to have larger forecast errors than for male winners, although the difference is not statistically significant. In model 2, we find that higher trait anxiety significantly worsens the performance of T winners. When we distinguish the effects for male and female winners, we find that trait anxiety affects female players but not male players. The Wald test accordingly shows that trait anxiety affects male and female winners differently.

From model 3 of Table 7.5, the significant effect of trait anxiety for female T winners confirms competiveness as one of the factors that could be driving gender performance differences in the T treatment. While men are not affected by their self-reported levels of trait anxiety, women perform worse as their trait anxiety levels increase. Since higher levels of trait anxiety is associated with lower competitiveness, our findings here show that the performance of women is lower if they are less competitive, or conversely, that the performance of women is higher as they are more competitive. In fact, from model 3 of Table 7.5, the trait anxiety effect is so salient that once it is interacted by gender, the Female dummy is now negative and significant – indicating that female winners in the T treatment actually outperform male winners once the effect of competitiveness is controlled for.

In models 4 to 6, we focus on gender differences for losers in the T treatment. Regression models 4 and 5 show women performing worse than men. The gender difference is no longer significant when gender-interacted trait anxiety terms are included in model 6 – although both trait anxiety terms are insignificant. Overall, it appears that competitiveness has little effect on the performance of losers.

In review, we find evidence that gender differences in competitiveness are driving gender differences in performance for certain people. For winners in the T treatment, competitiveness has no effect for men, while significantly affects the performance of women, where lower levels of competitiveness leads to lower performance.

Result 7.2. *Gender differences in how competitiveness affects players can, at least in part, explain the post-intervention gender differences in the T treatment.*

Competitiveness, however, cannot fully explain the gender differences that we observe in the T treatment. The elicited trait anxiety scores, which we use to proxy competitiveness, does not seem to affect T losers.

7.5. Summary and Discussion

This chapter has found that there exist post-intervention gender differences in performance in all treatments, when the data is pooled. In most treatments, these post-intervention gender differences are accounted for by differences in ability. However in the T treatment, post-intervention gender differences in performance cannot be explained by ability. Rather, these differences are attributed to the differential effects the T treatment intervention have on participants of different gender.

Competition is a central element in the T treatment. As such, we ask whether gender differences associated with competitiveness could explain the post-intervention gender performance differences that we observe in the T treatment. With participant's self-reported trait anxiety scores as a proxy of their underlying competitiveness, we find that competitiveness can partially explain the gender performance differences that we observe.

8. Conclusion

8.1. Summary

This thesis studied the effect payoffs and feedback has on the performance of experimental workers in a cognitively challenging forecasting task. Our three initial research questions were: 1) which of the three pay schemes of piece rates, tournaments and salaries bring about the best performance from workers; 2) what role do relative performance feedback and rank-dependent payoffs have on the performance of tournaments; and 3) and how do the different pay schemes and relative performance feedback impact on learning?

These three research questions were addressed in an experiment with a real-effort forecasting task, where participants were asked to predict the underlying price of a hypothetical stock based on the observation of two numerical cue values for each of twenty rounds. There are two versions of the task: the single cue task has one of the cue values fixed at a particular value in every rounds, while both cue values change in the dual cue task. Forecast errors, the absolute difference between the predicted value and the actual underlying stock price, are our primary measure of performance in this task, where they indicate the accuracy of the prediction.

Our experimental treatments vary along two dimensions of pay schemes and feedback. We study three different pay schemes: piece rates, winner-takes-all tournaments and salaries. Piece rates pay players more for better individual performance. Tournaments pay a fixed prize to the winner of a matched pair, while the loser receives nothing. Salaries pay players a pre-announced fixed payment irrespective of their performance.

Treatments also differ in terms of the feedback that is provided to players. In the default feedback protocol, players only observe their forecast errors at the end of every round, allowing them to gauge their individual performance. The other feedback protocol provides additional information about whether players have performed better or worse than a random partner. This feedback is context-loaded, informing players whether or not they have 'won' or 'lost'. This relative feedback simulates competition between players.

Our treatments are as follows. The Piece Rate (PR) treatment pays participants a piece rate on their forecast errors and does not provide relative performance feedback to them. The Piece

Rate Win Lose treatment (PRWL), pays participants piece rates while providing such feedback. This feedback has no effect on the piece rate payoffs in the PRWL treatment. The PR and PRWL treatments differ only in terms of whether relative feedback is provided or not. In the Tournament (T) treatment, players are randomly matched with a partner, and is subsequently rematched every round. The player who has the smaller forecast error of the two 'wins' while their partner 'loses'. The winner receives $1 for winning, while the loser receives nothing.[48] Feedback is identical in the PRWL and T treatments, but differ in terms of pay scheme. The Salary (S) treatment pays participants a pre-announced amount of $20, so payoffs do not depend on performance.

Our results show that in terms of pay schemes, salaries perform particularly well – with forecast errors lower in the S treatment than in both the PR and T treatments. This finding is consistent with Cognitive Evaluation Theory (Deci & Ryan, 1985). We also find that tournaments perform similarly to piece rates, commensurate to the Piece Rate Equivalence property of tournaments (Lazear & Rosen, 1981).

Decomposing tournaments, we find that the mere act of competing motivates performance, where the PRWL treatment performs better than the PR treatment. Controlling for this effect and looking solely in terms of payoffs, we also find that tournaments no longer perform as well as piece rates. The PRWL treatment performs better than the T treatment, suggesting that the rank-dependent payoffs are not as effective as piece rates in eliciting effort. This suggests that competition plays a crucial role in motivating performance under tournaments, where it appears that it is driving Piece Rate Equivalence.

Our results show that, overall, the PRWL and S treatments perform particularly well compared to the PR and T treatments. When we disaggregate our results by ability, we find that most of these effects are borne out by low performers. Forecast errors of high performers do not differ across treatments.

[48] We also have the Tournament-No-Info (TNI) treatment, whereby players engage in tournament play as in the T treatment, but have relative feedback withheld from them. During play, they are unaware of how they are performing relative to their partners. Relative feedback is provided only at the end of the game. We only have data for the TNI treatment for the dual cue task.

When we turn our attention to performance dynamics, we find that learning is only present in the T treatment. This is represented by both improved forecast accuracy and forecast consistency over time. The learning under tournaments is attributed to the stark differences in payoffs between winning and losing, which provide a strong impetus for players to improve their performance, in turn improving their chances of winning. In the single cue T treatment, the rate of learning is significantly higher amongst low performers than high performers, where we observe significant catching up and convergence of performance.

In contrast, while the provision of relative feedback motivates performance, it does not affect players' learning. Winning improves players' perceptions of competency, but has no influence on their future performance. The PRWL treatment does not exhibit any learning.

When analysing effects by gender, we find that women perform significantly worse than men in every treatment. In most treatments, these post-intervention differences in performance can be explained by differences in participants' ability. However, this is not the case for the T treatment, where post-intervention gender differences in performance are observed while the corresponding differences in ability are absent. This gender performance difference in the T treatment is attributed to women reacting adversely to the T treatment intervention. Such gender differences in performance in the T treatment can, in part, be explained by female participants 'shying away' from competition by reducing their performance.

8.2. Implications for Productivity

Our study has important implications for the productivity of workers. Based on our findings, there are readily available tools for managers to improve worker productivity. The easiest way to do so is by introducing an element of competition. This competition can be simulated by feedback on how well a worker is performing compared to his peers. As such, it is a relatively cost-effective way to improve worker performance. The motivating effect associated with relative feedback kicks in immediately and remains constant over time. This feedback, however, does not spur learning – meaning that it is well suited for short term positions, or positions with high turnover. It is also suited to jobs where the task is mechanical and there is little scope for learning and long term improvement.

Tournaments are more appropriate if learning is particularly valued. We found that although tournaments did not perform well overall, it is associated with a high rate of learning. In long tenured positions, tournaments will enable workers to be most productive. If the task at hand is not overly difficult, tournaments are especially effective in motivating low performing workers to learn and perform. There is a caveat however. While men perform particularly well under tournaments, women underperform.

Aside from relative performance feedback, productivity can be improved by paying workers a fixed salary instead of performance pay schemes. Neither piece rates nor tournaments perform as well as a flat salary. However, like the provision of feedback, the motivating effects of salaries are not dynamic.

8.3. Limitations and Directions for Future Research

Our study utilised a laboratory experiment to study the effect of pay schemes and feedback on the productivity of experimental workers. The design of the experiment restricts the scope of our analyses and affects how the findings are interpreted. Here we reflect on our experiment and discuss things that – with the benefit of hindsight – we might have done differently. These also indicate areas where future research could be directed.

One limitation of our study is the number of participants that took part in our experiment. In places, results pointed in the direction that we would expect, but were nevertheless insignificant at conventional levels. In some instances, these coefficients were large in magnitude but were also accompanied by large standard errors. A larger number of observations would have reduced these standard errors, making our results more salient.

Another limitation of our study is that our focus is on final performance, rather than effort. There is therefore an inherent disconnect with theoretical models, where agents are modelled to exert effort and effort in turn leads to performance via a production function. It is empirically possible that our interventions increase effort but this does not translate into higher performance. Upon reflection, we could potentially proxy effort by measuring the duration of time a participant takes before they make their forecast. This amount of time can be thought of as the time which participants have devoted towards thinking and processing the available information to them, with the goal of making an accurate forecast. To our knowledge, there are few studies that study the effect of incentives and feedback on both effort and performance.

A significant part of our analyses was dedicated to learning. Our experiment consisted of 20 rounds, for which 15 studied the effects associated with the treatment interventions. It would have been appropriate to study learning over a longer time horizon. A longer time horizon would also provide a greater number of observations to work with.

Appendix 1. Participant Recruitment Message

To: Students at the University of Auckland

My name is Tony So. I am a Ph.D. student in the Department of Economics. I am writing to invite you to take part in a research project that I am currently working on. In order to recruit participants for my study I am approaching students at the University and I am making this announcement in a number of different courses with the permission of the course instructor. The findings of this study will be published in scholarly journals at a date in the future.

The University of Auckland is providing the funds for this study. I have received approval from the University of Auckland Human Participants Ethics Committee to undertake this project.

Your participation in this study is completely voluntary and you are free to withdraw at any point during the study if you wish to do so. You do not have to provide a reason for your withdrawal and you will not incur any penalty for doing so.

Your participation in the study will not have any effect on your grade for the course. You do not need any discipline-specific knowledge in order to take part.

My research looks at individual decision making under conditions of uncertainty. I will provide you with more detailed instructions and explain the task that you will be expected to perform if you sign up to take part.

There will be a financial remuneration for your participation. You will get $5.00 just for showing up but in order to get this you must arrive on time. You can expect to make around $20.00 in total for participating in one session. The actual amount will vary based on the decisions you make in the experiment. However you cannot lose money and everyone will make a positive sum of money. On average participants will make around $20.00. The money that you make in the session is private information and will not be revealed to any other participant. We encourage you to not reveal this information to any other participant.

Each of you will take part in only one session of the study which will last about 90 minutes. The experiments will be run in a University Computer Lab (listed below). You will be in a group with other participants. Once all participants signed up for a session have assembled, I will read you the instructions describing what you have to do. After the instructions are read to you, and all questions answered, I will ask you to make a decision for a number of rounds.

Typically this decision will involve picking a number for each round and entering that number using the computer keyboard. Based on the number you pick and the numbers picked by others in your group, you will earn a certain amount of money. After a certain number of rounds the session will come to an end and you will be paid your earnings in the experiment in cash.

There is also no physical or psychological risk or discomfort involved. Participants are free to make decisions at their own pace with no pressure on your time. There is no audio or videotaping involved.

Only the researchers involved will have access to the data collected during these experiments. They will be kept in a locked filing cabinet or as password protected computer files in my office to which only we have the keys. The data is collected confidentially. You will never be identified by your name. Every participant in the experiment will be assigned an ID number and all the data will be filed using that ID number. There is no way to connect you personally to the decisions you made in the experiment. All the computer files will be permanently deleted after six years.

Currently we are recruiting participants for the following dates.

[a list of session dates]

All sessions will be held at the CISCO Lab (Lab 05) on Level Zero of the Owen G Glenn Building (the business school building), located at 12 Grafton Road.

All sessions will start at 4:00 PM and last for about 90 minutes.

NOTE: Please sign up for ONLY ONE of the following sessions! HOWEVER THESE ARE NEW EXPERIMENTS AND YOU CAN SIGN UP FOR THESE EVEN IF YOU HAVE PARTICIPATED IN AN EXPERIMENT THIS SEMESTER OR THIS YEAR.

Please click on the following like to sign-up.

[sign up link]

(You should be able to click or "Ctrl+Click" on this link to get to the relevant page. If that does not work then cut and paste the URL into your browser window. You will need your UPI and NetAccount/Cecil password to access this page.)

Enter your first and last name, email address and choose ONE of the dates from the drop-down list. ***Please do NOT sign up for more than one session.*** You will be sent a reminder to the email address you provide a day before the session you have signed up for.

If you think you need more information before you can participate then please feel free to contact me at *[contact number]* or e-mail me *[contact email]*.

The Head of the Department of Economics is: Professor Basil Sharp, Department of Economics, The University of Auckland, Private Bag 92019, Auckland. Telephone: *[contact number]*, e-mail: *[contact email]*

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN SUBJECTS ETHICS COMMITTEE on 28/02/2011, for a period of 3 years, from 28/02/2011, Reference 2011/007

Appendix 2. Instructions for Forecasting Game

A2.1. General Instructions

The University of Auckland

Instructions for the Experiment

WELCOME.
PLEASE TURN YOUR CELL PHONES OFF NOW

This is a study examining the manner in which people make decisions. The University of Auckland has provided the funds to conduct this research. If you follow the instructions and make good decisions you might earn a considerable amount of money.

At the beginning of the session each person will be given an Earnings Account with $5.00 in it. You will participate in a decision making task for each of 20 rounds. You will have the chance to earn money each round, with your earnings for each round being added to your Earnings Account. At the end of the experiment, the balance of your Earnings Account will be paid to you in cash.

[In the Salary treatment, the previous paragraph was replaced by the following:

At the beginning of the session each person will be given an Earnings Account with $5.00 in it. You will participate in a decision making task for each of 20 rounds. You will be paid $20 for participating in the study. At the end of the experiment, the $20 will be paid to you in cash.]

DESCRIPTION OF THE TASK:

In each round you will be asked to predict the future value of a fictitious **'stock'**. The value of this stock is unknown to all participants, but you will be able to observe two CUES that can help you form your forecast. These cues can be used to predict the stock's value much the same way that the amount of rainfall and the average temperature can be used to predict the quality of a corn crop, the number of unoccupied apartments and student enrolment this year can be used to predict next year's rent increases, or the demand for sports cars can be used to predict their future price.

In each round you will be shown the values for the two CUES.

NOTE: One of the CUE values will always be fixed at 150 for each of the 20 rounds. The other cue value will change each round. But the relation of the cue values to the stock's price will remain the same.

Example:

For example let the value of Cue A is fixed at 150. Suppose the values for the cues in a round were given as:
CUE A = 150
CUEB = 100

You will be asked to predict the price of the stock given these two cue values.
The next round one of the cues will take on a different value, such as:
CUE A = 150
CUEB = 450

You will then predict that round's price using these new cue values. Remember that even though the values of the cues change, the underlying relation between the cue values and the stock's price remains the same. Thus, in order to make accurate forecasts you will need to determine the relation between the cues and the price of the stock.

YOUR FORECASTING ERROR

After making your forecast, the computer will calculate the distance between your forecast and that round's actual price (your absolute forecasting error). This amount will be your **forecast error**.

Example:

Suppose your forecast was 230. If the actual price of the stock was 200 then your forecast error would be 30:
Your forecast error = 230 – 200 = 30

Suppose your forecast was 148. If the actual price of the stock was 200 then your forecast error would be 52:

Your forecast error = 200 – 148 = 52

YOUR EARNINGS IN EACH ROUND:

In each round, your earnings will depend on your forecast error. Your earnings in each round will equal $1 **less** your forecast error for that round.

That is, your earnings (E) in each round will be given by E = $1.00 – (forecast error).

Example:

Suppose your forecast error in a particular round is 30. Then you will earn $0.70 in that round. This is because:
$1.00 – $0.30 = $0.70

Suppose in another round your forecast error is 8. Then you will earn $0.92 in that round. This is because:
$1.00 -- $0.08 = $0.92

Note that if your error is 100 or over, then you will earn nothing in that round. The minimum amount you can earn in a round is $0.00.

Suppose in another round your forecast error is 102. Then you will earn $0.00 because:
$1.00 --$1.00 = $0.00

[This passage was also included in the Salary Treatment:
Note that your actual earnings for the experiment will be $20.00, regardless of your forecasting accuracy.]

SPECIFIC INSTRUCTIONS:

Before Round 1:
You will be shown 10 examples of cues and stock prices. You will have 5 minutes in which to examine these examples.

Round 1:
At the end of the 5-minute example round you will be shown the first two cue values and asked to forecast the price of the stock in Round 1. You will have **90 seconds** to make your forecast.

End of Round 1:
At the end of the **90 seconds** all participants will have entered their forecasts. After all earnings have been calculated you will be shown your results for Round 1. The computer will then show you your earnings for the round, including:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this Round	Total Earnings

Please record this information on the RECORD SHEET provided to you.

Beginning of Round 2:
After examining and recording the earnings from round 1, you will be shown the values of CUE A and CUE B in Round 2. You will have **90 seconds** to form your forecast.

Subsequent Rounds:
Each subsequent round proceeds in the same way and will be repeated for each of the 20 rounds. In each round, you will make a forecast based on two new cue values. At the end of round 20, you will receive a cash payment in the amount indicated by the earnings account.

[In the Salary treatment, the previous paragraph was replaced by the following:

Each subsequent round proceeds in the same way and will be repeated for each of 20 rounds. In each round, you will make a forecast based on two new cue values. At the end of round 20, you will receive $20.00 in cash.

Note that your actual earnings for the experiment will be $20.00, regardless of your forecasting accuracy. However, we would like you to try and earn as much as possible by forming as accurate of forecasts as possible.]

However, after you have finished the first five rounds of play, we will have a pause. It is possible that there will be a change in the way in which you earn money for the subsequent rounds 6 through 20. If there is no change then we will tell you so and ask you to simply continue playing the game in the same manner as in the first five rounds. However, if there is a change in payment, then we will provide you with further instructions at that point and explain these changes and also answer any questions you may have.

Do not use calculators. Your forecasts and your earnings are your private information. **It is important that you do not talk or in any way try to communicate with other people during the experiment. If you violate the rules, you will be asked to leave the experiment.**

GOOD LUCK!!

A2.2. Piece Rate Win Lose Treatment Specific Instructions

THE UNIVERSITY OF AUCKLAND

Rounds 6 to 20

Rounds 6 to 20 are played exactly as rounds 1 to 5 but with the following exceptions:

- Each period you will be paired with another participant in the session today. Your partner will change each round, so you will never be paired with the same partner more than once;
- After you have made your forecast, the computer will compare your forecast error to your partner's forecast error in that round;
- Your results will show whether your forecast error was greater or less than your partner's for that round;
- If your error is less than your partner's, then you will be told you WIN that round. If your error is more than your partner's, you will be told you LOST that round. If your error is equal to your partner's, then the computer will randomly decide the winner and loser.
- Your payment will remain unchanged. That is, each round you will continue to be paid:
  Earnings = $1.00 – Forecast Error

You will also be shown your partner's forecast and forecast error at the end of the round. That is, at the end of each round you will observe:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE

Example: Suppose the actual price was 210, your forecast was 168, and your partner's forecast was 163. Your forecast error would be 42 and your partner's forecast error would be 47. You would see the following results for that round:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
		168	210	42	$0.58	WIN

Example: Suppose the actual price was 210, your forecast was 168, and your partner's forecast was 173. Your forecast error would be 42 and your partner's forecast error would be 37. You would see the following results for that round:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
		168	210	42	$0.58	LOSE

Do you have any questions?

A2.3. Tournament Treatment Specific Instructions

THE UNIVERSITY OF AUCKLAND

Rounds 6 to 20

Rounds 6 to 20 are played exactly as rounds 1 to 5 but with the following exceptions:

- You will have an additional $4.00 added to your Earnings Account;
- Each period you will be paired with another participant in the session today. Your partner will change each round, so you will never be paired with the same partner more than once;
- After you have made your forecast, the computer will compare your forecast error to your partner's forecast error in that round;
- Your results will show whether your forecast error was greater or less than your partner's for that round;
- If your error is less than your partner's, then you will be told you WIN that round. If your error is more than your partner's, you will be told you LOST that round. IF your error is equal to your partner's, then the computer will randomly decide the winner and loser;
- Your payment will depend upon whether your forecast error is greater or less than your partners. That is, each round you will earn either $1.00 or $0.00. You will be paid either:

Earnings = $1.00 if you WIN
Or
Earnings = 0 if you LOSE

Example: Suppose your forecast error was 42 and your partner's forecast error was 47. You would see the following results for that round:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42	$1.00	WIN

Example: Suppose your forecast error was 42 and your partner's forecast error was 37. You would see the following results for that round:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42	$0.00	LOSE

Do you have any questions?

A2.4. Tournament-No-Info Treatment Specific Instructions

THE UNIVERSITY OF AUCKLAND

Rounds 6 to 20

Rounds 6 to 20 are played exactly as rounds 1 to 5 but with the following exceptions:

- You will have an additional $4.00 added to your Earnings Account;
- Each period you will be paired with another participant in the session today. Your partner will change each round, so you will never be paired with the same partner more than once;
- After you have made your forecast, the computer will compare your forecast error to your partner's forecast error in that round;
- If your error is less than your partner's, then you will be told you WIN that round. If your error is more than your partner's, you will be told you LOST that round. If your error is equal to your partner's, then the computer will randomly decide the winner and loser;
- Your payment will depend upon whether your forecast error is greater or less than your partner. That is, each round you will earn either $1.00 or $0.00. You will be paid either:

  Earnings = $1.00 if you WIN
  Or
  Earnings = 0 if you LOSE

You will not know whether you won or lost until the end of the 20th round. That is, at the end of each round you will see the following information:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error

At the end of the 20th round, you will see the following information for each round:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
						WIN

You will only know whether you won or lost at the end of the 20th round.

Example: Suppose your forecast error was 42 and your partner's forecast error was 47. At the end of that round you would observe:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42		

At the end of the 20th round, you will observe:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42	$1.00	WIN

Example: Suppose your forecast error was 42 and your partner's forecast error was 37. At the end of that round you would observe:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42		

At the end of the 20th round, you will observe:

Cue A	Cue B	Your Forecast	Actual Price	Forecast Error	Earnings this round	WIN or LOSE
				42	$0.00	LOSE

Do you have any questions?

A2.5. Instructions for Logging into the Computerised Game

PLEASE FOLLOW THESE INSTRUCTIONS CAREFULLY

- Log onto the computer using your EC account. You will not be charged for the activity.

- Open the Internet Explorer Browser. You might need to make sure that you have activated your NetLogin.

- In the address bar of the browser, enter the following address:
  http://econresearch6.eco.auckland.ac.nz:8080/gameservlet

- PLEASE NOTE:
  - DO NOT USE THE 'BACK KEY'
  - ALWAYS USE THE MOUSE…DO NOT USE THE ENTER KEY

- You should now see a PLAYER LOGIN SCREEN

WAIT
DO NOT PROCEED UNTIL INSTRUCTED

- When instructed to do so, please enter an identification name in the box at the lower left side of the screen.

  NOTE: This name can be anything you like, such as 'boy' or 'max774' or 'ShR1Ely'. It is the name the computer will use to identify your computer throughout the session.

- You should now see a screen showing the Practice Rounds.

WAIT
DO NOT PROCEED UNTIL INSTRUCTED

A2.6. Pre-Task Questionnaire

THE UNIVERSITY OF AUCKLAND
BUSINESS SCHOOL
ECONOMICS

Player ID ____________

PLEASE ANSWER ALL OF THE FOLLOWING QUESTIONS

A number of statements which people have used to describe themselves are given below. Read each statement and, using the scale below, tick the appropriate number indicating **how you generally feel**. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe **how you generally feel**.

1	2	3	4
Almost never	Sometimes	Often	Almost always

	Almost Never	Sometimes	Often	Almost always
	1	2	3	4
1. I feel pleasant				
2. I tire quickly				
3. I feel like crying				
4. I wish I could be as happy as others seem to be				
5. I am losing out on things because I can't make up my mind soon enough				
6. I feel rested				
7. I am "calm, cool and collected"				
8. I feel that difficulties are piling up so that I cannot overcome them				
9. I worry too much over something that doesn't really matter				
10. I am happy				
11. I am inclined to take things hard				
12. I lack self-confidence				
13. I feel secure				
14. I try to avoid facing a crisis or difficulty				
15. I feel blue				
16. I am content				
17. Some unimportant thoughts run through my mind and bother me				
18. I take disappointments so keenly that I can't put them out of my mind				
19. I am a steady person				
20. I get in a state of tension or turmoil as I think over my recent concerns and interests				

A2.7. Post-Task Questionnaire

THE UNIVERSITY OF AUCKLAND
BUSINESS SCHOOL
ECONOMICS

Player ID ____________

PLEASE ANSWER ALL OF THE FOLLOWING QUESTIONS

A) For each of the following statements, please indicate how true the statement is for you using the following scale:

1 2 3 4 5 6 7
Not at all true Somewhat true Very true

	Not at all true			Somewhat true			Very true
	1	2	3	4	5	6	7
1. I enjoyed this activity very much							
2. I think I am pretty good at this activity							
3. I put a lot of effort into this activity							
4. I did not feel nervous at all which doing this activity							
5. This activity was fun to do							
6. I think I did pretty well at this activity, compared to other participants							
7. I did not try very hard to do well at this activity							
8. I felt very tense while doing this activity							
9. I thought this activity was boring							
10. After working at this activity for a while, I felt pretty competent							
11. I tried very hard on this activity							
12. I was very relaxed doing this activity							
13. This activity did not hold my attention							
14. I am satisfied with my performance at this task							
15. It was important to me to do well at this task							
16. I was anxious while working on this task							
17. I would describe this activity as very interesting							
18. I was pretty skilled at this activity							
19. I did not put much energy into this							
20. I felt pressured while doing this activity.							
21. I thought this activity was quite enjoyable.							
22. This was an activity that I could not do very well.							
23. While I was doing this activity, I was thinking about how much I enjoyed it.							

B) The following items ask about how you felt about the other participants during the session.

	Not at all true			Somewhat true			Very true
	1	2	3	4	5	6	7
1. I felt really distant to them							
2. I really doubt they and I would ever be friends							
3. I felt I could really trust them							
4. I'd really like the chance to interact with them more often							
5. I'd really prefer not to interact with them in the future							
6. I don't feel like I could really trust them							
7. It is likely that they and I could become friends if we interacted a lot							
8. I felt close to them							

C) How many of the people in this session did you know before the experiment? ________

D) Basic information about you:

Your Gender (Male/ Female) ________

Age: ______

Major: ____

Year in School (e.g., Stage 2) ______________________________

Ethnicity (Please circle one):

Maori Pacific Island NZ European
Asian Other _______________

Country where you were born? __________________

If you were born outside of New Zealand, at what age did you move here? _________

Appendix 3. Proof of Standardised Forecast Error Properties

A3.1. Standardised Forecast Errors have Zero-Mean

Standardised forecast errors, z, are defined to be:

$$z_{it} = \frac{e_{it} - \bar{e}_t}{\sigma_t(e)}$$

The mean of this across participants i in any particular round:

$$\bar{z}_t = \frac{1}{n}\sum_{i}^{n} z_{it}$$

$$\bar{z}_t = \frac{1}{n}\sum_{i}^{n}\left(\frac{e_{it} - \bar{e}_t}{\sigma_t(e)}\right)$$

And since the mean and standard deviation of forecast errors are invariant with individuals i,

$$\bar{z}_t = \frac{1}{n}\cdot\frac{1}{\sigma_t(e)}\cdot\left(\left(\sum_{i}^{n} e_{it}\right) - n\bar{e}_t\right)$$

$$\bar{z}_t = \frac{1}{n}\cdot\frac{1}{\sigma_t(e)}\cdot\left(\left(\sum_{i}^{n} e_{it}\right) - n\left(\frac{\sum_{i}^{n} e_{it}}{n}\right)\right)$$

$$\bar{z}_t = \frac{1}{n}\cdot\frac{1}{\sigma_t(e)}\cdot\left(\sum_{i}^{n} e_{it} - \sum_{i}^{n} e_{it}\right)$$

$$\bar{z}_t = 0$$

■

A3.2. Standardised Forecast Errors have Unit Standard Deviation

$$\sigma_t(z) = \sqrt{\sum_i^n (z_{it} - \bar{z}_t)^2}$$

From previous proof, the mean of standardised forecast errors is zero:

$$\sigma_t(z) = \sqrt{\sum_i^n (z_{it})^2}$$

$$\sigma_t(z) = \sqrt{\sum_i^n \left(\frac{e_{it} - \bar{e}_t}{\sigma_t(e)}\right)^2}$$

Since the standard deviation of forecast errors amongst participants does not differ by participants i, it can be factored out:

$$\sigma_t(z) = \sqrt{\left(\frac{1}{\sigma_t(e)}\right)^2 \cdot \sum_i^n (e_{it} - \bar{e}_t)^2}$$

$$\sigma_t(z) = \frac{1}{\sigma_t(e)} \cdot \sqrt{\sum_i^n (e_{it} - \bar{e}_t)^2}$$

Substituting in the calculation of $\sigma_t(e)$, and cancelling out the terms:

$$\sigma_t(z) = \frac{1}{\sqrt{\sum_i^n (e_{it} - \bar{e}_t)^2}} \cdot \sqrt{\sum_i^n (e_{it} - \bar{e}_t)^2}$$

$$\sigma_t(z) = 1$$

■

9. References

Akerlof, George A., & Yellen, Janet L. (1990). The Fair Wage-Effort Hypothesis and Unemployment. *Quarterly Journal of Economics, 105*(2), 255-283.

Ariely, Dan, Bracha, Anat, & Meier, Stephan. (2009). Doing Good of Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review, 99*(1), 544-555.

Azmat, Ghazala, & Iriberri, Nagore. (2010). The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students. *Journal of Public Economics, 94*(7-8), 435-452.

Azmat, Ghazala, & Iriberri, Nagore. (2016). The Provision of Relative Performance Feedback: An Analysis of Performance and Satisfaction. *Journal of Economics & Management Strategy, 25*(1), 77-110.

Balzer, William K., Doherty, Michael E., & O'Connor, Raymond. (1989). Effects of Cognitive Feedback on Performance. *Psychological Bulletin, 106*(3), 410-433.

Bandiera, Oriana, Barankay, Iwan, & Rasul, Imran. (2005). Social Preferences and the Response to Incentives: Evidence from Personnel Data. *Quarterly Journal of Economics, 120*(3), 917-962.

Bandiera, Oriana, Larcinese, Valentino, & Rasul, Imran. (2015). Blissful Ignorance? A Natural Experiment on the Effect of Feedback on Students' Performance. *Labour Economics, 34*, 13-25.

Baumeister, Roy F., & Leary, Mark R. (1995). The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological Bulletin, 117*(3), 497-529.

Bellemare, Charles, Lepage, Patrick, & Shearer, Bruce. (2010). Peer Pressure, Incentives, and Gender: An Experimental Analysis of Motivation in the Workplace. *Labour Economics, 17*(1), 276-283.

Bénabou, Roland, & Tirole, Jean. (2006). Incentives and Prosocial Behavior. *American Economic Review, 96*(5), 1652-1678.

Bereby-Meyer, Yoella, & Roth, Alvin E. (2006). The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation. *American Economic Review, 96*(4), 1029-1042.

Blanes i Vidal, Jordi, & Nossol, Mareike. (2011). Tournaments Without Prizes: Evidence from Personnel Records. *Management Science, 57*(10), 1721-1736.

Bowles, Samuel, & Polanía-Reyes, Sandra. (2012). Economic Incentives and Social Preferences: Substitutes or Complements? *Journal of Economic Literature, 50*(2), 368-425.

Brown, Paul M. (1995). Learning from Experience, Reference Points, and Decision Costs. *Journal of Economic Behavior & Organization, 27*(3), 381-399.

Brown, Paul M. (1998). Experimental Evidence on the Importance of Competing for Profits on Forecasting Accuracy. *Journal of Economic Behavior & Organization, 33*(2), 259-269.

Bruno, Bruna. (2013). Reconciling Economics and Psychology on Intrinsic Motivation. *Journal of Neuroscience, Psychology and Economics, 6*(2), 136-149.

Bull, Clive, Schotter, Andrew, & Weigelt, Keith. (1987). Tournaments and Piece Rates: An Experimental Study. *Journal of Political Economy, 95*(1), 1-33.

Cadsby, C. Bram, Engle-Warnick, Jim, Fang, Tony, & Song, Fei. (2010). Psychological Incentives, Financial Incentives, and Risk Attitudes in Tournaments: An Artefactual Field Experiment *Working Paper*.

Cameron, Judy, & Pierce, W. David. (1994). Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis. *Review of Educational Research, 64*(3), 363-423.

Card, David, Mas, Alexandre, Moretti, Enrico, & Saez, Emmanuel. (2012). Inequality at Work: The Effect of Peers Salaries on Job Satisfaction. *American Economic Review, 102*(6), 2981-3003.

Cardella, Eric. (2012). Learning to Make Better Strategic Decisions. *Journal of Economic Behavior & Organization, 84*(1), 382-392.

Charness, Gary, & Gneezy, Uri. (2009). Incentives to Exercise. *Econometrica, 77*(3), 909-931.

Charness, Gary, & Levin, Dan. (2005). When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect. *American Economic Review, 95*(4), 1300-1309.

Charness, Gary, Masclet, David, & Villeval, Marie Claire. (2014). The Dark Side of Competition for Status. *Management Science, 60*(1), 38-55.

deCharms, Richard. (1968). *Personal Causation: The Internal Affective Determinants of Behavior*. New York: Academic Press.

Dechenaux, Emmanuel, Kovenock, Dan, & Sheremeta, Roman M. (2015). A Survey of Experimental Research on Contests, All-Pay Auctions and Tournaments. *Experimental Economics, 18*(4), 609-669.

Deci, Edward L. (1971). Effects of Externally Mediated Rewards on Intrinsic Motivation. *Journal of Personality and Social Psychology, 18*(1), 105-115.

Deci, Edward L. (1972). Intrinsic Motivation, Extrinsic Reinforcement, and Inequity. *Journal of Personality and Social Psychology, 22*(1), 113-120.

Deci, Edward L., Koestner, Richard, & Ryan, Richard M. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin, 125*(6), 627-668.

Deci, Edward L., & Ryan, Richard M. (1985). *Intrinsic Motivation and Self-Determination in Human Behaviour*. New York: Plenum Press.

Delfgaauw, Josse, Dur, Robert, Non, Arjan, & Verbeke, Willem. (2014). Dynamic Incentive Effects of Relative Performance Pay: A Field Experiment. *Labour Economics, 28*(1-13).

Dohmen, Thomas, & Falk, Armin. (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences and Gender. *American Economic Review, 101*(2), 556-590.

Dutcher, E. Glenn, Balafoutas, Loukas, Lindner, Florian, Ryvkin, Dmitry, & Sutter, Matthias. (2015). Strive to be First or Avoid being Last: An Experiment on Relative Performance Incentives. *Games and Economic Behavior, 94*, 39-56.

Ehrenberg, Ronald G., & Bognanno, Michael L. (1990). Do Tournaments have Incentive Effects? *Journal of Political Economy, 98*(6), 1307-1324.

Ellingsen, Tore, & Johannesson, Magnus. (2007). Paying Respect. *Journal of Economic Perspectives, 21*(4), 135-149.

Epstein, Jennifer A., & Harackiewicz, Judith M. (1992). Winning is Not Enough: The Effects of Competition and Achievement Orientation on Intrinsic Interest. *Personality and Social Psychology Bulletin, 18*(2), 128-138.

Erev, Ido, Bereby-Meyer, Yoella, & Roth, Alvin E. (1999). The Effect of Adding a Constant to All Payoffs: Experimental Investigation, and Implications for Reinforcement Learning Models. *Journal of Economic Behavior & Organization, 39*(1), 111-128.

Erev, Ido, & Roth, Alvin E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review, 88*(4), 848-881.

Eriksson, Tor. (1999). Executive Compensation and Tournament Theory: Empirical Tests on Danish Data. *Journal of Labor Economics, 17*(2), 262-280.

Eriksson, Tor, Poulsen, Anders, & Villeval, Marie-Claire. (2009). Feedback and Incentives: Experimental Evidence. *Labour Economics, 16*(6), 679-688.

Falk, Armin, & Fehr, Ernst. (2003). Why Labour Market Experiments? *Labour Economics, 10*(4), 399-406.

Falk, Armin, & Ichino, Andrea. (2006). Clean Evidence on Peer Effects. *Journal of Labor Economics, 24*(1), 39-57.

Falk, Armin, & Kosfeld, Michael. (2006). The Hidden Costs of Control. *American Economic Review, 96*(5), 1611-1630.

Fehr, Ernst, & Schmidt, Klaus M. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics, 114*(3), 817-868.

Feltovich, Nick. (2000). Reinforcement-based vs Belief-based Learning Models in Experimental Asymmetric-Information Games. *Econometrica, 68*(3), 605-641.

Feltovich, Nick, & Ejebu, Ourega-Zoé. (2014). Do Positional Goods Inhibit Saving? Evidence from a Life-Cycle Experiment. *Journal of Economic Behavior & Organization, 107*, 440-454.

Fershtman, Chaim, & Gneezy, Uri. (2011). The Tradeoff between Performance and Quitting in High Power Tournaments. *Journal of the European Economic Association, 9*(2), 318-336.

Festré, Agnès, & Garrouste, Pierre. (2015). Theory and Evidence in Psychology and Economics about Motivation Crowding Out: A Possible Convergence? *Journal of Economic Surveys, 29*(2), 339-356.

Frey, Bruno S., & Jegen, Reto. (2001). Motivation Crowding Theory. *Journal of Economic Surveys, 15*(5), 589-611.

Fryer, Roland G. (2011). Financial Incentives and Student Achievement: Evidence from Randomised Trials. *Quarterly Journal of Economics, 126*, 1755-1798.

Fryer, Roland G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics, 31*(2), 373-407.

Fu, Qiang, Ke, Changxia, & Tan, Fangfang. (2015). "Success Breeds Success" or "Pride Goes Before a Fall"? Teams and Individuals in Multi-Contest Tournaments. *Games and Economic Behavior, 94*, 57-79.

Furrer, Carrie, & Skinner, Ellen. (2003). Sense of Relatedness as a Factor in Children's Academic Engagement and Performance. *Journal of Educational Psychology, 95*(1), 148-162.

Gneezy, Uri, Meier, Stephan, & Rey-Biel, Pedro. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives, 25*(4), 191-210.

Gneezy, Uri, & Rustichini, Aldo. (2000a). A Fine is a Price. *Journal of Legal Studies, 29*(1), 1-17.

Gneezy, Uri, & Rustichini, Aldo. (2000b). Pay Enough or Don't Pay at All. *Quarterly Journal of Economics, 115*, 791-810.

Green, Jerry R., & Stokey, Nancy L. (1983). A Comparison of Tournaments and Contracts. *Journal of Political Economy, 91*(3), 349-364.

Hannan, R. Lynn, Krishnan, Ranjani, & Newman, Andrew H. (2008). The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans. *The Accounting Review, 83*(4), 893-913.

Harackiewicz, Judith M. (1979). The Effects of Reward Contingency and Performance Feedback on Intrinsic Motivation. *Journal of Personality and Social Psychology, 37*(8), 1352-1363.

Holmås, Tor Helge, Kjerstad, Egil, Lurås, Hilde, & Straume, Odd Rune. (2010). Does Monetary Punishment Crowd Out Pro-Social Motivation? A Natural Experiment on Hospital Length of Stay. *Journal of Economic Behavior & Organization, 75*(2), 261-267.

Hyde, Janet Shibley, Fannema, Elizabeth, & Lamon, Susan J. (1990). Gender Differences in Mathematics Performance: A Meta-Analysis. *Psychological Bulletin, 107*(2), 139-155.

James, Harvey S. (2005). Why Did You Do That? An Economic Examination of the Effect of Extrinsic Compensation on Intrinsic Motivation and Performance. *Journal of Economic Psychology, 26*(4), 549-566.

Kandel, Eugene, & Lazear, Edward P. (1992). Peer Pressure and Partnerships. *Journal of Political Economy, 100*(41), 801-817.

Kluger, Avraham N., & DeNisi, Angelo. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin, 119*(2), 254-284.

Knoeber, Charles R., & Thurman, Walter N. (1994). Testing the Theory of Tournaments: An Empirical Analysis of Broiler Production. *Journal of Labor Economics, 12*(2), 155-179.

Kosfeld, Michael, & Neckermann, Susanne. (2011). Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics, 3*(3), 86-99.

Kräkel, Matthias. (2008). Emotions in Tournaments. *Journal of Economic Behavior & Organization, 67*(1), 204-214.

Kuhn, Peter, Kooreman, Peter, Soetevent, Adriaan, & Kapteyn, Arie. (2011). The Effects of Lottery Prizes on Winners and Their Neighbours: Evidence from the Dutch Postcode Lottery. *American Economic Review, 101*(5), 2226-2247.

Kuhnen, Camelia M., & Tymula, Agnieszka. (2012). Feedback, Self-Esteem, and Performance in Organizations. *Management Science, 58*(1), 94-113.

Lazear, Edward P. (1986). Salaries and Piece Rates. *Journal of Business, 59*(3), 405-431.

Lazear, Edward P. (2000). Performance Pay and Productivity. *American Economic Review, 90*(5), 1346-1361.

Lazear, Edward P., & Rosen, Sherwin. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy, 89*(5), 841-864.

Lepper, Mark R., Greene, David, & Nisbett, Richard E. (1973). Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the "Overjustification" Hypothesis. *Journal of Personality and Social Psychology, 28*(1), 129-137.

Levin, Andrew, Lin, Chien-Fu, & Chu, Chia-Shang James. (2002). Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. *Journal of Econometrics, 108*(1), 1-24.

Ludwig, Sandra, & Lünser, Gabriele K. (2012). Observing Your Competitor - The Role of Effort Information in Two Stage Tournaments. *Journal of Economic Psychology, 33*(1), 166-182.

Mas, Alexandre, & Moretti, Enrico. (2009). Peers at Work. *American Economic Review, 99*(1), 112-145.

Masclet, David, Peterle, Emmanuel, & Larribeau, Sophie. (2015). Gender Differences in Tournament and Flat-Wage Schemes: An Experimental Study. *Journal of Economic Psychology, 47*, 103-115.

Merlo, Antonio, & Schotter, Andrew. (1999). A Surprise-Quiz View of Learning in Economic Experiments. *Games and Economic Behavior, 28*(25-54).

Merlo, Antonio, & Schotter, Andrew. (2003). Learning by Not Doing: An Experimental Investigation of Observational Learning. *Games and Economic Behavior, 43*(1), 116-136.

Müller, Wieland, & Schotter, Andrew. (2010). Workaholics and Dropouts in Organizations. *Journal of the European Economic Association, 8*(4), 717-743.

Nalebuff, Barry J., & Stiglitz, Joseph E. (1983). Prizes and Incentives: Towards a General Theory of Compensation and Competition. *Bell Journal of Economics, 14*(1), 21-43.

Niederle, Muriel, & Vesterlund, Lise. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *Quarterly Journal of Economics, 122*(3), 1067-1101.

Pesaran, M. Hashem, & Smith, Ron. (1995). Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics, 68*(1), 79-113.

Prendergast, Canice. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature, 37*(1), 7-63.

Promberger, Marianne, & Marteau, Theresa M. (2013). When do Financial Incentives Reduce Intrinsic Motivation? Comparing Behaviors Studied in Psychological and Economic Literatures. *Health Psychology, 32*(9), 950-957.

Reeve, Johnmarshall, & Deci, Edward L. (1996). Elements of the Competitive Situation That Affect Intrinsic Motivation. *Personality and Social Psychology Bulletin, 22*(1), 24-33.

Rick, Scott, & Weber, Roberto A. (2010). Meaningful Learning and Transfer of Learning in Games Played Repeatedly without Feedback. *Games and Economic Behavior, 2010*, 716-730.

Roth, Alvin E., & Erev, Ido. (1995). Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior, 8*, 164-212.

Ryan, Richard M. (1982). Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *Journal of Personality and Social Psychology, 43*(3), 450-461.

Ryan, Richard M., Mims, Valerie, & Koestner, Richard. (1983). Relation of Reward Contingency and Interpersonal Context to Intrinsic Motivation: A Review and Test Using Cognitive Evaluation Theory. *Journal of Personality and Social Psychology, 45*(4), 736-750.

Ryan, Richard M., & Powelson, Cynthia L. (1991). Autonomy and Relatedness as Fundamental to Motivation and Education. *Journal of Experimental Education, 60*(1), 49-66.

Schotter, Andrew, & Weigelt, Keith. (1992). Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results. *Quarterly Journal of Economics, 107*(2), 511-539.

Segal, Jan D., & Weinberg, Robert S. (1984). Sex, Sex Role Orientation and Competitive Trait Anxiety. *Journal of Sport Behavior, 7*(4), 153-159.

Shearer, Bruce. (2004). Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *Review of Economic Studies, 71*, 513-534.

Spielberger, Charles D., Gorsuch, Richard L., Lushene, Robert E., Vagg, Peter R., & Jacobs, Gerard A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Tran, Anh, & Zeckhauser, Richard. (2012). Rank as an Inherent Incentive: Evidence from a Field Experiment. *Journal of Public Economics, 96*(9-10), 645-650.

Vandegrift, Donald, & Brown, Paul M. (2003). Task Difficulty, Incentive Effects, and the Selection of High-Variance Strategies: An Experimental Examination of Tournament Behavior. *Labour Economics, 10*(4), 481-497.

Vandegrift, Donald, & Brown, Paul M. (2005). Gender Differences in the Use of High-Variance Strategies in Tournament Competition. *Journal of Socio-Economics, 34*(6), 834-849.

Vandegrift, Donald, Yavas, Abdullah, & Brown, Paul M. (2007). Incentive Effects and Overcrowding in Tournaments: An Experimental Analysis. *Experimental Economics, 10*(4), 345-368.