Feedback and Performance: Experiments in Behavioral Economics

by

William Gilje Gjedrem

Thesis submitted in fulfillment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

Faculty of Social Sciences
UiS Business School
2016

University of Stavanger
N-4036 Stavanger
NORWAY
www.uis.no

ISBN: 978-82-7644-685-2
ISSN: 1890-1387
PhD thesis no. 321

Preface

This thesis is submitted in fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the University of Stavanger, Faculty of Social Sciences, Norway. The project was funded by the Norwegian Research Council (227004). The financial support from the Norwegian Research Council is gratefully acknowledged.

The thesis consists of four separate essays which are summarized in the introduction, including a brief discussion concerning the concept of causality and the use of experiments as an empirical strategy. Chapter 2 is the first essay, which has been written together with Mari Rege. Chapter 3 is my single author essay. Chapter 4 is joint work with Ola Kvaløy. Finally, chapter 5 is written in collaboration with Kristoffer Wigestrand Eriksen and Jon Kristian Heimdal.

Acknowledgements

I would like to extend my warmest gratitude to my supervisor Mari Rege. She has been very important during the four years that I have worked on this thesis, and without her support the bumpy road towards a PhD would certainly have been bumpier. Her encouragement was also influential in my decision to pursue a PhD in the first place. I am also very thankful to my co-supervisor, Ola Kvaløy, for his continued support, the encouraging discussions we have had, and the constructive feedback he has provided. The contribution of my second co-supervisor, Mark Votruba, is also appreciated.

All of my colleagues at the University of Stavanger, and especially those whom I have collaborated with the most, deserve a big thank you. This also includes the social aspect of the work environment; an extremely important factor to stay motivated on a daily basis. I am also very grateful to all of my co-authors.

Finally, I would like to extend some thankful words to my family. I would certainly not have completed the PhD without their understanding and support. Especially all the support from my wife, Ingrid, and our son, Edwin, have been cardinal for me. Words are not strong enough to signify how indebted I am to them.

Contents

Introduction

Understanding behavioral phenomenon in the social sciences is complex as many factors influence the relationships we observe. In recent decades, economists have more commonly used experiments to face the key challenge of identifying causal effects of human behavior. In this thesis, I use a natural quasi-experiment to identify the causal relationship between autonomy and productivity in the field, as well as three laboratory experiments to identify the causal effects of feedback on behavioral responses in the workplace and on investment decisions.

In all of the essays that follow, I seek to identify how different levels of performance feedback influence human behavior. Performance feedback is the simple provision of information about outcomes. There are numerous ways to create variation in performance feedback, and in the following essays I either vary the frequency of the provision of feedback, or between absolute performance feedback (APF) and relative performance feedback (RPF). I think of APF as information about the absolute outcome of a certain situation, i.e., information about the performance of oneself and with no information regarding what others have performed. On the other hand, RPF is information about the relative outcome of a certain situation, i.e., information about the performance of oneself relative to another. As an example from a workplace, APF could be how much revenue an employee generated one day, and RPF could be how much revenue an employee generated relative to a colleague, an average of all colleagues, etc. For an investment manager, APF could be the daily stock market return, and RPF could be the stock market return of the other employees in the investment company or the market index.

Therefore, a fundamental question is: How should we expect people to respond to performance feedback in general? From a scientific point of view, we may approach such a question theoretically and conduct empirical studies to explore real outcomes. In standard economic theory, where we assume that individuals are maximizing payoff, performance feedback itself should not matter. As individuals only care about maximizing payoff, additional information about performance is simply neglected. However, we may also use a theoretical approach that allows for behavioral aspects, and lets these aspects influence individuals who now seek to maximize utility. Such behavioral concerns may be

a person's self-esteem or social considerations. This makes the utility maximizing function complex, as there are potentially many factors that could explain responses to variation in performance feedback. For example, knowing how one performed relative to another may, for example, generate feelings of inadequacy, invoke competitive preferences or entail self-criticism, which are all likely to affect future motivation and performance. Instead, if the feedback was only absolute, and not relative, these concerns may not have been invoked. From empirical studies, we know that variations in the frequency of feedback, or varying between APF and RPF, sometimes affect future decisions about investments and performance levels. However, there are still many unanswered questions with respect to how people respond to performance feedback, and the results so far suggest certain contextual dependency.

The remaining part of this first section starts with a brief summary of the essays that follow in later sections of the thesis. After that, I will argue that experiments have important properties that make them a good alternative as an empirical strategy to identify causal relationships.

1. Summary of Essays

Chapter two "The Effect of Less Autonomy on Productivity in Retail: Evidence from a Quasi-Natural Field Experiment" (with Mari Rege).

In this essay we study how relative performance feedback, as part of a larger change in managerial practice, affects the overall sales of a retail chain. The feedback policy that was introduced provided employees with sales data at the store level so that the store's performance was compared relative to others. Moreover, it recorded how many times the staff actively approached customers and compared this directly to the development of sales.

A primary question which is presently debated is: Who should be responsible for decision-making in modern workplaces; the manager or the employees themselves? Managers may prefer to make decisions on a centralized level to ensure compliance with the overall strategy, as well as to maintain control and consistency. On the other hand, employees are often the ones possessing the specializations and qualifications needed to make decisions for complex tasks. Several studies have shown that decentralized decision-making is important

for productivity, innovation, and motivation in high-skilled complex jobs (e.g. Hackman and Oldham, 1976; Milgrom and Roberts, 1995; Ichniowski and Shaw, 1999; Caroli and Van Reenen, 2001; Tambe and Hitt, 2012). Importantly though, many jobs still do not require any special skills, and decision-making can therefore benefit from having been made strategically by managers. This relates to an important topic in personnel economics; how to efficiently delegate decisions within a firm (Lazear and Gibbs, 2014)? This question is of crucial importance for the long-term profitability of a business. Managers who strategically define best practices in such low-skilled jobs, may increase both productivity and motivation (Lazear and Gibbs, 2014).

This paper uses a quasi-experimental approach to examine the causal relationship between autonomy and productivity in a low-skilled and narrowly defined job, utilizing weekly sales data over two years for stores in a Norwegian consumer electronics retail chain. Specifically, we investigate a change in management practice to more detailed job instructions for sales staff, in addition to increased systematic control and feedback. A crucial decision which sales staff has to make, many times a day, is whether or not to approach a customer entering the store. Some customers appreciate immediate contact with sales staff, whereas others prefer to be left alone, and sometimes these preferences are signaled by body language. The new practice required sales staff to always actively approach customers.

We use a triple difference empirical approach to examine the effects of this change on management practice. This design of the experiment benefits from the fact that the change was only introduced in some stores and at different points in time. Identification of the treatment effect arises from differential changes in sales in treated stores relative to controls in the weeks before and after treatment introduction. This was compared relative to the same double-difference during the previous year with no treatment. The analysis is based on the following key assumption; differential trends in sales across treated and non-treated stores are identical in both the year before and the year of the treatment introduction. This assumption may not hold, and we do specification tests and placebo analysis to address the validity.

Our results indicate that the change caused an average increase in sales of about 4.3 percent and in the number of transactions by about 3.3 percent. This

suggests that more detailed job instructions, based on best practice, may increase productivity in low-skilled, narrowly defined jobs. Placebo analysis shows no differences in sales between treated and comparison stores in the absence of the treatment. In addition, our analyses are not sensitive to shifting the period of analysis. This strengthens our belief in that the main results are effects actually caused by the treatment. Moreover, differential analysis suggests that the treatment benefits smaller stores more than larger stores, and that stores reporting high treatment-compliance increased sales more than low treatment-compliance stores. Finally, the effect seems to be persistent.

Chapter three "Relative Performance Feedback: Effective or Dismaying?".

The rapid technological development of recent decades has made it easier for employers to collect and analyze employee performance data. Some organizations use this information to provide employees with relative performance feedback (RPF) in an attempt to increase motivation and performance. There are, however, reasons to question whether such information always improves productivity. For example, competition between employees for higher ranks may drive performance up, but for others it may be demotivating to always perform worse than others. A particular worry is that some mechanisms "crowd out" the intrinsic motivation of employees to work (Deci, 1971; Frey and Oberholzer-Gee, 1997).

This paper continues to explore the relationship between feedback and productivity, but herein the experiment is conducted in a the more controlled study environment of a laboratory. Two aspects of peoples' social concerns are likely to be important reasons why RPF affects motivation. Firstly, people have competitive preferences and secondly, they care about their relative competence levels. The latter aspect is considered the core of intrinsic motivation (Deci and Ryan, 2000), and learning about the performance of others may adjust the perception of one's own competence. However, people may have competitive preferences too, which are strengthened with the introduction of relative performance feedback. In an effort to disentangle these social concerns, this chapter presents an experiment that includes treatments designed to feature each concern separately, which should provide us with insight into how people respond differently to RPF in various environments.

Two treatments are used to feature each social concern separately. The first treatment (CPF treatment) uses the past performances of participants as benchmarks to rank the current subjects' performance. Importantly, subjects in this treatment do not learn anything about the performance of any other subject in the same session. Thus, the environment is designed to reduce the competitiveness to a minimum, and rather provide a signal about the general competence level of others to solve the specific task. The second treatment (TPF treatment) uses the performance of three others working alongside the subject as the ranking benchmark. This should raise competitiveness to a higher level as subjects compete against each other for the high ranks. In contrast to the former treatment, in the latter there is only mere 'talk' about the general competence level of others. The two treatments are compared to a baseline where subjects only learn about their own absolute performance.

The overall results, using non-parametric tests, suggest no performance difference between the baseline and treatments under any pay-scheme. However, regression analysis is required to adequately control for the subjects' ability and to test for heterogeneous reactions. These analyses show that, when payment is fixed, the average performance of subjects is greater in both treatments compared to the baseline, but this is only significant in the CPF treatment. Large variations in performance exist, especially in the CPF treatment where subjects with low self-assessed ability (SAA) reduce their performance substantially when RPF is provided. For the equivalent group of subjects in the TPF treatment, no such negative response was identified. Moreover, those who report high ability perform better in both treatments. In the performance pay conditions of the experiment no average treatment effects have been identified. However, differential analysis shows that males and females respond differently depending on their reported ability.

Chapter four "Smells Like Team Spirit: An Experiment on Relative Performance Feedback" (with Ola Kvaløy).

People prefer high rank to low rank. Even when rank is independent from monetary outcomes, people are willing to take costly actions in order to climb the ladder. Modern organizations utilize this basic human insight by providing employees with feedback on their relative performance in order to motivate them to work harder. However, although rank and relative performance feedback

(RPF) are such basic ingredients in competitive environments, more recently economists have systematically studied how people respond to rank and RPF.

The experimental literature on RPF has thus far concentrated on individual behavior and feedback. However, not only individuals receive RPF, but also groups of individuals, such as firms, or teams within firms, who compete against each other and receive feedback about their relative performance. Sales or R&D teams, for instance, are benchmarked against similar teams in other firms. Moreover, firms often create internal competitions between teams in order to sell or innovate more (see e.g., Birkinshaw, 2001; Marino and Zabojnik, 2004; Baer et al., 2010). Successful teams are typically compensated by some monetary rewards, but the team competitions per se are also potentially motivating.

Ultimately, this paper contributes to the existing literature by investigating how teams respond to relative performance feedback and explores whether teams suffer from free-riding activities, and to what extent RPF mitigates this problem. There are several reasons why people might respond differently to team feedback compared to individual feedback. The joy of winning together with a team might be different from the joy of winning alone. Similarly, the costs of losing as a team might be different from the costs of losing alone. Moreover, repeated RPF may create peer effects within the team, which again establishes a different response to team RPF compared to individual RPF. We thus investigate to what extent and under which conditions teams respond to RPF, as well as compare how individuals respond differently to team RPF than individual RPF.

We do this by conducting a controlled laboratory experiment consisting of six treatments. In each treatment, subjects work on a real-effort task for six periods. We primarily vary treatments along two dimensions: team or individual incentives, and team or individual feedback. However, to establish a "baseline" of performance, we have two treatments in which subjects only receive absolute performance feedback. Under RPF, individuals (teams) are always compared with two other individuals (teams), i.e., after each period, each individual or team is ranked as either number 1, 2 or 3. Each team consists of three subjects, and so each subject earns one third of total team output when provided with team incentives. The monetary outcomes are independent from feedback rankings.

Our main results can be summarized as follows: We find that when subjects are exposed to team incentives the RPF on how the team is doing compared to

the two others increases its average performance by almost 10 percent. Team incentives without RPF give rise to a free-rider problem, but RPF to teams more than offsets this problem. We find that the treatment effect is driven by the teams' top performers. The average individual performance of the top performers within each team is almost 20 percent higher when the teams receive RPF compared to when the teams only receive absolute performance feedback. These effects more or less disappear under individual incentives and/or individual RPF. Our experiment thus suggests that top performers are particularly motivated by the combination of team incentives and team RPF. In fact, team incentives trigger significantly higher performance than individual incentives when subjects are exposed to team RPF.

Chapter five "Feedback and Risk-Taking with Own and Other People's Money" (with Kristoffer W. Eriksen and Jon Kristian Heimdal).

People often take risk on behalf of others. For example, politicians decide on behalf of the local or national population, and CEOs make decisions associated with risk-taking on behalf of employees and owners. In finance, investment managers trade on behalf of their customers. In 2015, U.S. registered investment companies managed assets for more than $ 18 trillion, and this was on behalf of more than 90 million retail investors (ICI, 2016). Their clients' willingness to take risk is often unknown or uncertain to the investment manager, and he may also choose different investment portfolios on behalf of others than what he does with his own wealth. Furthermore, their interests in the outcome of the investments do not necessarily align as investment managers often bear limited direct consequences of the investment outcomes.

Even though investments on behalf of others are extensive, research offers only limited guidance as to how people choose to make such investments, and it is particularly scant on how feedback on investment outcomes affects these decisions. The frequency of such outcomes has previously shown to affect investment decisions with own money (see e.g., Gneezy and Potters, 1997), and frequent feedback is natural for investment managers who closely monitor portfolios.

People who invest and take risk with their own money are affected by the frequency of feedback on investment outcomes. Benartzi and Thaler (1995) introduced the behavioral hypothesis termed myopic loss aversion (MLA) as a

possible explanation to the famous equity premium puzzle (Mehra and Prescott, 1985). It suggests that investors move towards less risky investments the more frequently they receive and evaluate feedback on investment outcomes. While the experimental literature over the last 20 years has shown that people respond to feedback manipulation when investing their own money (starting with Gneezy and Potters, 1997; Thaler et al., 1997), private investors often delegate wealth management to investment managers. Such professionals are also found to exhibit behavior consistent with MLA in experimental settings using their own money (Haigh and List, 2005; Eriksen and Kvaløy, 2010), however less is known about how and whether the bias transfer to those investment decisions on behalf of others.

In this chapter, we investigate whether feedback frequency affects decision making for individuals regarding investment for both themselves and others. We make use of the standard investment game first introduced by Gneezy and Potters (1997), and employ a within-between subjects design. That is, while we vary the feedback frequency between subjects (high and low frequency), the same subject makes risky decisions with both his/her own money and others'.

The within-subject part of the experiment allows us to shed some light on how people adapt their investment decisions when facing situations where the choices regard both their own money and that of someone else, and to what extent the manipulation of feedback frequency affects this adaption. The between part of the experiment allows us to study whether subjects exhibit MLA with their own and other people's money, and the within part allows us to study how much risk they take for both options (within). Combining these dimensions, we can also study the relative investment of subjects, i.e., how much they choose to invest with their own money relative to how much they choose to invest with other people's money, and whether the manipulation of feedback frequency affects this.

Our results show that when people invest on behalf of others, feedback frequency on investment outcomes matter. The amount they invest is the same across low and high feedback frequency. However, the relative investment is different across feedback frequency. When the frequency is low, subjects invest significantly less with other people's money compared to their own money. When feedback frequency is high, they invest about the same amount with own money

as with other people's money. In general, people do seem to exhibit MLA when they invest their own money, but not when they invest other people's money. Thus, manipulating feedback frequency does not seem to make people less afraid of risk when they invest other people's money, and therefore average risk-taking is less than with own money. Consequently, in terms of maximizing expected earnings, people who make investment choices on behalf of others may fail to perform any better than what their clients' would have done themselves.

2. Using Experiments to Unveil Causalities

Researchers aim to unveil causalities rather than simply show correlations. In this thesis, I aim to find the causal relationship between performance feedback and human behavior. In particular, I will explore how variations in performance feedback affect employee motivations, and how feedback frequency affects investment decisions. However, what exactly does causality mean? Generally, causality occurs when an event (cause, explanans) brings about another event (effect, explanandum). A causal mechanism is the configuration (event) that always (or most often) leads to another event through the properties and power of the events (Little, 2011).[1] Furthermore, it is commonly considered that the cause must precede the effect. We should also clearly separate causal relationships and correlations. If what we observe is merely a correlation, it may just be a set of events that tend to occur simultaneously or sequentially, and not one causing another. Rather, it may be from their common relation to some third variable that is the true underlying cause. For example, in this thesis I ask whether performance feedback affects productivity; however, there may be another event that represents the true underlying reason behind any observed change, and it may also be that performance feedback does not always lead to this change in productivity.

Whether or not causalities exist in the social sciences can too be considered. This question requires a very lengthy discussion that is far beyond the purpose of this subsection. The answer depends, amongst others, on whether the causal relationship needs to satisfy the property of necessity or the closely related prop-

[1] There are many other similar ways to formulate the causality definition, some of these are more conservative.

erty of lawfulness (see e.g., Hempel, 1965; Hume, 2012), which is challenging to argue for in the social sciences. However, Elster (2007) emphasizes that a causal explanation is to give an account of why it happened as it happened, which detaches causal explanations from the necessity criteria. Some philosophers use the term social mechanisms (see e.g., Little, 1991; Hedström and Swedberg, 1998) to argue for causality in the social sciences. More specifically, in complex social environments, patterns of individual behavior that have causal properties may exist, which is to say that it has the ability to produce a regular series of events. Mechanisms are often assumed to be complexities that underlie and account for aggregate social regularities (Steel and Guala, 2011). Guala (2005) writes that consensus today, in order to have more informative accounts to what it means for X to cause Y, must be possible to articulate causes and effects. This is to say that X caused Y in given circumstances. Any claims that are made about causality must be seen in such a framework. For example, in chapter three, providing subjects with relative performance feedback may both lower and increase productivity, but not necessarily. In chapter four, providing teams with relative performance feedback has the ability to increase the productivity of subjects, but not necessarily. If it does not, it may just reflect that the characteristics of the particular situation have changed.

This thesis consists of one natural quasi-experiment from the field and three randomized laboratory experiments. Why I have used experiments to investigate social casual effects may be questioned. The major benefit of experiments, in contrast to many other empirical strategies, is that they explicitly manipulate the cause, making it easier to identify the effect and eliminate disturbances (Guala, 2005). Some consider experiments as the strongest tool to infer causality in the social sciences (Shadish et al., 2002; Christensen, 2004). Pearl writes that "this is the only scientifically proven method of testing causal relations from data, and to this day, the one and only causal concept permitted in mainstream statistics" Pearl (2000, p. 340). A causal effect in experiments is considered to be the difference in outcome of being exposed to some treatment and not being exposed. If the treatment can be properly identified, the effect (difference in outcome) of the cause (treatment) can be measured. The obvious problem is that we cannot observe the same event simultaneously under two different conditions. Therefore, the treated event is compared to a counterfactual event. Only the

presence of the treatment is allowed to vary across treated and non-treated (Heckman, 2008). This is challenging as the social context is continuously changing. A completely unchanged social context is unrealistic and counterfactuals are therefore considered as similar to the treated as possible (Shadish et al., 2002).

Experiments allow for randomization which is a key inherent property. Randomization of experiments means that samples of subjects are drawn from a population and then are randomly divided into treatment or control groups. Random assignment should provide unbiased estimates of the average treatment effect, instead of trying to control all extraneous variables (Dane, 2010). If properly conducted, and the sample size is large enough, randomization will make the characteristics of the two groups close to equal, or probabilistically similar Shadish et al. (2002). Any differences in outcome between the treated and control groups are then likely to have been caused by the treatment, and not by any other correlated background variable (Guala, 2005). Hence, randomized experiments have, compared to alternative empirical strategies, less of a challenge to convincingly argue that there are no other underlying reasons (correlated variables) behind the identified effects. The randomization process, and the strict isolation (control) of the difference between treatment and control, ensures that this is less likely.

Another benefit of experiments is that, through the manipulation of causes, one can be more certain that the cause actually leads to the effect, and not the other way around. Hence, the experimental approach is suited to avoid the detection of a reversed causality. Furthermore, in the framework of understanding causalities as regularities rather than laws, experiments are great at providing statistical evidence. By having a large enough sample, one can statistically show that the treatment regularly provides a difference in outcome compared to the control. For example, in chapter two we show that the treated stores on average increase sales (a regularity) compared to the non-treated counterfactual stores, however the sub-sample analysis shows that not all stores benefited much from the treatment. Hence, the casual argumentation in the social sciences is based on a regularity statement. In chapter two, this regularity statement could be that increasing autonomy in the workplace has the ability to regularly produce higher sales.

Causes from experiments typically rely on the INUS condition, meaning that

they are insufficient but non-redundant parts of an unnecessary but sufficient condition (Shadish et al., 2002; Guala, 2005). Insufficient, as any cause from an experiment cannot alone create the effect. Non-redundant, as the cause makes a difference, it adds something to the situation. Unnecessary, as other factors could create the same effect. Sufficient, as they can be used together with the full context to create the effect. For example, in chapter two, the more detailed instructions on how to act in the workplace is not alone sufficient to increase productivity. However, it constitutes a real change in the organization; other factors could too potentially affect productivity to a similar extent, but more detailed instructions, together with the context in the workplace, could potentially increase productivity.

Despite the good arguments to conduct experiments as previously discussed, experiments certainly have some challenges too. Pure randomization is not always easy. The process itself could be problematic. For example, pressure from third parties may not allow for perfect randomization (political interests, management in firms get involved, etc.), lab experiments allow people to sign up on any available slot (which may lead to differences between groups), and so on. Moreover, the social sciences are affected by the subjects' personal experiences and endless varieties of social contexts. Whether randomization effectively avoids controlling for all extraneous variables and purely identifies causalities, may still be debated. The randomization in so-called "randomized experiments" may still suffer from the fact that the nature of the randomization may affect participants' behavior, or there may be imperfect compliance because of the existence of control status (Pearl, 2000). The most optimal randomization occurs in the field when participants are not aware of their participation in an experiment, and where treatment and control status are perfectly randomized. However, such a design is likely to have less experimental control (compared to a lab experiment where control is considered higher).

There are several other limitations to experiments. Manipulation of the cause may be impossible to conduct. Moreover, the experiment may not necessarily give an answer as to why the effect occurred, and it is often very context specific and hard to generalize (Shadish et al., 2002; Guala, 2005). Another common critique to experiments is the question of external validity. For example, the laboratory experiments in this thesis may have sufficient internal validity, but it is

less certain that the external validity holds. In chapter four, subjects seem to positively respond to team relative performance feedback in an abstract setting within a computer lab, but what about the similar situation in the field? Experiments also have strict moral and legal constraints. For example, they are costly to run, require solid cooperation with participants (firms, government, etc.) that may have their own agenda, and so on. Therefore, research allowing for self-selection or non-randomization of treatment could in some instances be preferable (Pearl, 2000; Angrist and Pischke, 2008).

Quasi-experiments are those that do not randomize into treatment or control conditions. Instead the aim is to construct control groups that are as similar as possible to the treated group. As the social context and properties vary across experiments, exact replications are impossible. However, the use of similar experiments to provide replications enable us to move towards a causal understanding of the phenomenon. To be able to draw causal inference, the design must satisfy the basic requirements for all causal relationships. Manipulating the treatment and statistical analysis ensures that the cause precedes the effect and that they covariate. The challenging part is to rule out alternative explanations of the effect (Shadish et al., 2002). Angrist and Pischke (2008) seeks experiments that mimic a randomized trial to exploit cheaper and more readily available sources of variation. It may also be that the decision to evaluate the treatment is made after being implemented, such that randomization is implausible (Bingham and Felbinger, 2002).

One commonly used analysis in quasi-experiments is the difference-in-differences (DD) approach, which is partly the empirical strategy in chapter two of this thesis. The first difference is the difference in the average outcome variable before and after the treatment, i.e., the difference in sales for treated stores before and after the time of treatment. This difference is likely biased as some unobserved characteristic correlates with the treatment status and the outcome variable. The second difference is that in the average outcome variable before and after the "treatment" of control subjects, i.e., the difference in sales for control stores before and after the time of treatment. Combining these differences cancels out common trends in the outcome variable and the effect of unobserved variables. The following model illustrates the difference in the outcome variables across the treated and control group:

$$\Delta Y_{i,t} = \beta_0 + \delta d_{i,t} + a_i + \lambda_t + \varepsilon_{i,t},$$

Where $Y_{i,t}$ is the outcome variable for entity i in period t, $d_{i,t}$ is an indicator of whether the entity is treated or not and $\varepsilon_{i,t}$ is the idiosyncratic error. δ is the measured effect of being treated (the effect of the cause), the variable of interest. All unobserved effects on the outcome variable that are time invariant for the entity (a_i), and all effects over time that are common to all entities (λ_t), are essentially differenced out. This is known as the fixed effect (Angrist and Pischke, 2008). There are two main assumptions of the DD-approach; the trend between treated and control entities would have been similar in the absence of the treatment, and that no other event systematically occurs only to one of the groups (Blundell and Dias, 2009).

The DD-approach may also be extended to a triple difference approach (DDD-approach). There may be calendar effects that differ across treated and control entities, violating the first assumption in the DD-approach, thereby biasing the estimated treatment effect. The DDD-approach addresses this concern by controlling for differential calendar effects across treated and control entities. By having sufficiently many observations prior to the treatment period, one can estimate such common differential calendar effects in the absence of the treatment, and control for this in the overall analysis. The first assumption of the DDD-approach is therefore slightly different from the DD-approach; the differential trend between treated and control entities would have been similar in the absence of the treatment. The second assumption remains unchanged. There is a more detailed explanation and discussion on this in the empirical strategy section of chapter two.

To summarize, I have highlighted beneficial properties that experiments have, which makes them a strong option as an empirical strategy to unveil causalities. It enables manipulation of the cause, ensuring the right direction of the causal relationship, and makes it easier to study the precisely defined relationships of interest. Moreover, through the statistical power of randomization, experiments have less challenges with correlated events than alternative empirical strategies.

References

Angrist, J. and Pischke, J. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Baer, M., Leenders, R. T. A., Oldham, G. R., and Vadera, A. K. (2010). Win or lose the battle for creativity: The power and perils of intergroup competition. *Academy of Management Journal*, 53(4):827–845.

Benartzi, S. and Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics*, 110(1):73–92.

Bingham, R. and Felbinger, C. (2002). *Evaluation In Practice: A Methodological Approach, 2nd Edition*. CQ Press.

Birkinshaw, J. (2001). Why is knowledge management so difficult? *Business strategy review*, 12(1):11–18.

Blundell, R. and Dias, M. C. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3):565–640.

Caroli, E. and Van Reenen, J. (2001). Skill-biased organizational change? Evidence from a panel of British and French establishments. *Quarterly Journal of Economics*, 116(4):1449–1492.

Christensen, L. (2004). *Experimental Methodology*. Allyn and Bacon.

Dane, F. (2010). *Evaluating Research: Methodology for People Who Need to Read Research*. SAGE Publications.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, 18(1):105–115.

Deci, E. L. and Ryan, R. M. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67.

Elster, J. (2007). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Eriksen, K. W. and Kvaløy, O. (2010). Do financial advisors exhibit myopic loss aversion? *Financial Markets and Portfolio Management*, 24(2):159–170.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding- out. *The American Economic Review*, 87(4):746–755.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge University Press.

Hackman, J. R. and Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(2):250–279.

Haigh, M. S. and List, J. A. (2005). Do professional traders exhibit myopic loss aversion? an experimental analysis. *The Journal of Finance*, 60(1):523–534.

Heckman, J. J. (2008). Econometric causality. *International statistical review*, 76(1):1–27.

Hedström, P. and Swedberg, R. (1998). *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press.

Hempel, C. (1965). *Aspects of scientific explanation: and other essays in the philosophy of science*. Free Press.

Hume, D. (2012). *A Treatise of Human Nature*. Dover Philosophical Classics. Dover Publications.

Ichniowski, C. and Shaw, K. (1999). The effects of human resource management systems on economic performance: An international comparison of US and Japanese plants. *Management Science*, 45(5):704–721.

ICI (2016). *Investment Company Fact Book*. Investment Company Institute (ICI), www.icifactbook.org.

Lazear, E. P. and Gibbs, M. (2014). *Personnel Economics in Practice*. Wiley-Blackwell. Wiley.

Little, D. (1991). *Varieties of Social Explanation: An Introduction to the Philosophy of Social Science*, volume 103. Westview Press.

Little, D. (2011). Causal mechanisms in the social realm. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*. Oxford University Press.

Marino, A. M. and Zabojnik, J. (2004). Internal competition for corporate resources and incentives in teams. *RAND Journal of Economics*, 35(4):710–727.

Mehra, R. and Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145–161.

Milgrom, P. and Roberts, J. (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics*, 19(2):179–208.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Steel, D. and Guala, F. (2011). *The Philosophy of Social Science Reader*. Routledge.

Tambe, P. and Hitt, L. M. (2012). The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, 23(3):599–617.

Thaler, R. H., Tversky, A., Kahneman, D., and Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics*, 112(2):647–661.

The Effect of Less Autonomy on Productivity in Retail: Evidence from a Quasi-Natural Field Experiment*

William Gilje Gjedrem[1] and Mari Rege[2]

Abstract: This paper investigates a causal relationship between autonomy and productivity in retail, utilizing store level weekly sales data from a large consumer electronics retail chain in Norway. In 2011 the retail chain made it a mandatory part of the job instruction to approach every customer who entered the store. To ensure compliance, the chain also adopted a system for feedback and monitoring. Critical to our empirical strategy, this change in management practice was introduced in some stores only and at different points in time. This allows us to estimate the effects of the change on productivity in a quasi-natural field experiment using a triple-difference approach. We find that the change in management practice increased sales by 4.3 percent and transactions by 3.3 percent. The effect seems to be persistent, suggesting that a more detailed job instruction, combined with systematic feedback and control, may increase productivity in low-skilled narrowly defined jobs.

*We are grateful to Alexander Cappelen, Robert Dur, Venke Furre Haaland, Christine Harbring, Ola Kvaløy, John List, Scott Shane, Bertil Tungodden, Mark Votruba, and a number of seminar participants for helpful comments and suggestions. Financial support from the Norwegian Research Council (227004) is gratefully acknowledged. We are also grateful to the consumer electronics retail chain studied in this paper, and its consulting firm, for providing us with data and treatment information.

[1]University of Stavanger, 4036 Stavanger, Norway. E-mail: william.g.gjedrem@uis.no
[2]University of Stavanger, University of Oslo, ESOP. E-mail: mari.rege@uis.no

1. Introduction

An important question in personnel economics is how to efficiently delegate decision-making within a firm (Lazear and Gibbs, 2014). Should the manager make most of the decisions for consistency and control, or should the manager delegate the decisions in order to let the employees exploit the specific knowledge of time and place? Many studies suggest that decentralized decision-making is important for productivity, innovation and motivation in high-skilled complex jobs (e.g., Hackman and Oldham, 1976; Milgrom and Roberts, 1995; Ichniowski and Shaw, 1999; Caroli and Van Reenen, 2001; Tambe and Hitt, 2012). Notably, however, there are still a lot of low-skilled and narrowly defined jobs. For these jobs, figuring out best practice, and have all employees following best practice, may increase both productivity and motivation (Lazear and Gibbs, 2014).

This paper investigates a causal relationship between autonomy and productivity in a low-skilled and narrowly defined job. Specifically, we investigate a change in management practice to more detailed job instructions, in addition to more systematic control and feedback, for sales staff in a large consumer electronics retail chain in Norway. A crucial decision sales staff has to make, many times a day, is whether or not to approach a customer who enters the store. Some customers appreciate immediate contact with sales staff, whereas others prefer to be left alone, and sometimes their preferences are signaled by body language. Prior to the change in management practice, the managers had identified that the sales staff who followed a very simple rule-of-thumb; make contact with every single customer who enters the store – ignoring any signals from the customer – had higher sales than those who did not follow this strategy. Still, it was a challenge to induce the majority of employees to follow this simple rule-of-thumb. It had not been sufficient to train and encourage the employees, nor did monetary incentives in terms of sales bonuses suffice. These observations, in addition to low sales figures, prompted a change in management practice in 2011: The retail firm made it an obligatory part of the job instruction to make contact with every single customer who enters the store. Moreover, to ensure compliance and validation, they adopted a feedback and monitoring technology.

There are many behavioral mechanisms through which this change in man-

agement practice may affect sales. On the one hand, the change could decrease sales because the sales staff is no longer allowed to utilize his or her specific knowledge about the individual customer, or because the salespersons are feeling controlled and monitored, which leads to lack of motivation or higher turnover (Hackman and Oldham, 1976; Spector, 1986). On the other hand, there are several mechanisms through which this change in management practice may increase productivity, by forcing everybody to adhere to best practice. In a behavioral model in which it is costly to make decisions, the decrease in autonomy could make the sales staff more effective because the simple rule-of-thumb allows them to spend less time and energy on making decisions (Simon, 1955; Tversky and Kahneman, 1973). Moreover, in a behavioral model with time-inconsistent preferences (Akerlof, 1991; Loewenstein and Prelec, 1992; Laibson, 1997), the decrease in autonomy may help the employees fight procrastinating behavior. A salesperson may often feel that it is uncomfortable to approach some of the customers – especially those who look like they prefer to be left alone – and decide that she is not up for it today, even if she knows that making contact with the customer may increases sales and thereby her future earnings. With the simple rule-of-thumb and the monitoring technology, such procrastinating behavior is more difficult to carry out. Finally, in a behavioral model in which individuals care about positive feedback from the management (Ellingsen and Johannesson, 2007; Kosfeld and Neckermann, 2011), or care about doing the right thing (Andreoni, 1990; Coleman and Coleman, 1994), the change in management practice could make the sales staff more effective by increasing their motivation. The change has made the "right thing to do" well defined: Approach every customer. As long as a salesperson is doing this, he can experience an intrinsic reward of feeling that he is doing the right thing and gaining the management's approval through the monitoring and feedback system.

To examine the effect of the change in management practice on productivity, we exploit the fact that the change was only introduced in some stores and at different points in time. This allows us to estimate the effects in a quasi-natural field experiment using a triple-difference model. Identification of the treatment effect arises from differential changes in sales in treated stores relative to control stores, in weeks before and after treatment introduction, during the year of treatment introduction (Treatment Year), relative to the same double-difference

during the previous year (Control Year). The crucial identifying assumption in our triple-difference approach is that differential changes in sales across treated and non-treated stores are identical in Control and Treatment Year in the absence of treatment. There are several reasons why this may not be true. For example, treated and non-treated stores may experience different trends in sales because they have a different customer base or focus on different products. Importantly, the long time horizon in the data set allows us to run Placebo tests investigating the validity of this assumption. The empirical results suggest that the change in management practice to more detailed job instructions increased sales by 4.3 percent and transactions by 3.3 percent. The effects seem to be increasing and persistent; measuring 9 percent for sales and 11 percent for transactions after 25 weeks. This indicates that a more detailed job instruction, based on best practice, may increase productivity in low-skilled narrowly defined jobs.

This paper relates to several strands of literature. The idea of figuring out best practice through industrial engineering, and have everyone do it that way is the essence of Taylorism, originating from the book 'The Principles of Scientific Management' written by the US industrial engineer Frederick Winslow Taylor (Taylor, 1911). Taylorism may seem outdated as jobs are becoming more and more knowledge-intensive and complex. As noted above, several papers suggest that in high-skilled complex jobs decentralized decision-making is important for productivity, innovation and motivation. However, there are still a lot of low-skilled and narrowly defined jobs, and hence, important to better understand the relationship between autonomy and productivity in these types of jobs.

Importantly, this paper adds to the strand of literature, in the intercept between behavioral and empirical labor economics, investigating causal effects of different human resource management practices, utilizing data from the field (for reviews see e.g., List and Rasul, 2011; Bandiera et al., 2011; Bloom and Van Reenen, 2011). For example, Hamilton et al. (2003) demonstrate that the introduction of team incentives in a large textile company improves worker productivity; Gneezy and List (2006) demonstrate, both in the contexts of data entry and door-to-door fundraising, that employees reciprocate a higher wage with greater effort during the early hours of the task, but the effect is not persistent; Blanes i Vidal and Nossol (2011) demonstrate that introducing relative performance feedback led to a large and long-lasting increase in productivity

for workers picking up customer orders;[1] Hossain and List (2012) demonstrate that the productivity of workers and teams of workers in a high-tech manufacturing facility respond to the framing of incentives; Bradler et al. (2016) and Kvaløy et al. (2015) demonstrate, in the context of data entry, that unannounced, public recognition on employee performance and motivational talk can increase productivity; and Kosfeld and Neckermann (2011) demonstrate, also in the the context of data entry, that offering a congratulatory card from the organization honoring the best performance have a large effect on productivity.[2] Our paper contributes to this literature by demonstrating that a change in management practice, to more detailed job instructions, in addition to more systematic control and feedback, can increase productivity.

2. Treatment and Hypotheses

2.1 The Consumer Electronics Retail Chain

We investigate effects of a change in management practice to more detailed job instructions, in addition to more systematic control and feedback, in a large consumer electronics retail chain in Norway (hereby referred to as CE). The change in management practice took place in 2011, and at the time CE was one of the leading distributors of consumer electronics in Norway, with a market share of approximately 30 percent. As of April 2012 CE had approximately 1500 employees and consisted of 166 stores, of which 61 were self-owned and 105 were franchised stores, in addition to an online store.[3] CE was facing sharp competition with other consumer electronics chains and an increasing number of online stores, and their change in management practice was prompted by the fact that the development in annual revenues had not been satisfactory.

[1] See also Berger et al. (2013), Bandiera et al. (2013), Delfgaauw et al. (2013), Delfgaauw et al. (2014) and Ashraf et al. (2014) for examples of evidence on tournament incentives.

[2] See also the investigation of how management practice matters for productivity in Bloom et al. (2013).

[3] We have data for all 61 self-owned stores and 60 franchised stores. As CE was not in charge of the financial reporting of the remaining franchised stores, we do not have access to data on these. 4 self-owned stores opened during the last parts of 2011, and did not conduct a change in management practice (they opened as treated), therefore only 57 self-owned stores are relevant for the analysis.

The change in management practice, which we in the following will refer to as treatment, was introduced in all stores that were self-owned by CE, henceforth referred to as treated stores. The other stores did not undertake the change, and are henceforth referred to as control stores.[4] Both treated and control stores were geographically located throughout Norway at various locations in city centers and shopping malls. Customers were supposed to get the same experience regardless of which type of store they visited. They all offered the same electronic brands and products, had the same weekly offers, and benefited from the centralized branding of CE.

Notably, the change in management practice did not affect the employee's monetary incentives. Store managers and division leaders had a basic wage and a bonus system dependent on store sales. The bonus was based on the the actual performance of the store/division, relative to performance targets for costs and sales revenues, and could more than double the wage. As such, CE had strong monetary incentives at the store management level to improve sales and cut costs. The sales staff received a tariff wage, in addition to a bonus dependent on store and individual sales. Every second month the aggregated sales in a store was compared to the store budget and some of the possible surplus was allocated to the salespersons. One quarter of this amount was split equally between all salespersons, and three quarters was distributed based on each individual salesperson's sales record. In addition, the sales staff received a commission on insurance sales.

2.2 Treatment

In an evaluation of their operative activities in 2010 the management of the self-owned CE stores made three key observations. 1) Many customers leave the store without buying anything (76 percent), 2) There are large differences in sales performance among the salespeople, and 3) A survey of the salespeople revealed that the best performers focus on establishing some kind of contact with every single customer. The management had regularly encouraged, trained and coached salespeople to approach the customers entering the store. However, the evaluation demonstrated clearly that in practice this encouragement and coach-

[4]The online store is excluded from our analyses.

ing did not suffice. As such, CE made approaching every customer who enters the store an explicit and obligatory part of the job instruction in their self-owned stores.

To ensure compliance to and validation of the new job instruction, the retail chain also adopted a system for feedback and monitoring. Specifically, every time a salesperson had contact with a customer, she was supposed to click once on a device she was carrying in her pocket, called a clicker. A customized software program combined information from the clicker with the information from a customer counter at the entrance. First and foremost, this software gave the salespeople daily information about the "click rate", which was the share of customers entering the store with whom salespeople had been in contact. The salespeople were expected to have a hundred percent click rate.[5] The software also provided information about sales and the hit rate, which was the share of customers entering the store who actually bought something. All this information was conveyed in a stylized graphical sales report as illustrated in Figure 1.

The report was utilized every morning in a 15 minute morning pep talk meeting for all employees. Here the management gave the salespeople feedback on performance for click rate, hit rate and sales in comparison to performance goals and performance in other CE stores. The purpose was to let the salespeople see in retrospect how the strategy of more actively approaching customers increased their sales figures. In addition to the feedback during the morning meetings, the store manager met weekly with each salesperson to provide individual performance feedback.

This research project started several years after treatment introduction. At the time when the treated stores made the above changes in management practice, they were not a part of any study. As the decision to undertake this study happened ex-post, a Hawthorne effect or demand induced effect is not a major concern. One could, however, still imagine other reasons why the treatment only has short run effects; for example, the employees think the clickers and new technology are cool in the beginning, but then the excitement tapers of. As we will see below, our analysis measures performance several months after treatment introduction.

[5]The average click rate of all treated stores was 120 percent.

2.3 Timing of Treatment

The treatment was introduced at different times between week 25 and 38 of 2011 for the 49 self-owned stores in our sample.[6] The timing of treatment is illustrated in Figure 2. The treatment introduction was stretched over time due to stores' limited capacity during summer holidays and capacity constraints of regional managers to participate on the day of the treatment introduction. As such, the timing of treatment among the self-owned stores was not randomized; it was an administrative decision based on practicalities. Moreover, treatment was also not randomized, as it was based on CE ownership. Nevertheless, the fact that the treatment was only introduced in self-owned stores, and at different times, allows us to address selection in a quasi-experimental design utilizing a triple-difference approach. This will be carefully described in Section 3.2 Empirical Strategy.

2.4 Hypotheses

The key decision sales staff has to make, many times a day, is whether or not to approach a customer who enters the store. Some customers appreciate immediate contact with sales staff, whereas others prefer to be left alone, and sometimes this is signaled by body language. In crude words, the treatment is telling the sales staff to ignore these signals and approach every single customer.

There are good theoretical arguments both for a negative and for a positive treatment effect on sales. On the one hand, the treatment could decrease sales because the sales staff is no longer allowed to utilize his or her specific knowledge about the individual customer. For example, a customer may be signaling with her body language that she absolutely wants to be left alone to look at product displays, and approaching this customer may result in a lost purchase because the customer leaves the store in annoyance. Also, the treatment could decrease sales because the salespersons feel controlled and monitored. Several studies suggest that a reduction in autonomy can decrease people's motivation (Hackman and Oldham, 1976; Spector, 1986).

[6] We exclude the self-owned stores that participated in piloting the treatment to ensure a clean definition of treatment. See Section 3.1 for sample selection.

On the other hand, there are several mechanisms through which the treatment could increase productivity, by forcing everybody to adhere to best practice. In a behavioral model in which it is costly to make decisions, the treatment could make the sales staff more effective because they spend less time and energy on making decisions (Simon, 1955; Tversky and Kahneman, 1973). Instead of looking at customers for signals of whether or not they want help, the sales staff follows the simple rule-of-thumb; make contact with every single customer who enters the store. This way they spend their time and energy on helping customers, instead of trying to decide whether to offer help.

Moreover, in a behavioral model with time-inconsistent preferences (Akerlof, 1991; Loewenstein and Prelec, 1992; Laibson, 1997), the treatment may help the employees fight procrastinating behavior. A salesperson may often feel that it is uncomfortable to approach some of the customers – especially those who do not seem friendly or seem to prefer to be left alone. Then, if the salesperson cares disproportionally more about what is happening right now, compared to what is happening in the future, she may decide that she is not up for it today, even if she knows that making contact with the customer likely increases sales and thereby her future earnings. Instead, she hopes to start approaching customers more actively the next day, when she hopefully feels more like interacting with people. With the simple rule-of-thumb and the monitoring technology, such procrastinating behavior is no longer possible.

Finally, in a behavioral model in which individuals care about positive feedback from the management (Ellingsen and Johannesson, 2007; Kosfeld and Neckermann, 2011), or care about doing the right thing (Andreoni, 1990; Coleman and Coleman, 1994), the treatment could make the sales staff more effective by increasing their motivation. The treatment has made the "right thing to do" well defined: Approach every customer. As long as a salesperson is doing this, she can experience an intrinsic reward or a "warm glow" of feeling she is doing the right thing, which is even emphasized by a click on her clicker. Moreover, she will feel the approval from management through the extensive monitoring and feedback system set up to reinforce the message of approaching every customer.

As there are good theoretical reasons both for negative and positive treatment effects on sales, we do not have hypotheses for how we expect the treat-

ment to affect sales. Our estimated treatment effect will give us the net effect of possibly many different mechanisms.

3. Data and Empirical Strategy

3.1 Data and Summary Statistics

CE has given us access to weekly store level sales data from week 1 in 2009 until week 52 in 2012 for all CE stores. It is the total gross weekly sales revenue of the store, without considering profit margins or taxes. The sales data has high reliability, as it is collected from the same source as the stores' financial reporting system, and is subject to certain legal requirements and an annual inspection by auditors. In addition to the sales data, we also have access to weekly number of transactions. This is the weekly number of purchases in the store. If a customer buys several goods in one purchase, it is recorded as one transaction. If the customer first purchases some goods, and then decides to purchase some other goods, it is recorded as two transactions. Transaction data is only available from week 1 in 2010, and there are some missing values.

Sales is a key indicator of performance in retail; it is important for the firm's cash flow and profitability. Notably, however, sales does not transform linearly to profitability. For example, sales staff may be able to sell more of lower priced items (e.g. HDMI cord) with relatively large profit margins, rather than expensive products with relatively low profit margins (e.g. TVs). This would barely be noticeable on the overall sales data, but still be important for the store's profitability. Therefore, including number of transactions would potentially capture something that overall sales does not, that could still tell us something about the effectiveness of the treatment. In particular, from the transactions measure we learn whether the treatment increased sales by increasing the hit rate.

In our main analysis we only utilize data up until week 5 of 2012, as CE in the spring of 2012 reorganized and closed many of the self-owned stores. As such, we define the year starting at week 6 in 2011 to week 5 in 2012 as the Treatment Year. This allows a large observation window of treated stores both before and after the treatment introduction during the weeks 25-38 of 2011.[7] To

[7] In Table 5 we investigate robustness to this definition of Treatment Year.

control for differential calendar effects across treated and non-treated stores in a triple-difference approach, we define the year prior to the Treatment Year (week 6 in 2010 to week 5 in 2011) as the Control Year.

To assure a clean definition of treatment and control, we exclude 8 self-owned stores that participated in piloting the treatment, and 9 stores that were not self-owned, but adopted parts of the treatment. Moreover, we exclude 9 stores that closed during Control or Treatment Year, or within 6 weeks after end of Treatment Year; and we exclude the first 3 weeks of observations for stores that opened during Control or Treatment Year. Finally, we exclude 3 non-treated stores missing all but a few observations on transactions.[8]

Making these restrictions, we are left with 49 treated stores and 39 control stores, providing us with total observations of 8905 on sales and 8730 on transactions. Summary statistics for these stores are provided in Table 1 (inflation adjusted to 2011 Norwegian kroner). We can see in Panel A that the average weekly sale is about Norwegian kroner 546K in treated stores, and about 351K in control stores. For transactions, the corresponding figures are 483 and 334, respectively. Panel B provides summary statistics for the Control Year only, allowing a comparison of treated and non-treated stores prior to treatment. In the last column we can see that, prior to treatment, transactions and sales are substantially larger on average in treated stores compared to control stores.

In Figures 3 and 4 we illustrate the development in sales and transactions during Control and Treatment Year for treated and non-treated stores. We can see substantial calendar effects, and even if the lines for treated and non-treated stores often move in parallel, this is not always consistent. The grey area in Figures 3 and 4 marks the period of treatment introduction. Due to the calendar effects, it is hard to spot any treatment effects with the naked eye. The triple-difference approach, carefully outlined in the next section, will control for store, week and year fixed effects, in addition to differential calendar effects across treated and non-treated stores. Additionally we add controls for different time trends across store size and store location. We define store size based on average weekly sales volume up until two weeks prior to the first stores being treated, and categorize them into three tertiles. We define a store to be located in a mall if the store is located in the same building as other stores, does not have its own

[8] In Table 6 we investigate robustness to all these sample restrictions.

designated parking area and has no separate entrance directly from the outside of the building.[9]

In addition to sales and transactions data, we have data on the hit rate for treated stores. In Figure 5 we illustrate the development in the hit rate for treated stores in Treatment and Control Year. We can see that during our study period between 20 and 30 percent of the customers who enter the store end up actually purchasing something. Moreover, the hit rate was larger in the Treatment Year than in the Control Year, and particularly so after treatment introduction. This is consistent with a positive treatment effect on performance, but it could also be changing trends. Unfortunately, as we do not have hit rates for the control stores, we are not able to use the hit rate as an outcome in our triple-difference analysis.

3.2 Empirical Strategy

To explain our empirical strategy, assume first that we only utilize data from Treatment Year, and consider the following difference-in-differences model for log sales in store i in week w ($sales_{i,w}$):

$$sales_{i,w} = \alpha + \beta treatment_{i,w} + store_i + week_w + \varepsilon_{i,w} \quad (1)$$

where $treatment_{i,w}$ is an indicator for whether or not store i is treated in week w; $store_i$ is a vector of store fixed effects; $week_w$ is a vector of week fixed effects (52); and $\varepsilon_{i,w}$ is the error term. The vector $store_i$ controls for time-invariant observable and unobservable store characteristics, as for example number of parking spots outside the store, the location of the store, store size and friendliness of staff. The vector $week_w$ controls for store-invariant time characteristics, such as the Christmas season, macro economic demand shocks and marketing campaigns.

Identification of the treatment effect β in Equation (1) arises from differential change in sales in treated stores relative to control stores before and after treatment. Estimates of β produced under Equation (1) are undermined if calendar effects differ across treated and non-treated stores. For example, if

[9] About 33 percent of the non-treated stores and 43 percent of the treated stores are located in a mall.

Christmas season increases sales more in treated than non-treated stores, this would bias the estimates. To address this concern our empirical strategy applies a triple-difference approach, controlling for differential calendar effects across treated and non-treated stores. To do this we utilize data from both Control and Treatment Year and estimate the following model for log sales in store i in week w and year y ($sales_{i,w,y}$):

$$\begin{aligned} sales_{i,w,y} &= \alpha + \beta treatment_{i,w,y} + store_i + week_w + year_y \\ &\quad + week_w * year_y + week_w * treat_i + year_y * treat_i + \varepsilon_{i,w,y} \end{aligned} \quad (2)$$

where $year_y$ is an indicator for Treatment Year. Notably, Equation (2) controls for differential calendar effects across Control and Treatment Year by including the interaction term $week_w * year_y$, and across treated and non-treated stores by including the interaction terms $year_y * treat_i$ and $week_w * treat_i$, where $treat_i$ is an indicator for whether or not store i is a treated store. As such, identification of the treatment effect β in Equation (2) arises from differential change in sales in treated stores relative to control stores in weeks before and after treatment introduction during the Treatment Year, relative to the same double-difference during the Control Year. The crucial identifying assumption in our triple-difference approach is that differential changes in sales across treated and non-treated stores are identical in Control and Treatment Year in the absence of the treatment. This may not be true if, for example, treated and non-treated stores experience different trends in sales because they have a different customer base or focus on different products. To address this concern, we control for different time trends across store location and store size and estimate the following model:

$$\begin{aligned} sales_{i,w,y} &= \alpha + \beta treatment_{i,w,y} + store_i + week_w + year_y \\ &\quad + week_w * year_y + week_w * treat_i + year_y * treat_i \\ &\quad + week_w * size_i + year_y * size_i + week_w * year_y * size_i \\ &\quad + week_w * mall_i + year_y * mall_i + week_w * year_y * mall_i \\ &\quad + \varepsilon_{i,w,y} \end{aligned} \quad (3)$$

where $size_i$ is a vector of store size fixed effects $(MediumSize, LargeSize)$, and

$mall_i$ is an indicator for whether or not store i is located in a mall. Now our identifying assumption is that differential changes in sales across treated and non-treated stores – not due to different time trends across store location and store size – are identical in Control and Treatment Year in the absence of the treatment. Importantly, the long time horizon in the data set allows us to run a Placebo test investigating the validity of this assumption.

4. Results

4.1 Main Results

In Table 2, we investigate the effect of treatment on sales (Panel A) and transactions (Panel B). First, in Column 1 we only utilize data from Treatment Year and estimate the double-difference model in Equation (1). The estimates suggest that the treatment leads to a 4.4 percent increase in sales and a 5.6 percent increase in transactions. As discussed in Section 3.2, the estimates in Column 1 are biased if calendar effects differ across treated and non-treated stores. In Column 2, we utilize data from Treated and Control Year and estimate the triple-difference model in Equation (2). We can see that controlling for differential calendar effects across treated and non-treated stores somewhat reduces the estimates. We also discussed in Section 3.2 that the estimates in Column 2 are possibly biased if treated and non-treated stores experience different trends in sales because they have a different customer base or focus on different products. In Column 3 and 4 we address this concern by controlling for different time trends across store size and location. Column 4 corresponds to Equation (3). We can see that controlling for differential trends somewhat increases the estimate on sales, but does not change the magnitude of the estimate on transactions.

Finally, in Column 5 we investigate the plausibility of the identifying assumption in a Placebo analysis. As discussed in Section 3.2, the estimates in Column 4 are reliable only if differential changes in sales across treated and non-treated stores – not due to different time trends across store location and store size – are identical in Control and Treatment Year in the absence of treatment. In the Placebo analysis we move all the sample criteria and treatment

definitions one year back. Then, since the "Treatment Year" is now prior to the treatment introduction, we should not see any treatment effect, unless our treatment effect is picking up diverging trends between treated and non-treated stores (not due to different time trends across store location and store size). Consistent with our identifying assumption, the Placebo analysis demonstrates a very small and insignificant treatment effect. Unfortunately, we do not have transaction data sufficiently back in time to do Placebo analysis with transactions as a dependent variable.

In Table 3 we investigate differential treatment effects across store size (measured in sales volume), treatment compliance and store location. We can see in Column 1 that, due to power issues, there are no significant differences across store size, but the results are suggestive of larger treatment effect in stores with smaller sales volume. In Column 2 we can see that there is a large and significantly different treatment effect on transactions between stores located in malls and stores not located in malls. Indeed the treatment effect seems to be entirely driven by the stores located in malls. In terms of sales the estimate goes in the same direction, but it is not significant. Stores located in malls are likely to have more customers stopping by to peek without a carefully planned intention because the time cost of stopping by is so low. One possible explanation for the differential treatment effects estimated in Column 2 is that the treatment succeeds in making transactions with these customers.

In Column 3 we have created two indicators for medium (MediumCompliance) and high (HighCompliance) compliance if treatment compliance scored in the second or third tertile, respectively. A store compliance score was calculated as the mean score of a weekly quality survey during the treated period. The purpose of the quality survey was to evaluate the treatment and ensure compliance at the management level. The survey assessed the quality of morning meetings, morning routines, and sales observations, and had to be filled in by the store manager. In Column 3, we can see that the treatment only had a significant effect in stores with high compliance rating. The estimates suggests that the treatment increased the sales by 7 percent in these stores. This estimate must, however, be interpreted with caution, as treatment compliance is possibly an endogenous variable.

In Table 4 we investigate differential treatment effects across time. As dis-

cussed in Section 2.2, a potential concern is that the treatment only has short run effects if there is initial excitement with the clickers and new technology, which eventually tapers of. It may also be that there is a cost of the decrease in autonomy, materializing in employees quitting. Such an increase in turnover will likely not affect sales until several weeks after treatment introduction. In order to investigate persistence, we create four indicators for treatment 1-8, 9-16, 17-24, and 25-and-more weeks after treatment introduction. In Column 1 we see no evidence of the treatment effect tapering off. Indeed, the effect seems to be small in the first weeks after treatment introduction. However, after week 17 it becomes large and significant, and 25 weeks after treatment introduction sales in treated stores have increased by 9 percent and transaction by 11 percent. This suggests that there is a phase-in period after treatment introduction, in which the stores learn to use the new feedback and monitoring system, and employees learn effective strategies for approaching every customer.

In Column 2 of Table 4 we introduce a second type of Placebo analysis in order to investigate the plausibility of our identifying assumption. We do this by creating two indicators for treatment 1-8 and 9-16 weeks *prior* to treatment introduction. If our identifying assumption is true, the estimated coefficients for these indicators should be small and insignificant, which is confirmed in Column 2.

4.2 Robustness Analyses

A key choice in our empirical strategy was the choice of Treatment and Control Year. In Table 5 we demonstrate that our results are robust to these choices. In Column 1 we present our preferred model 4 from Table 2. Then in Column 2 we shift the definition of the Treatment and Control Year 4 weeks back in time, and in Column 3 we shift it 8 weeks back. We see that the estimates change very little. In Column 4 and 5 we shift the definition of the Treatment and Control Year 4 and 8 weeks forward in time, respectively. Moving the study period forward seems to have a modest effect, in particular for transactions, which likely reflects that we are getting into the period of major reorganization and store closure that started in the spring of 2012. As discussed in Section 3.1 we chose only to utilize data up until week 5 of 2012 due to the start of this major

reorganization.

Finally, in Table 6 we demonstrate how robust our results are to the exclusion restrictions we made when constructing our sample in Section 3.1. First, Column 1 presents our preferred model from Table 2. Then Column 2 drops the exclusion restriction of 3 non-treated stores missing all but a few observations; Column 3 drops the exclusion restriction of the first 3 weeks of observations for stores that opened during Control or Treatment Year; and Column 4 drops the exclusion restriction of 9 stores that closed during Control or Treatment Year, or within 6 weeks after end of Treatment Year. We can see that our estimates of interest are very robust to dropping all the exclusion restrictions. In Column 5 we drop the exclusion restriction of 8 self-owned stores that participated in piloting the treatment. This reduces the magnitude of our estimates, the treatment effect on sales is no longer significant, and the treatment effect on transactions is only significant at the ten percent level. This may reflect that the treatment was still under development and, hence, we do not have a sharp starting date for full treatment for these stores. It could also be due to the fact that the pilot stores are very large stores, and the treatment may not be effective in large stores (see Table 3 Column 1). Finally in Column 6 we drop the exclusion of 9 stores that were not self-owned, but adopted parts of the treatment. Somewhat surprising, since we are now including partly treated stores in the control group, this increases the magnitude of our estimates. In Column 7 we drop all the exclusion restrictions, and we see that the estimates are very similar to the estimates in the preferred model in Column 1.

5. Conclusions

An important question in personnel economics is how to efficiently delegate decision-making within a firm (Lazear and Gibbs, 2014). This paper investigates a causal relationship between autonomy and productivity in a low-skilled and narrowly defined job. Specifically, we investigate a change in management practice to more detailed job instructions, in addition to more systematic control and feedback, for sales staff in a large consumer electronics retail chain in Norway. To examine the effect of the change in management practice on productivity, we exploit the fact that the change was only introduced in some stores and at

different points in time. This allows us to estimate the effects in a quasi-natural field experiment using a triple-difference model. Identification of the treatment effect arises from differential changes in sales in treated stores relative to control stores, in weeks before and after treatment introduction during the year of treatment introduction (Treatment Year), relative to the same double-difference during the previous year (Control Year). The empirical results suggest that the change in management practice to more detailed job instructions increased sales by 4.3 percent and transactions by 3.3 percent. The effect seems to be persistent, suggesting that a more detailed job instruction, based on best practice, may increase productivity in low-skilled narrowly defined jobs.

As discussed in this paper, for low-skilled and narrowly defined jobs, there are good theoretical arguments suggesting that figuring out best practice, and have all employees following best practice, may increase both productivity and motivation. Our empirical evidence is consistent with these theories. However, this research is not conclusive. First, even if we address causality very carefully in a quasi experimental design, and the Placebo analyses support our identifying assumption, we cannot completely rule out that our results are driven by differential trends between treated and not treated stores. Second, our results pertain to one large retail chain in Norway, and we do not know how this extrapolates to other samples. More research is needed to fully understand the link between employees' autonomy and productivity, and different moderators of this link, in low skilled narrowly defined jobs. Future studies should also strive to obtain measures of worker satisfaction and turnover.

References

Akerlof, G. A. (1991). Procrastination and obedience. *The American Economic Review*, 81(2):1–19.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Bandiera, O., Barankay, I., and Rasul, I. (2011). Field experiments with firms. *The Journal of Economic Perspectives*, 25(3):63–82.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Berger, J., Harbring, C., and Sliwka, D. (2013). Performance appraisals and the impact of forced distribution - An experimental investigation. *Management Science*, 59(1):54–68.

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? Evidence from India. *The Quarterly Journal of Economics*, 128(1):1–51.

Bloom, N. and Van Reenen, J. (2011). Human resource management and productivity. *Handbook of Labor Economics*, 4:1697–1767.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016). Employee recognition and performance: A field experiment. *Management Science*.

Caroli, E. and Van Reenen, J. (2001). Skill-biased organizational change? Evidence from a panel of British and French establishments. *Quarterly Journal of Economics*, 116(4):1449–1492.

Coleman, J. S. and Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Delfgaauw, J., Dur, R., Non, A., and Verbeke, W. (2014). Dynamic incentive effects of relative performance pay: A field experiment. *Labour Economics*, 28:1–13.

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.

Ellingsen, T. and Johannesson, M. (2007). Paying respect. *The Journal of Economic Perspectives*, 21(4):135–150.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Hackman, J. R. and Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(2):250–279.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, 111(3):465–497.

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Ichniowski, C. and Shaw, K. (1999). The effects of human resource management systems on economic performance: An international comparison of US and Japanese plants. *Management Science*, 45(5):704–721.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Kvaløy, O., Nieken, P., and Schöttner, A. (2015). Hidden benefits of reward: A field experiment on motivation and monetary incentives. *European Economic Review*, 76:188–199.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–477.

Lazear, E. P. and Gibbs, M. (2014). *Personnel economics in practice*. John Wiley & Sons.

List, J. A. and Rasul, I. (2011). Field experiments in labor economics. *Handbook of Labor Economics*, 4:103–228.

Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597.

Milgrom, P. and Roberts, J. (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics*, 19(2):179–208.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Spector, P. E. (1986). Perceived control by employees: A meta-analysis of studies concerning autonomy and participation at work. *Human Relations*, 39(11):1005–1016.

Tambe, P. and Hitt, L. M. (2012). The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, 23(3):599–617.

Taylor, F. W. (1911). *The principles of scientific management*. Harper & Brothers.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232.

Table 1: Summary Statistics

	N	Treated	N	Non-Treated	N	Total	Difference
Panel A: Control- and Treatment Year							
Sales	5042	546.1	3863	350.7	8905	461.3	
		(335.7)		(273.8)		(325.1)	
Transactions	4876	482.6	3854	334.2	8730	417.1	
		(270.0)		(216.0)		(258.4)	
Panel B: Control Year							
Sales	2494	538.8	1835	358.4	4329	462.3	-180.4***
		(327.3)		(283.6)		(322.1)	[9.5]
Transactions	2338	463.4	1835	335.3	4173	407.1	-128.1***
		(247.3)		(212.6)		(241.2)	[7.3]
Number of stores	49		39		88		

Notes: Standard deviation in parenthesis. Standard error in brackets. Sales is denoted in thousands. Control Year is from week 6 in 2010 to week 5 in 2011. Treatment Year is from week 6 in 2011 to week 5 in 2012. Panel A provides summary statistics for both the Control and Treatment Year, whereas Panel B provides summary statistics for the Control Year only, allowing a comparison of treated and non-treated stores prior to treatment. * p<0.10, ** p<0.05, *** p<0.01.

Table 2: Main Results: Treatment Effect on Sales and Transactions

	(1)	(2)	(3)	(4)	(5)
Panel A: Sales					
Treatment	0.044**	0.035*	0.045**	0.043**	0.004
	(0.0188)	(0.0192)	(0.0207)	(0.0214)	(0.0245)
Observations	4576	8905	8905	8905	8523
Adjusted R^2	0.589	0.592	0.594	0.594	0.602
Panel B: Transactions					
Treatment	0.056***	0.034*	0.036**	0.033*	
	(0.0196)	(0.0180)	(0.0180)	(0.0187)	
Observations	4557	8730	8730	8730	
Adjusted R^2	0.682	0.715	0.718	0.718	
Fixed effects included:					
Store	Y	Y	Y	Y	Y
Week	Y	Y	Y	Y	Y
Year		Y	Y	Y	Y
Week x Year		Y	Y	Y	Y
Week x Treated		Y	Y	Y	Y
Year x Treated		Y	Y	Y	Y
Week x Year x Size			Y	Y	Y
Week x Year x Mall				Y	Y
Periods included:					
Treatment Year	Y	Y	Y	Y	
Control Year		Y	Y	Y	Y
Placebo					Y

Notes: Dependent variable is log of weekly sales (Panel A) and log of number of weekly transactions (Panel B). Control variables are the fixed effects indicated in the table. When the triple interactions are included in Columns 3 and 4, the corresponding first order interactions are also included. OLS coefficients reported with robust standard errors in parentheses, corrected for clustering across stores. Number of observations in panel B is lower because of more missing values. We do not possess data to conduct Placebo analysis in Panel B. * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Differential Effects

	(1)	(2)	(3)
Panel A: Sales			
Treatment	0.053	0.019	0.013
	(0.0441)	(0.0265)	(0.0301)
Treatment x MediumSize	0.007		
	(0.0644)		
Treatment x LargeSize	-0.042		
	(0.0592)		
Treatment x Mall		0.059	
		(0.0508)	
Treatment x MediumCompliance			0.011
			(0.0354)
Treatment x HighCompliance			0.070*
			(0.0390)
Adjusted R^2	0.594	0.595	0.595
Panel B: Transactions			
Treatment	0.027	-0.010	0.004
	(0.0405)	(0.0250)	(0.0257)
Treatment x MediumSize	0.009		
	(0.0531)		
Treatment x LargeSize	0.009		
	(0.0541)		
Treatment x Mall		0.106***	
		(0.0373)	
Treatment x MediumCompliance			0.027
			(0.0312)
Treatment x HighCompliance			0.056*
			(0.0286)
Adjusted R^2	0.718	0.720	0.719

Notes: Dependent variable is log of weekly sales and log of number of sales transactions, in Panel A and B, respectively. OLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across stores. All regressions have the same fixed effects included as our preferred model (Table 2, Column 4) with identical number of observations.
* p<0.10, ** p<0.05, *** p<0.01.

Table 4: Persistence

	(1)	(2)
Panel A: Sales		
Pre-Treatment weeks (-16)-(-9)		-0.024
		(0.0188)
Pre-Treatment weeks (-8)-(-1)		0.005
		(0.0264)
Treatment weeks 1-8	0.024	0.022
	(0.0199)	(0.0299)
Treatment weeks 9-16	0.042	0.039
	(0.0265)	(0.0340)
Treatment weeks 17-24	0.081**	0.076*
	(0.0337)	(0.0398)
Treatment weeks 25 and longer	0.089**	0.085*
	(0.0409)	(0.0470)
Adjusted R^2	0.594	0.594
Panel B: Transactions		
Pre-Treatment weeks (-16)-(-9)		-0.003
		(0.0130)
Pre-Treatment weeks (-8)-(-1)		0.034
		(0.0209)
Treatment weeks 1-8	0.021	0.040
	(0.0171)	(0.0253)
Treatment weeks 9-16	0.029	0.044
	(0.0239)	(0.0290)
Treatment weeks 17-24	0.055**	0.069**
	(0.0265)	(0.0303)
Treatment weeks 25 and longer	0.107***	0.120***
	(0.0394)	(0.0427)
Adjusted R^2	0.719	0.719

Notes: Dependent variable is log of weekly sales and log of number of sales transactions, in Panel A and B, respectively. OLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across stores. All regressions have the same fixed effects included as our preferred model (Table 2, Column 4) with identical number of observations. * p<0.10, ** p<0.05, *** p<0.01.

Table 5: Robustness Analysis: Study Period

	(1)	(2)	(3)	(4)	(5)
Panel A: Sales					
Treatment	0.043**	0.044**	0.045**	0.036*	0.037*
	(0.0214)	(0.0214)	(0.0217)	(0.0212)	(0.0215)
Observations	8905	8879	8846	8929	8849
Adjusted R^2	0.594	0.598	0.604	0.597	0.601
Panel B: Transactions					
Treatment	0.033*	0.036**	0.041**	0.027	0.024
	(0.0187)	(0.0179)	(0.0181)	(0.0198)	(0.0204)
Observations	8730	8692	8418	8764	8684
Adjusted R^2	0.718	0.726	0.658	0.718	0.717
Sample period	Preferred Model	÷ 4 Weeks	÷ 8 Weeks	+ 4 Weeks	+ 8 Weeks

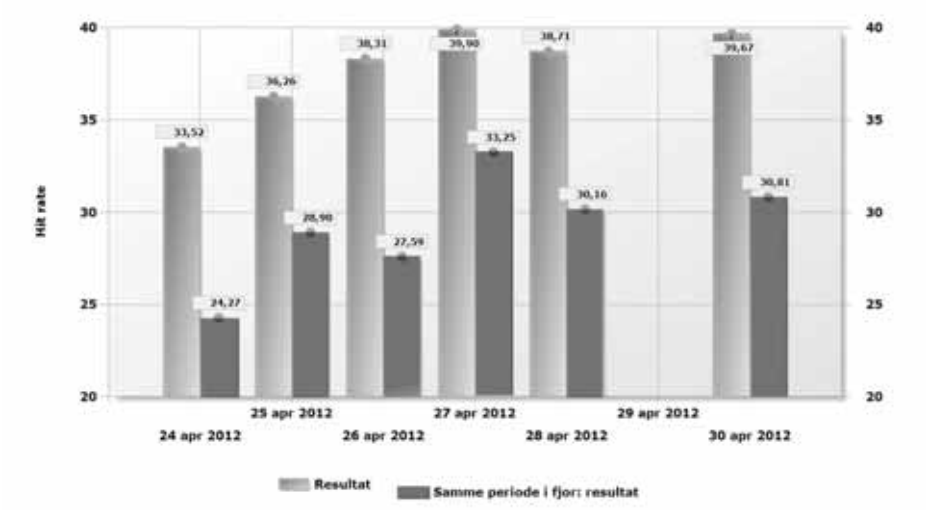
Notes: Dependent variable is log of weekly sales and log of number of sales transactions, in Panel A and B, respectively. OLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across stores. All regressions have the same specifications as our preferred model (Table 2, Column 4).
* p<0.10, ** p<0.05, *** p<0.01.

Table 6: Robustness Analysis: Exclusion Restrictions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Sales							
Treatment	0.043**	0.046**	0.042**	0.042*	0.028	0.054***	0.039**
	(0.0214)	(0.0214)	(0.0210)	(0.0216)	(0.0178)	(0.0200)	(0.0178)
Observations	8905	9114	8923	9329	9737	9841	11330
Adjusted R^2	0.594	0.593	0.594	0.572	0.606	0.599	0.588
Panel B: Transactions							
Treatment	0.033*	0.034*	0.032*	0.033*	0.025*	0.042**	0.030**
	(0.0187)	(0.0187)	(0.0183)	(0.0184)	(0.0144)	(0.0175)	(0.0136)
Observations	8730	8744	8747	8935	9557	9660	10723
Adjusted R^2	0.718	0.719	0.718	0.714	0.729	0.724	0.731
Restriction(s) dropped	None	1	2	3	4	5	All

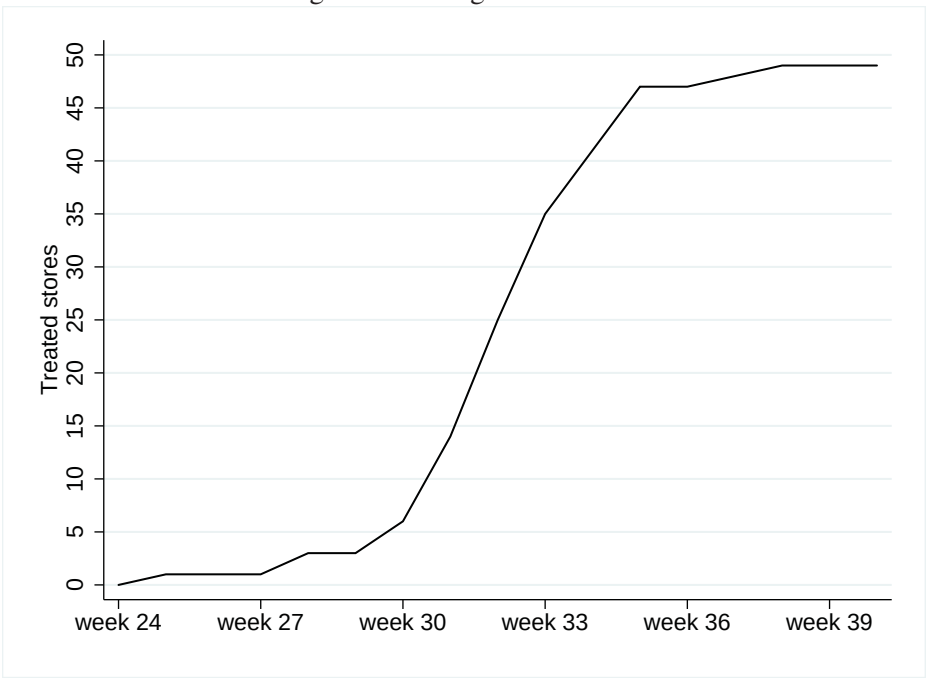
Notes: Dependent variable is log of weekly sales and log of number of sales transactions, in Panel A and B, respectively. OLS coefficients reported with robust standard errors in parentheses, corrected for clustering across stores. All regressions have the same fixed effects included as our preferred model (Table 2, Column 4). Exclusion restrictions referred to in table: 1) Exclude 3 non-treated stores missing all but a few observations; 2) Exclude the first 3 weeks of observations for stores that opened during Control or Treatment Year; 3) Exclude 9 stores that closed during Control or Treatment Year, or within 6 weeks after end of Treatment Year; 4) Exclude 8 self-owned stores that participated in piloting the treatment; 5) Exclude 9 stores that were not self-owned, but adopted parts of the treatment. * p<0.10, ** p<0.05, *** p<0.01.

Figure 1: Illustration of Graph in Stylized Sales Report



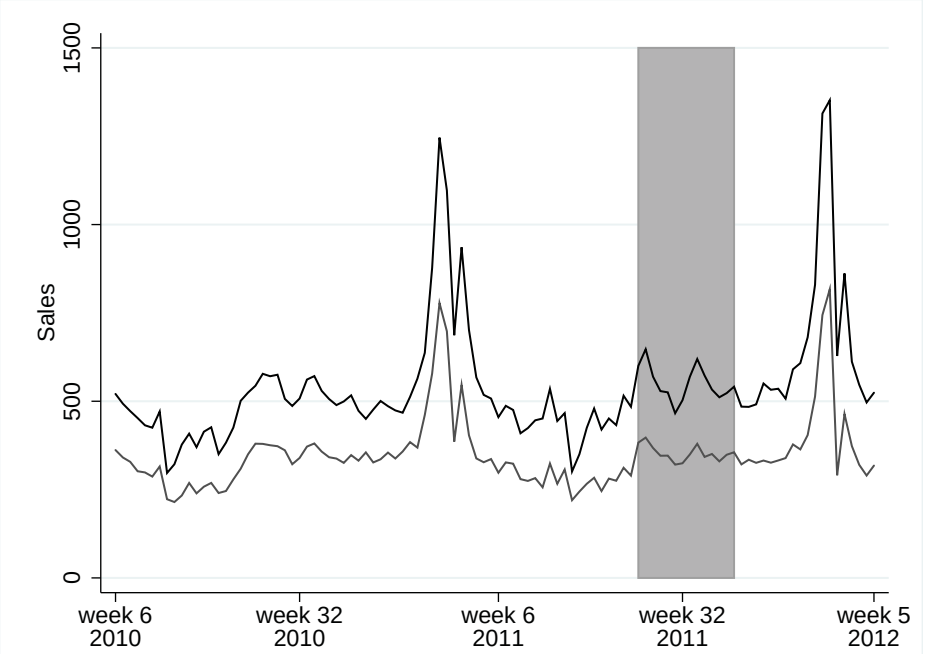
Notes: The graph illustrates daily hit rate in current year (light grey), in comparison with the same day in the previous year (dark grey).

Figure 2: Timing of Treatment



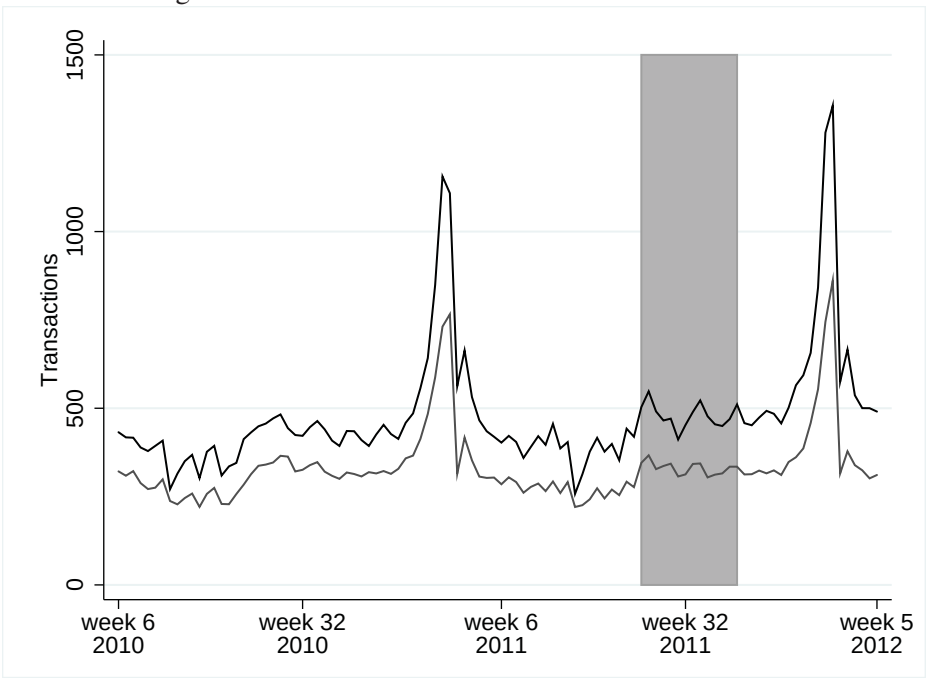
Notes: Number of treated stores by time. The graph illustrates how the treatment was introduced in different stores at different times between weeks 25 and 38 of 2011.

Figure 3: Sales in Control and Treatment Year



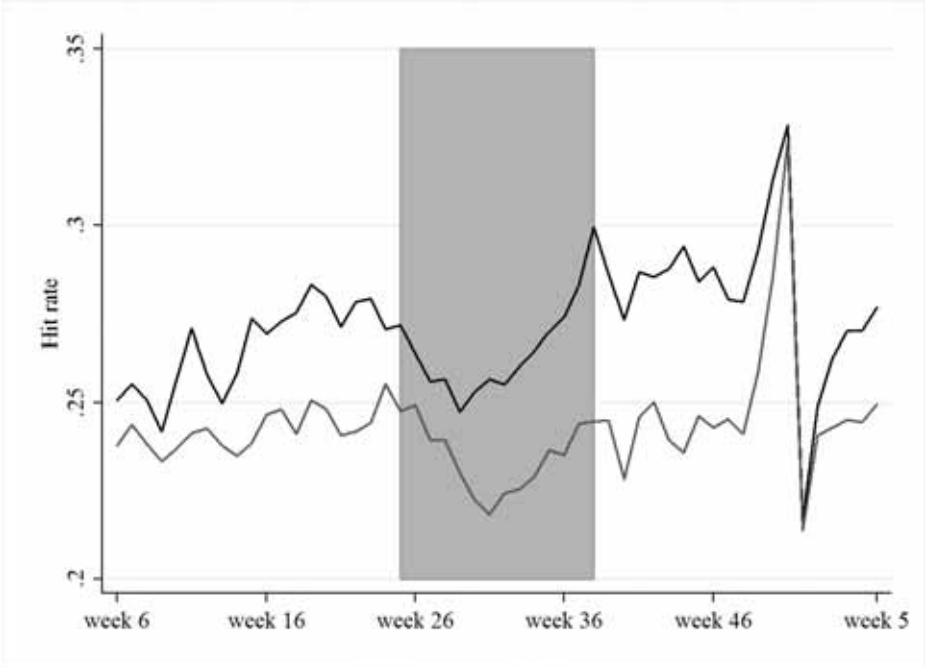
Notes: Sales in thousands by week for treated (black) and control (grey) stores. Grey area marks the period of treatment introduction (see Figure 2).

Figure 4: Transactions in Control and Treatment Year



Notes: Transactions by week for treated (black) and control (grey) stores. Grey area marks the period of treatment introduction (see Figure 2).

Figure 5: Hit Rate in Control and Treatment Year



Notes: Hit rate in treated stores by week for Treatment Year (black) and Control Year (grey). Grey area marks the period of treatment introduction in 2011 (see Figure 2).

Relative Performance Feedback: Effective or Dismaying?*

William Gilje Gjedrem

Abstract: In an experiment, I analyze whether the provision of relative performance feedback differently affects the performance of subjects when provided in various feedback environments. Subjects were ranked either relative to the performance of many subjects in the past or relative to three subjects working alongside themselves. Results indicate that the response from subjects in the former varies with how they perceived their own ability to solve the task. Those reporting low ability reduce their performance when provided with the feedback, whereas those reporting high ability improve. For subjects who were ranked relative to others working alongside themselves, no one respond negatively, but only those reporting high ability improve their performance. An important implication from this, especially for managers who design feedback policies in organizations, is that the way relative feedback is designed may lead to different behavioral reactions. In particular, the choice of benchmark used to relatively rank employees may result in responses that are not beneficial and lead to inefficient use of resources.

*I am grateful to Kristoffer W. Eriksen, Venke F. Haaland, Ola Kvaløy, David Laibson, Petra Nieken, Mari Rege, Marie C. Villeval, and seminar and conference participants for helpful comments and suggestions. Financial support from the Norwegian Research Council (227004) is gratefully acknowledged.

1. Introduction

Information technology has made it easier for firms to evaluate employee performance more precisely and to use these evaluations to rank employees in relation to each other. It might be tempting for firms to (uncritically) adopt these modern evaluation tools, believing that it will boost performance to new heights. Understanding the full extent of how relative performance feedback (RPF) affects employees is complex, as competing social mechanisms are likely to influence employees simultaneously. For example, while competition between employees may lead them to exert higher effort, it may also make them feel incompetent. A particular worry is that some mechanisms "crowd out" employees' intrinsic motivation to work (Deci, 1971; Frey and Oberholzer-Gee, 1997).

Two aspects of peoples' social concerns are likely to be important reasons why RPF affect motivation: they have competitive preferences and people care about whether they feel competent or not. The latter aspect is considered the core of intrinsic motivation (Deci and Ryan, 2000), and learning about the performance of others may adjust the perception of own competence. However, people may have competitive preferences too, which are strengthened with the introduction of relative performance feedback. These competitive preferences may arise from reasons such as joy of outperforming others or a desire for public recognition. In an effort to disentangle these social concerns, this paper presents an experiment that includes treatments designed to feature each concern separately, which should provide us with insight into how people respond differently to RPF in various environments.

In a lab experiment, two treatments are designed to feature each social concern separately. The first treatment, referred to as the CPF treatment, uses others' past performance as benchmark to rank the current subject's performance. Importantly, subjects in this treatment do not learn anything about the performance of any other subject in the same session. Thus, the environment is designed to reduce the competitiveness to a minimum, and it should rather provide a signal about the general competence level of others to solve the specific task. The second treatment, referred to as the TPF treatment, uses the performance of three others working alongside the subject as benchmark for ranking. This should raise the competitiveness to a higher level, as subjects directly compete against each

other for high ranks. In contrast to the former treatment, there are only noisy signals about the general competence level of others. These two treatments are compared to a baseline in which subjects only learn about their own absolute performance.[1] In addition to these three different feedback conditions, treatments are also varied across fixed and performance pay. Subjects work on a real effort task. Before being provided with any feedback, they are asked to self-assess their own ability to solve such tasks. The perception of own ability may prove important to future responses to RPF (Gibbons and McCoy, 1991; Abeler et al., 2011).

The overall results, using non-parametric tests, suggest no performance difference between the baseline and treatments under any pay-scheme. However, regression analysis is required to adequately control for subject's ability and to test for heterogeneous reactions. These analysis show that, when payment is fixed, the average performance of subjects is greater in both treatments compared to the baseline, but this is only significant in the CPF treatment. Large variations in performance exist, especially in the CPF treatment where those subjects with low self-assessed ability (SAA) reduce their performance substantially when RPF is provided. For the equivalent group of subjects in the TPF treatment, no such negative response has been identified. Moreover, those who report high ability perform better in both treatments. In the performance pay conditions of the experiment, no average treatment effects have been identified. However, differential analysis show that males and females respond differently depending on their reported ability.

This study has two main contributions. First, it highlights that behavioral reactions to RPF may differ depending on which social concerns the work environment features. This is important, as managers have numerous ways of designing feedback in their organization. In practice, RPF is likely to sometimes be based on instant performance data and other times on past performance data (e.g. last year/month/week), depending on the accessibility and type of performance data. Moreover, comparison of performance may often be based on a few selected employees, but other times it may be based on larger samples of employ-

[1] Absolute performance feedback is information about the output each individual produced during the previous working period.

ees (e.g. national/regional/branch/division).[2] Results from this experiment show that a benchmark that facilitates comparison of competence, may lead some to respond negatively. Second, it highlights that how one perceive own ability may play a key role in subsequent reactions to feedback about the performance of others. Whether you consider yourself competent or not at a particular task, may influence how susceptible you are to learning about the performance of others. Related studies have found evidence of higher performance when providing RPF, both from lab experiments (Hannan et al., 2008; Murthy and Schafer, 2011; Kuhnen and Tymula, 2012; Charness et al., 2014) and field data (Blanes i Vidal and Nossol, 2011; Bradler et al., 2016). However, there are still reasons to be concerned about deteriorating behavior, as some studies have not established a link between relative feedback and performance in certain contexts (Eriksson et al., 2009; Bellemare et al., 2010) or even identified negative effects (Barankay, 2012; Bandiera et al., 2013). This paper adds to this literature, and may suggest a reason for the divergence in the literature.

The remainder of this paper is organized as follows. Section 2 presents the experimental design. Section 3 provides an overview of the related literature and some hypotheses on the outcome of the experiment. Section 4 comprises the results of the experiment. Section 5 offers concluding remarks.

2. Experimental Design and Procedures

2.1 Task Description and Treatments

Subjects in the experiment are asked to solve a multiplication task, which is commonly used in related studies (see e.g., Kuhnen and Tymula, 2012; Hannan et al., 2012). Specifically, subjects are requested to find the product of a one-digit factor multiplied by a two-digit factor. They do this in five rounds, each round lasting 8 minutes.[3] After each round, they receive feedback on their performance.

The particular task was chosen for several reasons. It requires no prior knowledge other than basic math skills and it should be easy to understand.

[2] For example, a real-estate agency in Norway provide each employee with performance rank relative to all of the other employees in that agency.

[3] In related laboratory studies on RPF, the length of rounds have ranged from 1.5 to 5 minutes. I use slightly longer rounds to assure that the task is even more exhausting.

Moreover, it is important that performance depends on both ability and effort. This type of task induces heterogeneous ability levels, as math skills are expected to vary largely.[4] Therefore, some should feel competent performing the task, others not. Solving math questions is likely to be tiresome, especially when other activities are available in the lab. Specifically, subjects are allowed to engage in two alternatives: read newspapers[5] or surf the Internet using their mobile phone.[6] The combination of a simple and tiresome task itself and the alternative activities should induce disutility of effort. Finally, the task provides a stable and precise measure of performance.

Each session has the same sequence of multiplication tasks and all tasks are at about the same difficulty level, thus avoiding any dispersion of results due to variations in the task itself.[7] The screen displays how many minutes are left in each round. Subjects are not allowed to use any type of calculator or any other external remedies.[8] Subjects cannot continue to the next task until they have answered the current task correctly, to avoid strategic behavior of skipping tasks perceived as more difficult. If subjects answer incorrectly, they are told this and asked to try again.

[4] The subject pool includes students from various study programs that require no math skills to very high math skills.

[5] One fresh newspaper was available on each desk.

[6] A potential worry was that subjects would use their mobile phone to calculate the answers; however, the opportunity to cheat is limited in the lab and easy to detect. Subjects were informed that it was strictly prohibited to use any type of calculator and that they would receive no pay if detected. If someone used a mobile phone as a calculator, the subject would rapidly have to shift focus between the mobile phone and the computer screen, making it easy to detect. We balanced our attention to potential cheating with the concerns of remaining as neutral as possible to avoid any experimenter driven effects (such as subjects feeling monitored or pressured to work hard) (Zizzo, 2010).

[7] I would like to thank Camelia M. Kuhnen and Agnieszk Tymula for sharing the multiplication tasks they used in Kuhnen and Tymula (2012).

[8] Without these restrictions, the importance of ability level would be reduced.

Figure 1: Treatments in the Experiment

	Absolute performance feedback- (APF)	Competence performance feedback- (CPF)	Tournament performance feedback- (TPF)
Fixed pay	x	x	x
Performance pay	x	x	x

Figure 1 displays the treatments used in the experiment. Two dimensions are varied, the pay scheme and the type of feedback provided. Subjects are paid either a fixed amount or a piece rate (performance pay) for their participation.[9] The feedback dimension of the design varies across three different performance feedback conditions. In the baseline, subjects receive information on how many tasks they have solved correctly in the previous round. This is the same information as provided in the other two feedback conditions. The second feedback condition also provides feedback on their performance relative to the performances of participants in the past. The final feedback condition also provides feedback on their performance relative to three other subjects working alongside them on the same tasks. Feedback conditions are explained in detail below. With all these variations, there are six treatments. The overall experiment is a between-subjects design.

In an initial pre-round, all subjects have to work on the same real-effort task as in the main rounds. This round is used to measure (a proxy of) each subject's ability (an approach also used in e.g., Berger et al., 2013). In addition, subjects are also asked (on a scale from 1 to 4) to self-assess their ability to solve math tasks, a variable that to some extent reflects their prior self-esteem or a reference point in their perceived ability to solve such tasks.[10] This question is included, as such perceptions may prove important to how people respond to new information (Gibbons and McCoy, 1991; Koszegi, 2006; Abeler et al., 2011; Hannan et al., 2012). Whether subjects perceive this as an assessment of their ability in math relative to others or in absolute terms is unknown, but relative consideration is at least likely to influence it. Information about feedback is provided after the

[9] Subjects also earned NOK 100 in addition to the piece rate in the performance pay treatments.

[10] This self-assessment variable is recoded in the analysis (on a scale from 0 to 3, where 3 is highest assessment).

initial pre-round and this ability question.[11]

Subjects in the performance pay conditions know how much they earn per task solved, but they do not receive information about how much they have earned in each round until all five rounds are completed.[12] For all treatments, the sequence before the first round is as follows; first subjects have a trial period of one minute. Then they work for eight minutes on the math task in the pre-round, before they are asked to self-assess their own ability in solving math tasks.[13] After that, for the first time in the experiment, subjects are told about the performance feedback they will receive after each round and which type of feedback they will receive. Finally, they receive feedback based on the initial pre-round.

In the *absolute performance feedback (APF)* baseline, the performance feedback in each break consists of a text telling subjects that they solved *X* number of tasks in that round and a simple graph tracking their absolute performance across rounds. The use of a graphs in the provision of performance feedback is likely to be a close approximation of how feedback is presented to employees in firms, and some subjects may prefer graphical illustrations rather than just plain text. The information content in this treatment is also provided in the two treatments.

In the *competence performance feedback (CPF)* treatment, a rank of the subject's performance relative to many participants in the past is added to the feedback,[14] given as a rank from 1 to 4.[15] Subjects are informed that a rank of 1

[11] To ensure that subjects did not consider the pre-round as having a different purpose than the other rounds, the instructions said that there were 6 rounds in total (and not explicitly that the first round was an ability checker).

[12] This is to ensure that such information does not reinforce the strength of the feedback and to keep all conditions as similar as possible other than the deliberate variations. Subjects can still manually calculate their profit themselves.

[13] This question was not incentivized.

[14] I used real data from a session with 17 subjects who earned a fixed amount of 250 NOK, which I ran before the actual experiment was conducted. The session was identical to the baseline design. No significant differences (at the mean) were identified, neither in the pre-round or in any other round (both to fixed pay treatments and pooled data), for this session compared to the actual experiment. None of the participants in this session participated in the actual experiment, and they were not aware that their performance would be a benchmark for the experiment itself.

[15] Subjects in the CPF treatment did not know which conditions applied in the benchmark or any other information about it, except that it was based on a previous experiment. Although this might create some uncertainty about the benchmark for subjects in the CPF treatment, providing them with more details could easily open up for other questions or direct focus

means they are in the group with the 25% highest past performances and ranking of 4 implies that they are in the group with the 25% lowest past performances. As they are compared to many others, the feedback may invoke a strong signal about their ability relative to the general competence level. The environment is constructed such that there is no competition between the participants in the room; in fact, they are not told anything about any of the other participants. Hence, the competitiveness of this treatment is low. The graphical illustration also displays how many tasks the average of all people in the past solved in each of the respective rounds.

In the *tournament performance feedback (TPF)* treatment, subjects are also provided with feedback on their performance relative to three other randomly selected participants working alongside themselves on the same task. Subjects are given a rank from 1 to 4, and ranking 1 implies the highest performance in that round of the four participants, and so on.[16] The comparison group of four changes each round, which subjects are told explicitly. The competitiveness in this treatment should be high, as each subject's performance is ranked relative to three others sitting in the same room working on the same task. In contrast to the CPF treatment, the TPF treatment hardly reveals any information about the general competence level of the others, and if so, it is weak and noisy. The graphical illustration also displays the average of the four subjects' performance in that round.

2.2 Procedures

The experiment was conducted in the Business School at the University of Stavanger in Norway in early November 2013. It consisted of 12 sessions, two sessions for each treatment. Each session had up to 20 participants and lasted about one hour. Subjects were recruited from the whole student pool[17] at the

towards these details. Thus, knowing little may make it more plausible for subjects to assume that the benchmark had similar conditions. In any case, the information about the benchmark was the same for all subjects in the CPF treatment.

[16] In the regression analysis, this rank variable (for both the CPF and the TPF treatments) is re-coded, such that a higher numbered rank is a better performance.

[17] The student pool consists of a variety of students from faculty of Science and Technology, faculty of Social Sciences, and faculty of Arts and Education.

university using the recruitment program Expmotor.[18] Subjects were invited through their student email. In the experiment, subjects first received instructions about the work task and the pay scheme of the experiment. Instructions were given both in writing and read aloud, and subjects could ask questions before the experiment started.[19] The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). Payment was made in cash individually to each subject in a separate room by administrative staff after completion of the experiment.[20]

111 subjects participated in the fixed pay treatments earning 250 NOK (about $30) each; 35 subjects were in the APF baseline, 36 in the CPF treatment, and 40 in the TPF treatment. 110 subjects participated in the performance pay treatments, earning on average 240 NOK (about $29) each; 40 subjects were in the APF baseline, 30 in the CPF treatment, and 40 in the TPF treatment.

3. Related Literature and Hypotheses

3.1 Related Literature

Standard economic theory assumes that rational individuals seek to maximize their utility by obtaining the highest payoff at the least cost. It predicts that subjects choose not to work when wages are fixed, especially in a laboratory experiment where there are no reprimands if shirking. When paid a piece rate, subjects are expected to adapt according to their profit maximizing function, weighing utility of income against disutility of effort.

Despite these theoretical predictions, recent studies from personnel economics have shown that people exert costly effort even when wages are fixed (e.g., Kuhnen and Tymula, 2012; Charness et al., 2014). Other areas of research in economics, such as the gift-exchange literature starting with Fehr et al. (1993), have consistently found similar results (see e.g., Gneezy and List, 2006; Gächter and Thöni, 2010; Kube et al., 2012). Intrinsic motivation theory, first developed

[18] Developed by Choice Lab researcher Erik Sørensen and Trond Halvorsen at the Norwegian School of Economics (NHH).

[19] A translated copy is available in the appendix.

[20] A few sessions were paid electronically due to limitations in administrative capacity, and this was made public after completion of the session.

in the psychology literature by Deci (1971) and Deci and Ryan (1985), accounts for such other behavioral responses that go beyond what standard economic theory predicts. It argues that people also get utility (inherent satisfaction) for reasons other than what they get from external motivation, and that people often have an intrinsic drive to perform the task.

Assuming that people are motivated to work, this motivation may be crowded out or crowded in (reinforced) when firms carry through interventions (Frey and Oberholzer-Gee, 1997; Frey and Jegen, 2001; Falk and Kosfeld, 2006). For example, experiments have shown that if subjects are not paid enough they will produce less (Gneezy and Rustichini, 2000) and that symbolic awards can be used to increase productivity (Kosfeld and Neckermann, 2011; Ashraf et al., 2014; Bradler et al., 2016). A meta-analysis by Kluger and DeNisi (1996) shows that studies on feedback interventions have reported mixed effects on performance. One type of such feedback intervention is RPF. Intuitively, though, how an employee perceives RPF may vary largely, depending on factors such as its content and how it is presented. This suggests that there are competing mechanisms affecting employees when such an intervention is carried out, and that the net effect on performance is a mixture of these mechanisms.

The majority of recent experimental studies have shown positive effects of RPF (e.g., Murthy and Schafer, 2011; Blanes i Vidal and Nossol, 2011; Hannan et al., 2012; Bradler et al., 2016), suggesting that it reinforces motivation. For example, Blanes i Vidal and Nossol (2011) found a long-lasting increase in productivity of almost 7% after starting to inform employees about their relative performance. Indeed, some of these studies have even shown that the majority of subjects improve when RPF is provided (Azmat and Iriberri, 2010; Kuhnen and Tymula, 2012; Charness et al., 2014). Other studies have only found context specific effects or no link between RPF and performance (Hannan et al., 2008; Eriksson et al., 2009; Azmat and Iriberri, 2016). Barankay (2012), conducting a field experiment in a furniture retailer, found that removing RPF increased subsequent employee performance, even though pay was not linked to relative performance. Bandiera et al. (2013) also showed that ranking incentives reduced the performance of the lowest ranked teams. These mixed findings suggest that we have yet to disentangle different competing mechanisms in the feedback evaluation process. These studies have not looked at variations in the benchmark

used to rank people or other sorts of conditions the environment features. As individuals have social concerns (Festinger, 1954), their reaction likely depends on how they perceive the feedback, which again is likely to depend on factors such as the competitiveness and their perception of competence.

A range of reasons have been put forward to explain why employees may respond to relative performance feedback. Concerns with social comparison may induce high performers to do even better if their work effort is (publicly) recognized (e.g., Moldovanu et al., 2007; Kosfeld and Neckermann, 2011; Ashraf et al., 2014; Kvaløy et al., 2015) or spark the performance of those who lag behind (Charness et al., 2014). Moreover, employees' competitiveness may rise as some enjoy outperforming others (Dohmen and Falk, 2011) or feeling dominant (Abbink and Sadrieh, 2009). Concerning self-esteem and feelings of competence, the content of relative performance feedback may prove or disprove prior beliefs about one's status or simply bring undesirable attention to it. Some might also be concerned about being monitored and thus might feel more obliged to work harder; others may simply feel that RPF is a rigorous way of controlling employees.

3.2 Hypotheses

Economic theory offers very few predictions on the outcome of the experiment. The predictions from standard neoclassic theory would be that there should be no difference between subjects receiving RPF or APF, and that subjects work harder with a piece-rate compared to a fixed payment. However, from the related literature outlined above and knowledge on that people's utility function very often seem to include behavioral aspects that cannot be explained by standard theory, subjects may respond to feedback and they may work even though payment is fixed. Thus, I do not have any formal predictions or hypotheses related to the outcome of the experiment.

Beside the overall treatment effects, I want to study how people respond differently to relative performance feedback depending on where they are positioned in the ability distribution. To analyze this I use two proxies for ability. The first proxy is the self-assessed ability that subjects report prior to any knowledge about the treatment. The second is their performance in the pre-round, also

reported before any knowledge about the treatment. The initial idea when this project started was that people may respond differently to feedback depending on whether they perceieve own ability as low or high, and especially if the relative feedback allows them to compare their competence relative to others.

According to Deci and Ryan (2000), the key to enhancing intrinsic motivation is for employees to possess a feeling of being competent. In turn, however, feeling incompetent may rather undermine intrinsic motivation. The majority of related studies outlined above suggest a positive effect of RPF, and therefore I generally expect similar results. The CPF treatment uses a benchmark designed so that feedback may be perceived as information about the general level of competence, and hence make those ranking low (high) feel incompetent (competent). Moreover, this feedback may enable low performing subjects to see that they are positioned as a negative deviation from what might be considered as the social performance norm (Bernheim, 1994). Thus, there might be a positive relationship between higher SAA and receptiveness of the information revealed in this treatment (Burks et al., 2013). People may respond differently to feedback depending on how they perceive their own ability. In particular, subjects who feel competent may be positively affected by the performance feedback, as they may consider relative feedback as favorable information. On the other hand, those who do not feel competent might react negatively to it, as they prefer not to have this information at all.

When the environment features strong competition as in the TPF treatment, where subjects learn about the performance of three others working alongside themselves, the feedback may be perceived differently. As people are likely to have competitive preferences, providing such feedback may strengthen these preferences. Contrary to the CPF treatment, subjects are not likely to feel that the feedback represents a social performance norm or a measure of the general level of competence. This treatment is most similar to the existing lab evidence on RPF, and as they have often shown, most subjects tend to improve their performance (e.g., Charness et al., 2014). For high ranking subjects it might be motivating to maintain this winning position, whereas a loss may motivate to improve performance to win the next competition.

Other lab experiments have studied differential effects of RPF based on their rank. Some studies have found that initial worst performers often improve more

than initial best performers (Kuhnen and Tymula, 2012; Charness et al., 2014), whereas others have not established such pattern (Azmat and Iriberri, 2016). Just a few field studies have particularly looked at this issue. For example, Blanes i Vidal and Nossol (2011) found very similar effects of RPF for different types of preexisting levels of performance, and all overwhelmingly positive. Studies on RPF in school have shown that providing relative grade feedback did not negatively affect the subsequent grades of students (Azmat and Iriberri, 2010; Bandiera et al., 2015); however, grades are censored at both ends of the grading scale, making analysis of improvements restricted. In this experiment I can use ranks from the pre-round to study how subjects differently respond, depending on this rank. This approach is likely to provide similar results as studying different reactions depending on reported SAA, as both of these are proxies for true ability.

4. Experimental Results

4.1 Descriptive Analysis

Descriptive statistics from the experiment are presented in Table 1. The average number of tasks solved during a round is about 26, and it takes about 19 seconds on average to complete one task. The number of tasks solved in the pre-round is lower than in subsequent rounds, an important reason for this is likely to be learning effects from working on the task. There are some differences between baseline and treatments when it comes to demographic characteristics. For example, there are more males than females in the experiment and there are some gender differences across treatments. Moreover, it also seems to be some differences in the proxy variables for ability across treatments. To better account for this, running regression analysis that include control variables is important.

The main outcome variable in the analysis is the number of tasks subjects solve during a round. I also make use of two other alternative outcome variables. The first is the number of times a subject submit an answer, i.e. the sum of tasks solved and incorrect attempts. This variable may be considered a measure of subject's total effort, without the concern of quality (denoted *Total effort* in the tables). The second variable is the ratio between tasks solved and total effort, which may be a more complete measure of quality (denoted *Success rate* in the tables).

Non-parametric Mann-Whitney U-tests comparing the baseline to the treatments (for all three outcome variables) are reported in Table 1. All tests, independently of pay-scheme, show no difference in outcome between the baseline and any treatment.[21] However, as noted above, the pre-round variables and the reported SAA's suggest an imbalance in ability of subjects across treatments. In particular, means for both the pre-round variable and SAA variable are higher in the baseline compared to both treatments, under both fixed pay and performance pay.

[21] Note that these insignificant results should be interpreted cautiously, as any significant treatment difference would require a fairly large effect size. For example, under fixed-pay, given the sample size of 35 in the baseline and 36 in the CPF treatment, the minimum difference required to detect a significant difference at the 5% level with 80% probability is d=0.69. This implies that a difference would require almost 12 tasks more solved for a treatment group to be significantly different from the baseline.

Figure 2 display the average number of tasks that subjects solved during all five rounds (pre-round excluded) across treatments in the fixed pay conditions, depending on their reported self-assessment of ability. As expected, subjects with higher SAA also perform better. If we only look at the averages of subjects across treatments in the first working round in Figure 3, the pattern is similar. One particular thing to notice is that subjects with the lowest SAA in the CPF treatment perform worse than those in the comparable APF baseline. As self-assessment of ability increases, so does the performance of those in the CPF treatment and more than it does for those in the APF baseline. In the performance pay conditions of the experiment, displayed in Figure 4, there are no clear patterns. One particular weakness to this graph is that only two subjects in the APF baseline self-assessed their ability to the lowest level, and only a few subjects self-assessed their ability to the highest level (in all treatments).

The descriptive analysis and figures provide an overview of the experimental outcome. However, structuring the data as a panel, using multiple observations per subject, and adding controls to correct for some of the imbalances, allow for a more comprehensive use of the available data and for heterogeneity analysis. I will now present the results of the regression analysis, grouped into two subsections. The first part will be on conditions with fixed pay, and the second part on conditions with performance pay.

4.2 Treatment Effects - Fixed pay

4.2.1 Main Results: Treatment Effects

Results that follow are generally based on Random Effects generalized least squares (GLS) with standard errors clustered on session, using the following regression model:

$$\begin{aligned} Tasks_{i,t} = \alpha + \beta CPFtreatment_i + \delta TPFtreatment_i \\ + Round_t + SAA_i + PreRound_i + Z_i + \varepsilon_{i,t} \end{aligned} \quad (1)$$

where $Tasks_{i,t}$ is the number of tasks solved by subject i in period t, $CPFtreatment_i$ and $TPFtreatment_i$ are indicators of which treatment subject i is in and $Round_t$

is a linear trend capturing learning effects. Moreover, self-assessed ability (SAA_i) and pre-round performance ($PreRound_i$) of subjects are included as a proxy for their true ability to solve math tasks. Finally, Z_i is a set of individual predetermined characteristics and $\varepsilon_{i,t}$ is the idiosyncratic error.

The main results are presented in Table 2. Column 1 shows the overall treatment effects on the main dependent variable. Both the CPF treatment and the TPF treatment have a positive coefficient, but only the CPF treatment is weakly significant at the 10% level. Hence, adding controls provide a slightly different picture compared to the simple comparison of means. It suggests an overall weakly positive relationship between relative feedback and performance. The size of the effect does not seem to be different for the two treatments.

However, people may respond differently to relative feedback, depending on their prior belief about own ability. Therefore, in column 2, the treatment variables are interacted with the SAA variable. This should reveal whether there are some differences in response to the feedback, depending on this belief. The results reveal substantial heterogeneous effects. First, those with lowest reported SAA in the CPF treatment do significantly worse after being ranked relative to others, they solve about 5 tasks less ($p<0.01$).[22] This is relative to subjects reporting the lowest SAA in the baseline. Second, when the reported SAA linearly increases, the performance of subjects in the CPF treatment is higher than in the baseline. The interaction between the treatment and the SAA is highly significant ($p<0.01$).[23]

Results are different for subjects in the TPF treatment. No one seems to be negatively affected by the relative feedback. For subjects with reported SAA above the lowest level, the linear relationship is weakly positive ($p=0.071$), suggesting that the feedback improves the performance of these subjects. Comparing the two treatments directly, subjects with lowest reported SAA in the CPF treatment perform significantly less than subjects with the lowest reported SAA in the TPF treatment ($p=0.025$). Moreover, the interaction between the linear SAA and CPF treatment is significantly greater than the interaction between the linear

[22] In the table, this is reflected in the CPF treatment variable, as the other levels of reported SAA are included in the interaction between this and the treatment variable.

[23] Alternatively one could interact each level of reported SAA with each treatment, however this approach suffers from lack of power given the relatively small sample in this experiment. The only significant interaction is between the highest level of reported SSA and the CPF treatment .

SAA and the TPF treatment ($p<0.01$).

The pattern is similar for the alternative dependent variables, as can be seen in columns 3 to 6. For subjects in the CPF treatment, total effort does not seem to have increased overall. However, when studying different levels of reported SAA, there is a clear negative effect on effort for those who report the lowest SAA, whereas higher reported levels of SAA seems to improve total effort as well. The same applies to the success rate. For subjects in the TPF treatment, total effort seems to have increased slightly, and this seems to be particularly strong for those who reported the lowest level of SAA. This might suggest that the feedback itself motivated these subjects, but that their low ability constrained them from improving their performance in terms of tasks solved. The success rate seems unaffected by the TPF treatment.

For those who reported the lowest level of SAA in the CPF treatment, the relative feedback may seem to have crowded out their intrinsic motivation to work, and this is reflected in both how many tasks they solve and how many attempts they make to answer the tasks. This is consistent with the argument that these subjects may explicitly learn about their negative deviation from a social norm and/or that they feel less competent as a result of the relative feedback, thereby reducing their performance. In contrast, when the reported SAA is higher, the relative feedback seems to enhance their intrinsic motivation. This is very much in line with the behavioral reflections in section 3, in that only subjects who perceive own ability as high will increase performance. Moreover, when the feedback is competitive and subjects are evaluated relative to only a few others, no one seems to be negatively affected, and if anything performance and effort seem to increase. More generally these results suggest that it matters to people whether the feedback is presented in a competitive manner or more as a basis for evaluating their competence. There may be several reasons for the overall weak treatment effects, other than the observed differential treatment effects. For example, subjects may already perform close to their maximum potential (an upper ceiling) independent of treatment, or that the additional feedback information is simply too weak to induce a stronger response.

As a comparison and robustness check, a Random Effects Tobit model is also included in Table A1-1. These analysis provide results that are very similar to the findings presented above, but standard errors have increased somewhat,

making some of the coefficients less significant.[24]

4.2.2 Differential Effects of Rank

Another approach to analyze treatment effects, is to study how subjects respond to feedback information about their rank based on the pre-round. This rank is based on their performance prior to any knowledge about the treatment, as was the case for the SAA variable. Hence, this may also be viewed as an analysis of differential treatment effects across levels of ability, i.e. whether high skilled subjects respond differently than low skilled subjects. Thus, results are expected to be similar as in the previous table. Outputs can be seen in Table 3. Ranks from the pre-round are interacted with the treatment variables, hence the way to interpret coefficients are similar as in Table 2. Importantly though, subjects in the baseline do not actually learn about their true rank.[25] Results show that lowest ranked subjects perform significantly less when they learn what others have done in the past, and their success rate drops. Moreover, when rank is higher, subjects perform better and the success rate is higher. This is very consistent with the previous table. As in the previous table, even though lowest ranked subjects in the TPF treatment seem weakly negatively affected by their low rank, total effort have significantly increased. This indicates that subjects may want to increase effort, but do not have the ability to do so. For the remaining subjects who ranked higher in the TPF treatment, there are weakly positive association between rank and general performance.

4.2.3 Gender Analysis

Recent research has shown that genders tend to respond differently to competition and information (e.g., Gneezy et al., 2003, 2009; Marianne, 2011), and this also seems to apply in environments providing RPF Azmat and Iriberri (2016). Table 4 builds on the same type of analysis as in Table 2, only using sub-samples of gender. To ensure sufficient number of observations in each level of reported SAA, I merge the two lowest and the two highest levels of SAA, and use a binary variable to separate them. Males performance are shown in columns 1-3.

[24] The same applies for regressions with robust standard errors without clustering on sessions.
[25] I use the rank they would have received had they been in the CPF treatment.

The performance of those with low reported SAA is slightly below comparable males in the baseline. However, males who report high SAA perform about 6 tasks more (-3.086+8.837, p<0.01) in the CPF treatment than in the baseline. The pattern is similar for the alternative dependent variables, although mostly insignificant. There seems to be no effect on the performance of males in the TPF treatment, although the sign of the coefficient goes in the opposite direction than for males in the CPF treatment. Columns 4-6 are the comparable results for females. The pattern is similar, females with low reported SAA in the CPF treatment perform worse, especially when it comes to effort. Moreover, females seem to perform better when their reported SAA is higher, however these results are insignificant.

These results are in line with the treatment effects we observed in Table 2, however it suggest that differences are primarily driven by variations in the performance of males, as they seem to have stronger negative or positive response. Males appear to be particularly affected by feedback in the competitive environment. Some males with low SAA may feel dismayed by the feedback, whereas those with high SAA might feel they have to prove themselves competent by increasing their effort.

4.2.4 Responses on Stress, Feeling, and Motivation

After the experiment, subjects answered a questionnaire that included the following three questions. The first question asked whether they felt stressed when working on the task (from not stressed to very stressed). In column 1 of Table A1-2, subjects in the CPF treatment report lower levels of stress relative to subjects in the TPF treatment (p-value=0.065), suggesting that less competitive feedback is perceived less stressful. In column 2 we see that those who reported low SAA in the CPF treatment also tend to be less stressed, suggesting they are less compelled to work and perhaps become discouraged after receiving feedback information. For those who report higher SAA, there is a positive relationship between treatments and levels of stress. The second question asked how subjects felt working on the task (from very boring to very fun). Overall, subjects in the TPF treatment seem to enjoy the task more than the others, compared to both the baseline (p-value=0.023) and the CPF treatment (p-value=0.026). Subjects

with high SAA also report enjoying the task more than those with low SAA, independent of feedback condition. The final question asked whether subjects felt motivated or discouraged by the feedback during the breaks. In column 5, subjects in the CPF treatment significantly ($p<0.05$) report being less encouraged by the feedback. In column 6, this applies particularly to subjects with low SAA, who report significant discouragement from the feedback in the CPF treatment relative to these subjects in the baseline. Subjects with high SAA report higher motivation. For the TPF treatment, there is a weak positive relationship between treatment and reported motivation levels. This shows that subjects explicitly report disliking feedback when they consider their SAA low in an environment that features comparison of competence. These results, although possibly endogenous, illustrates that subjects indeed perceived feedback differently across the treatments.

4.3 Treatment Effects - Performance Pay

The following analysis are analogous to the empirical analysis in subsection 4.2. Results in Table 5 suggests no overall treatment effects or any differential effects on performance or effort for subjects in the performance pay conditions. The only evidence of any treatment effect is when success rate is used as the dependent variable. Then there are suggestive evidence of higher success rate in both treatments, primarily driven by those who reported lowest SAA. Overall, the lack of treatment effects suggest that under performance pay, RPF does not seem to matter in the context of this experiment. As Bellemare et al. (2010) similarly argued, a reason for the lack of effect when payment is conditioned on performance, is that the higher effort level associated with pay is likely to reduce the response to the social comparison process, or that the feedback is simply more conspicuous when payment is fixed. Put differently, it may be that when subjects are paid for higher performance, focus on earnings dominates feedback information. Another explanation would be that all subjects perform their maximum when performance is incentivised, and that they simply have no additional effort to offer. However, the performance in these treatments is at the same level as in the fixed pay treatments. Previous related studies have often shown a positive treatment effect of RPF under performance pay, but results

are mixed. For example, Eriksson et al. (2009) also found no effects of relative feedback under performance pay. The Random Effects Tobit regressions in appendix Table A1-3 show similar results.

Consistent with the previous paragraph, there are no differential treatment effects depending on the specific rank that subjects received after the pre-round, see Table 6. There are however some interesting gender differences, see Table 7. Males in the CPF treatment who reported low SAA perform about 6 tasks more (p-value<0.01), and males who reported high SAA perform about 5 tasks less (6.140-11.056, p-value<0.01), relative to comparable males in the baseline. Although much less significant, the same applies to males in the TPF treatment. For females in the CPF treatment, the treatment effect goes in the opposite direction, and cancel out the overall effect. Here, females who reported low SSA perform about 6 tasks less after having been provided with RPF (p-value<0.01), whereas females who reported high SSA improve performance by about 5 tasks (-6.149+10.658, p-value=0.081). Females in the TPF treatment seem unaffected by the treatment. The overall gender differences are striking, as they clearly seem to react very differently to the feedback information. However, one must interpret the results with some cautiousness given the few observations in the experiment.

Finally, look at subjects' responses from the questionnaire at the end of the experiment. For the question about stress, subjects in the TPF treatment report feeling significantly more stressed than subjects in the baseline, but this does not apply for subjects in the CPF treatment (column 1, Table A1-4). Differential analysis in column 2 tells us that particularly subjects in the CPF treatment with low SAA report not being stressed and subjects with high SAA feel more stressed. In terms of whether subjects enjoyed the task or not (column 3), subjects with high SAA report higher level of satisfaction in both treatments. Asking subjects whether the feedback motivated them or not (column 5), subjects in the CPF treatment report being less motivated, and subjects in the TPF treatment report no change in motivation level. Column 6 shows that particularly subjects with low SAA in the CPF treatment report being less motivated by the feedback.

5. Concluding Remarks

In the workplace, employees are often provided with some sort of relative performance feedback. However, there are reasons to believe that this might not always improve performance, as such information may be perceived differently from one person to another. Some are likely to put effort into improving their relative position for different reasons, whereas others may experience a drop in motivation after having seen the feedback. If so, research is scant on what causes different reactions, and to which extent the design of the feedback matters to how people respond. This experiment does not answer all of these questions, but investigates whether people do in fact respond differently to feedback, depending on who they are compared relative to and on the reported beliefs about own ability to solve the task. In particular, one environment features feedback designed to be competitive, the other environment features feedback designed to show the general competence of others. Any evidence of different reactions to the feedback would indicate that there may be different social concerns that work behind the scenes, which should be of interest for those designing feedback interventions in organizations and more generally to understand what motivates people.

This paper provides evidence that the provision of RPF may lead to a disparity in the performance of employees. In particular, when feedback is presented in an environment that features comparison of competence, those who perceive their ability to be low reduce their performance, suggesting that they are less susceptible of receiving such information. Their motivation may drop as a result of learning that their performance is below the social norm and/or that they feel less competent. On the other hand, when the environment features competition, and feedback is based on the relative comparison of just a few others, they seem less sensible to such information. Then no one seem to reduce their performance, and the overall performance seems to have increased.

There are several implications for those who design feedback interventions in organizations. In support of existing literature, it suggests that the average performance rise when RPF is provided, although the effects are modest in this experiment. Second, the results propose that managers have to think carefully about which type of employees they present feedback for and what type of benchmark they should use to rank them. Certain work environments are known to have

highly competitive workers who believe in their own skills. They might enjoy information about what others have done, and almost instant relative feedback may serve to increase productivity. On the other hand, if the work environment is characterized by workers who perhaps need more stable conditions and feedback that serve to boost their confidence, this paper suggests that an environment that facilitates comparison of competence perhaps is a bad idea. If anything, it should be feedback comparing only a few workers' current performances, making it more of a local competition. An alternative approach, which might avoid workers feeling less competent after seeing the performance feedback, is to form teams and then rank teams relative to each other instead. This way there is still a competitive element, without the individual focus that might be demotivating for certain workers. Whether or not this is an efficient way to design feedback interventions remains an empirical question for now.

References

Abbink, K. and Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, 105(3):306–308.

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, 101(2):470–492.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94:435 – 452.

Azmat, G. and Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy*, 25(1):77–110.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? evidence from a natural experiment on the effect of individual feedback on performance. *Labour Economics*, 34:13–25.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. Technical report, working paper, Wharton School, University of Pennsylvania.

Bellemare, C., Lepage, P., and Shearer, B. (2010). Peer pressure, incentives, and gender: An experimental analysis of motivation in the workplace. *Labour Economics*, 17(1):276–283.

Berger, J., Harbring, C., and Sliwka, D. (2013). Performance Appraisals and the Impact of Forced Distribution - An Experimental Investigation. *Management Science*, 59(1):54–68.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016). Employee recognition and performance: A field experiment. *Management Science*.

Burks, S. V., Carpenter, J. P., Goette, L., and Rustichini, A. (2013). Overconfidence and social signalling. *The Review of Economic Studies*, 80(3):949–983.

Charness, G., Masclet, D., and Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, 60(1):38–55.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1):105–115.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L. and Ryan, R. M. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review*, 101(2):556–590.

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679 – 688. European Association of Labour Economists 20th annual conference University of Amsterdam, Amsterdam, The Netherlands. 20 September 2008.

Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *The American Economic Review*, 96(5):1611–1630.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2):117–140.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5):589–611.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding- out. *The American Economic Review*, 87(4):746–755.

Gächter, S. and Thöni, C. (2010). Social comparison and performance: Experimental evidence on the fair wage-effort hypothesis. *Journal of Economic Behavior & Organization*, 76(3):531–543.

Gibbons, F. X. and McCoy, S. B. (1991). Self-esteem, similarity, and reactions to active versus passive downward comparison. *Journal of Personality and Social Psychology*, 60(3):414–424.

Gneezy, U., Leonard, K. L., and List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5):1637–1664.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Hannan, R. L., Krishnan, R., and Newman, A. H. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review*, 83(4):893–913.

Hannan, R. L., McPhee, G. P., Newman, A. H., and Tafkov, I. D. (2012). The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review*, 88(2):553–575.

Kluger, A. N. and DeNisi, A. (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–.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Koszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kube, S., Marechal, M. A., and Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *The American Economic Review*, 102(4):1644–1662.

Kuhnen, C. M. and Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science*, 58(1):94–113.

Kvaløy, O., Nieken, P., and Schöttner, A. (2015). Hidden benefits of reward: A field experiment on motivation and monetary incentives. *European Economic Review*, 76:188–199.

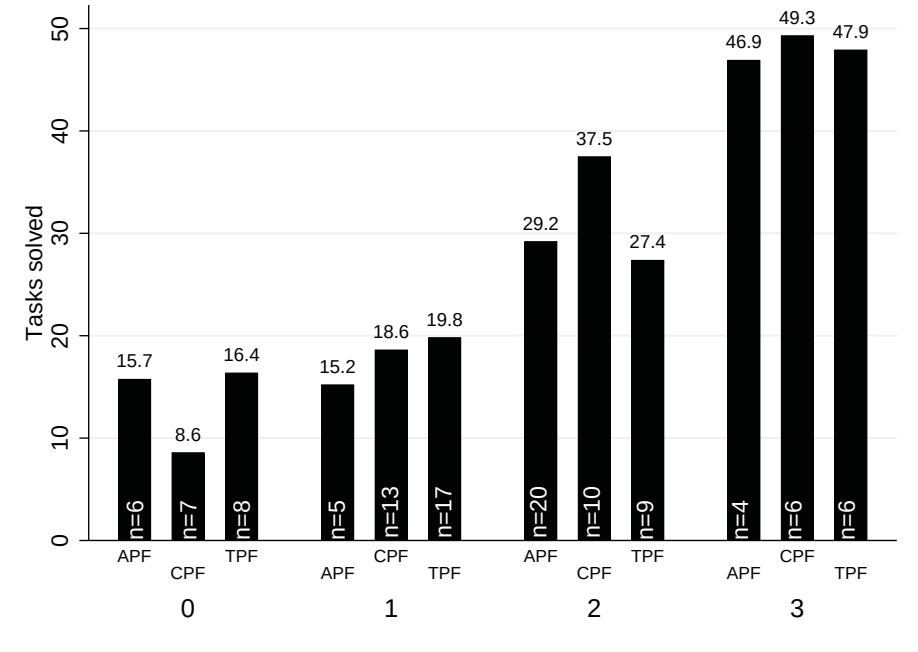
Marianne, B. (2011). Chapter 17 - new perspectives on gender. volume 4, Part B of *Handbook of Labor Economics*, pages 1543 – 1590. Elsevier.

Moldovanu, B., Sela, A., and Shi, X. (2007). Contests for status. *Journal of Political Economy*, 115(2):338–363.

Murthy, U. S. and Schafer, B. A. (2011). The effects of relative performance information and framed information systems feedback on performance in a production task. *Journal of Information Systems*, 25(1):159–184.

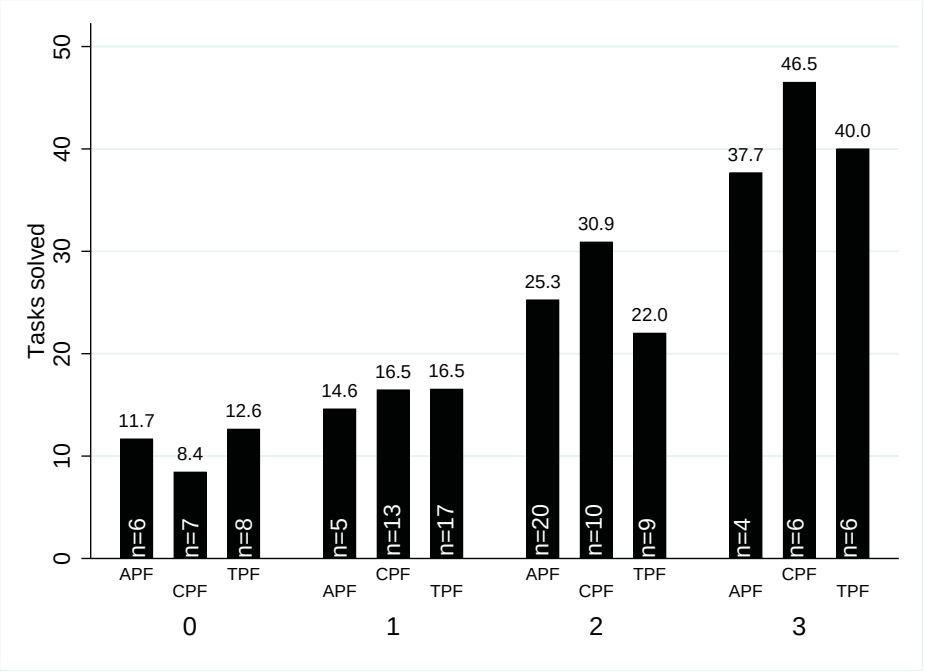
Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.

Figure 2: Tasks Solved over All Rounds (Fixed Pay)



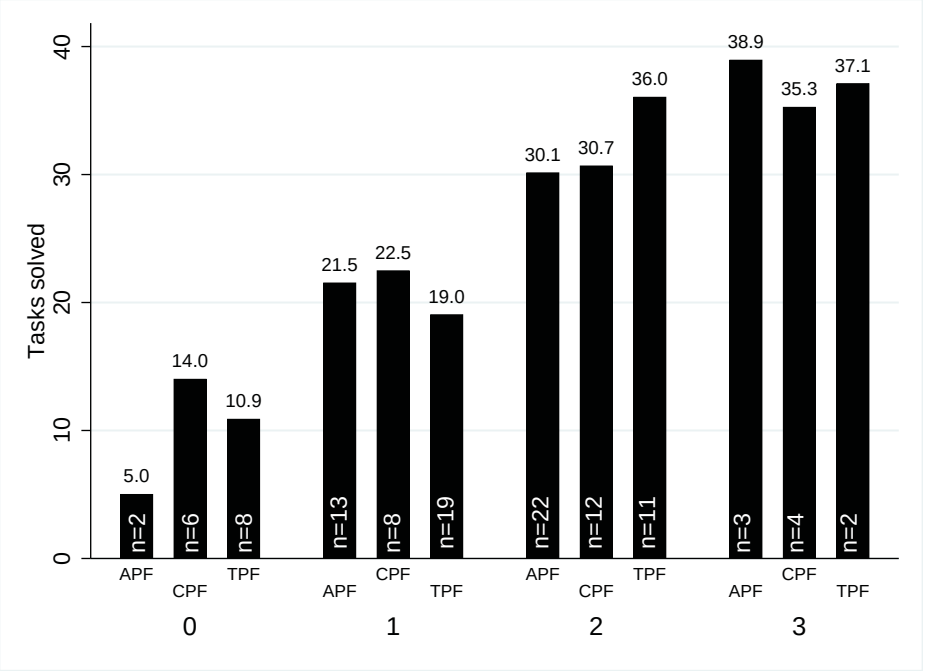
Notes: This histogram shows the average number of tasks solved over all rounds. Subjects are categorized by treatment and divided into different levels of reported SAA, which is coded from 0 (lowest) to 3 (highest).

Figure 3: Tasks Solved in 1st Round (Fixed Pay)



Notes: This histogram is equivalent to Figure 2, with the only difference that it displays average number of tasks solved in the 1st round only.

Figure 4: Tasks Solved over All Rounds (Performance Pay)



Notes: This histogram shows the average number of tasks solved over all rounds. Subjects are categorized by treatment and divided into different levels of reported SAA, which is coded from 0 (lowest) to 3 (highest).

Table 1: Descriptive Statistics

	(1) APF baseline	(2) CPF treatment	(3) TPF treatment	(1) vs. (2)	(1) vs. (3)
Panel A: Fixed pay					
Average tasks solved	27.00	27.01	25.03	0.97	0.72
	(15.54)	(18.31)	(14.56)		
Average total effort	34.88	33.00	33.11	0.67	0.92
	(17.19)	(18.41)	(13.78)		
Average success rate	0.738	0.731	0.721	0.295	0.924
	(0.134)	(0.218)	(0.182)		
Tasks solved in pre-round	18.43	16.94	15.35	0.61	0.24
	(12.30)	(12.17)	(10.20)	[2.90]	[2.60]
Total effort in pre-round	26.54	23.28	25.03	0.24	0.59
	(12.03)	(10.91)	(11.97)	[2.72]	[2.78]
Success rate in pre-round	0.648	0.649	0.600	0.983	0.344
	(0.197)	(0.272)	(0.236)	[0.056]	[0.051]
SAA	1.63	1.42	1.33	0.35	0.17
	(0.91)	(1.00)	(0.97)	[0.23]	[0.22]
Age	23.17	24.72	23.40	0.23	0.81
	(5.23)	(5.64)	(2.62)	[1.29]	[0.94]
Years of education	2.57	2.64	3.05	0.82	0.18
	(1.46)	(1.05)	(1.60)	[0.30]	[0.36]
Gender (1=males)	0.71	0.47	0.45	0.04**	0.02**
	(0.46)	(0.51)	(0.51)	[0.11]	[0.11]
Number of subjects	35	36	40	71	75

Table continues on the next page

Table 1: Descriptive Statistics (Continues)

	(1) APF baseline	(2) CPF treatment	(3) TPF treatment	(1) vs. (2)	(1) vs. (3)
Panel B: Performance pay					
Average tasks solved	27.01	25.76	22.99	0.78	0.32
	(15.33)	(12.82)	(14.01)		
Average total effort	35.87	33.03	29.95	0.38	0.13
	(16.40)	(13.32)	(14.54)		
Average success rate	0.710	0.744	0.700	0.687	0.962
	(0.198)	(0.140)	(0.202)		
Tasks solved in pre-round	18.20	17.30	14.53	0.74	0.16
	(12.98)	(8.56)	(10.22)	[2.73]	[2.61]
Total effort in pre-round	26.95	25.23	22.78	0.53	0.12
	(13.01)	(8.12)	(10.41)	[2.70]	[2.63]
Success rate in pre-round	0.636	0.652	0.589	0.738	0.367
	(0.194)	(0.207)	(0.258)	[0.048]	[0.051]
SAA	1.65	1.47	1.18	0.36	0.01***
	(0.70)	(0.97)	(0.81)	[0.200]	[0.17]
Age	23.98	26.83	23.65	0.03**	0.71
	(4.72)	(6.27)	(3.00)	[1.31]	[0.88]
Years of education	2.98	3.21	2.78	0.50	0.52
	(1.31)	(1.54)	(1.42)	[0.34]	[0.31]
Gender (1=males)	0.60	0.70	0.53	0.39	0.51
	(0.50)	(0.47)	(0.51)	[0.12]	[0.11]
Number of subjects	40	30	40	70	80

Notes: Panel A display statistics from the fixed pay conditions of the experiment. Panel B display statistics from the performance pay conditions of the experiment. Columns (1)-(3) are the APF baseline, CPF treatment and TPF treatment, respectively. Columns (4)-(5) show p-values on comparison of performance between baseline and each treatment. Non-parametric Mann-Whitney U-tests are conducted for the three outcome variables, other comparisons rely on simple t-tests. Standard deviation in parentheses. Standard error in brackets. * p<0.10, ** p<0.05, *** p<0.01.

Table 2: Main Results: Treatment Effects (Fixed Pay)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Tasks solved		Total effort		Success rate	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	1.589*	-5.003***	-0.441	-5.582***	0.004	-0.110***
	(0.954)	(1.070)	(1.427)	(1.267)	(0.008)	(0.018)
TPF treatment	1.947	-0.932	2.110	1.466**	0.007	-0.042
	(1.319)	(1.832)	(1.316)	(0.600)	(0.024)	(0.091)
SAA	1.944**	-0.111	1.379**	0.219	0.044**	0.009
	(0.835)	(0.776)	(0.646)	(0.580)	(0.017)	(0.015)
CPF treatment x SAA		4.323***		3.427***		0.074***
		(0.767)		(0.558)		(0.005)
TPF treatment x SAA		1.748*		0.251		0.030
		(0.968)		(0.620)		(0.046)
N	554	554	554	554	554	554

Notes: Columns (1)-(6) are estimated using Random Effects GLS, with robust standard errors clustered on session. The baseline in each column is subjects who self-assessed their ability to 0 (lowest) in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study and number of tasks solved in the pre-round. Standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Differential Effects of Rank (Fixed Pay)

	(1)	(2)	(3)
Dependent variable:	Tasks solved	Total effort	Success rate
APF baseline	Ref.	Ref.	Ref.
CPF treatment	-2.040**	-1.767	-0.089***
	(0.862)	(1.609)	(0.016)
TPF treatment	-1.589	1.672***	-0.066
	(1.943)	(0.554)	(0.079)
Rank PreRound	-0.642	-0.225	0.011
	(1.448)	(1.843)	(0.022)
CPF treatment x Rank PreRound	2.266***	0.856	0.059***
	(0.837)	(1.060)	(0.004)
TPF treatment x Rank PreRound	2.173*	0.252	0.043
	(1.187)	(1.216)	(0.032)
N	554	554	554

Notes: Columns (1)-(3) are estimated using Random Effects GLS, with robust standard errors clustered on session. The baseline in each column is subjects who would have ranked 0 in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study, SAA and number of tasks solved in the pre-round. Standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 4: Gender Analysis (Fixed Pay)

	(1)	(2)	(3)	(4)	(5)	(6)
Subsample:	Males			Females		
Dependent variable:	Tasks solved	Total effort	Success rate	Tasks solved	Total effort	Success rate
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	-3.086*	-2.966	-0.062	-1.762	-3.586***	-0.035
	(1.779)	(1.832)	(0.067)	(1.642)	(0.953)	(0.056)
TPF treatment	1.832	3.564*	0.056	0.651	0.840	-0.025
	(2.522)	(2.060)	(0.044)	(1.750)	(1.056)	(0.086)
High SAA	2.346	3.448	0.082***	-3.180	-1.153	-0.070
	(2.055)	(2.358)	(0.031)	(5.259)	(3.308)	(0.135)
CPF treatment x High SAA	8.837***	5.382*	0.119	4.637	4.817	0.060
	(3.285)	(2.851)	(0.094)	(5.894)	(3.642)	(0.138)
TPF treatment x High SAA	0.529	-1.573	-0.066*	2.589	2.037	0.048
	(2.311)	(2.742)	(0.040)	(7.457)	(4.794)	(0.168)
Observations	299	299	299	255	255	255

Notes: Columns (1)-(6) are estimated using Random Effects GLS, with robust standard errors clustered on session. Columns (1)-(3) are subsamples of males, columns (4)-(6) are subsamples of females. The baseline in each column is subjects who self-assessed their ability to be low in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, age, years of education, faculty of study, and number of tasks solved in the pre-round. Standard errors are in the parentheses.

* p<0.10, ** p<0.05, *** p<0.01.

Table 5: Main Results: Treatment Effects (Performance Pay)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Tasks solved		Total effort		Success rate	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	0.260	-0.374	-0.837	-1.100	0.045**	0.067
	(0.475)	(2.843)	(1.353)	(1.336)	(0.021)	(0.048)
TPF treatment	0.383	1.008	-1.390	-1.451	0.044***	0.063**
	(0.664)	(1.803)	(0.967)	(2.638)	(0.011)	(0.031)
SAA	0.828	0.775	0.215	0.130	0.044	0.053*
	(1.479)	(2.184)	(1.199)	(0.701)	(0.028)	(0.030)
CPF treatment x SAA		0.420		0.167		-0.013
		(1.858)		(0.995)		(0.021)
TPF treatment x SAA		-0.515		0.024		-0.012
		(1.354)		(1.636)		(0.018)
N	544	544	544	544	544	544

Notes: Columns (1)-(6) are estimated using Random Effects GLS, with robust standard errors clustered on session. The baseline in each column is subjects who self-assessed their ability to 0 (lowest) in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study and number of tasks solved in the pre-round. Standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Differential Effects of Rank (Performance Pay)

	(1)	(2)	(3)
Dependent variable:	Tasks solved	Total effort	Success rate
APF baseline	Ref.	Ref.	Ref.
CPF treatment	0.866	0.225	0.040
	(1.099)	(3.128)	(0.078)
TPF treatment	1.287	-2.142	0.038
	(1.235)	(3.407)	(0.075)
Rank PreRound	1.546	1.273	0.048
	(1.229)	(1.605)	(0.037)
CPF treatment x Rank PreRound	-0.534	-0.734	-0.002
	(0.797)	(1.568)	(0.035)
TPF treatment x Rank PreRound	-0.765	0.305	-0.004
	(1.021)	(1.612)	(0.039)
N	544	544	544

Notes: Columns (1)-(3) are estimated using Random Effects GLS, with robust standard errors clustered on session. The baseline in each column is subjects who would have ranked 0 in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study, SAA and number of tasks solved in the pre round. Standard errors are in the parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 7: Gender Analysis (Performance Pay)

	(1)	(2)	(3)	(4)	(5)	(6)
Subsample:	Males			Females		
Dependent variable:	Tasks solved	Total effort	Success rate	Tasks solved	Total effort	Success rate
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	6.140***	2.185	0.223***	-6.149***	-2.330**	-0.154***
	(1.980)	(6.279)	(0.035)	(0.690)	(1.124)	(0.033)
TPF treatment	2.628*	-1.661	0.165***	0.605	0.207	0.009
	(1.538)	(5.289)	(0.028)	(1.112)	(1.172)	(0.040)
High SAA	9.578***	6.415	0.223***	-9.300***	-8.017***	-0.197***
	(1.339)	(6.163)	(0.021)	(1.040)	(2.187)	(0.055)
CPF treatment x High SAA	-11.056***	-8.577	-0.222***	10.658***	4.099	0.129
	(2.156)	(6.837)	(0.031)	(3.147)	(4.267)	(0.081)
TPF treatment x High SAA	-3.593	-0.988	-0.135***	0.197	0.781	0.008
	(2.721)	(6.865)	(0.022)	(2.163)	(2.873)	(0.064)
Observations	324	324	324	220	220	220

Notes: Columns (1)-(6) are estimated using Random Effects GLS, with robust standard errors clustered on session. Columns (1)-(3) are subsamples of males, columns (4)-(6) are subsamples of females. The baseline in each column is subjects who self-assessed their ability to be low in the APF baseline. TThe dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, age, years of education, faculty of study, and number of tasks solved in the pre-round. Standard errors are in the parentheses.

* p<0.10, ** p<0.05, *** p<0.01.

Appendix

A1. Tables

Table A1-1: Main Treatment Effects (Fixed Pay) using RE Tobit

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Tasks solved		Total effort		Success rate	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	1.531	-5.146*	-0.450	-5.573*	0.003	-0.112*
	(1.476)	(2.687)	(1.712)	(3.168)	(0.034)	(0.062)
TPF treatment	1.933	-1.002	2.127	1.511	0.007	-0.043
	(1.435)	(2.576)	(1.665)	(3.038)	(0.033)	(0.060)
SAA	1.972**	-0.115	1.389	0.240	0.045**	0.009
	(0.874)	(1.259)	(1.014)	(1.485)	(0.020)	(0.029)
CPF treatment x SAA		4.377***		3.416*		0.075**
		(1.485)		(1.751)		(0.034)
TPF treatment x SAA		1.784		0.232		0.030
		(1.459)		(1.720)		(0.034)
Log-likelihood	-1761.8	-1757.5	-1836.9	-1834.4	210.2	212.6
Left-censored observations	9	9	4	4	9	9
N	554	554	554	554	554	554

Notes: Columns (1)-(6) are estimated using Random Effects Tobit. The baseline in each column is subjects who self-assessed their ability to 0 (lowest) in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study and number of tasks solved in the pre-round. Standard errors are in the parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table A1-2: Questionnaire Responses (Fixed Pay)

	(1)	(2)	(3)	(4)	(5)	(6)
Question asked:	Stressful		Joyfulness		Motivated	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	-0.176	-0.434*	0.015	-0.483	-0.131**	-0.473**
	(0.088)	(0.207)	(0.085)	(0.242)	(0.038)	(0.176)
TPF treatment	0.098	-0.030	0.513**	0.812***	0.093	0.069
	(0.128)	(0.204)	(0.158)	(0.149)	(0.235)	(0.136)
SAA	-0.152*	-0.239***	0.578**	0.553***	0.269**	0.192**
	(0.060)	(0.038)	(0.186)	(0.072)	(0.099)	(0.060)
CPF treatment x SAA		0.168		0.344**		0.228*
		(0.104)		(0.131)		(0.105)
TPF treatment x SAA		0.079		-0.226		0.004
		(0.068)		(0.137)		(0.075)
N	111	111	111	111	111	111

Notes: Columns (1)-(6) are estimated using OLS, with robust standard errors clustered on session. The dependent variable in columns (1)-(2) is a question of whether they felt the task was stressful (scaled from 0 to 3). The dependent variable in columns (3)-(4) is a question of whether they liked the work task (scaled from 0 to 4). The dependent variable in columns (5)-(6) is a question of whether they were motivated by the feedback (scaled from 0 to 4). All columns include the following control variables: gender, age, years of education, and faculty of study. Standard errors are in the parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table A1-3: Main Treatment Effects (Performance Pay) using RE Tobit

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Tasks solved		Total effort		Success rate	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	0.475	0.432	-0.802	-0.945	0.049	0.078
	(1.762)	(3.784)	(1.976)	(4.237)	(0.039)	(0.083)
TPF treatment	0.528	1.563	-1.355	-1.303	0.046	0.070
	(1.603)	(3.446)	(1.798)	(3.856)	(0.035)	(0.075)
SAA	0.910	1.102	0.229	0.199	0.045*	0.058
	(1.052)	(1.754)	(1.180)	(1.964)	(0.023)	(0.038)
CPF treatment x SAA		0.051		0.093		-0.018
		(2.098)		(2.351)		(0.046)
TPF treatment x SAA		-0.771		-0.050		-0.016
		(2.099)		(2.352)		(0.046)
Log-likelihood	-1740.6	-1740.5	-1730.4	-1730.4	205.1	205.2
Left-censored observations	12	12	2	2	12	12
N	544	544	544	544	544	544

Notes: Columns (1)-(6) are estimated using Random Effects Tobit. The baseline in each column is subjects who self-assessed their ability to 0 (lowest) in the APF baseline. The dependent variables are tasks solved, total effort (correct and incorrect attempts) and the rate of tasks solved over total effort, respectively. All columns include the following control variables: round trend, gender, age, years of education, faculty of study and number of tasks solved in the pre-round. Standard errors are in the parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

Table A1-4: Questionnaire Responses (Performance Pay)

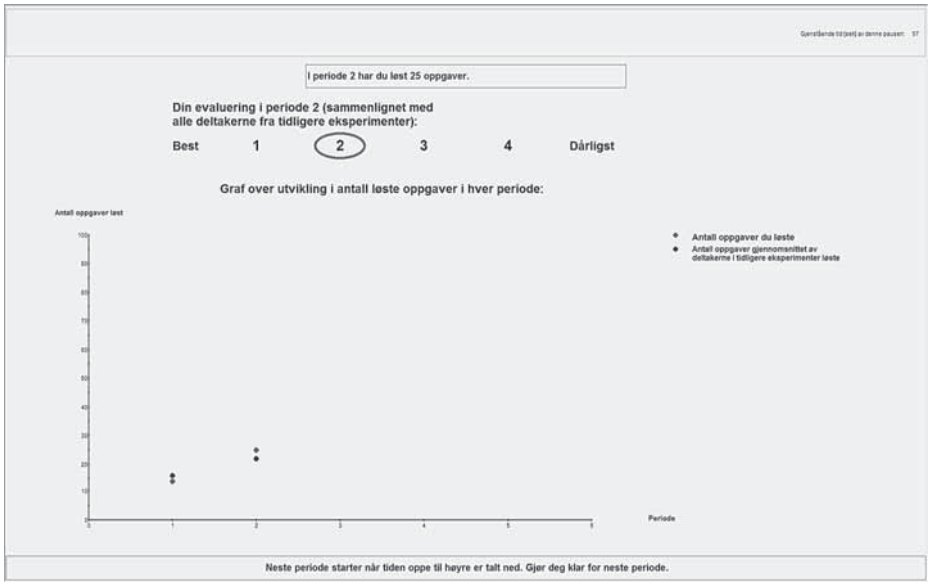
	(1)	(2)	(3)	(4)	(5)	(6)
Question asked:	Stressful		Joyfulness		Motivated	
APF baseline	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
CPF treatment	-0.058	-0.806***	0.049	0.013	-0.386**	-0.526*
	(0.072)	(0.184)	(0.092)	(0.468)	(0.114)	(0.253)
TPF treatment	0.262**	0.030	0.219**	-0.127	-0.053	-0.177
	(0.084)	(0.229)	(0.068)	(0.554)	(0.083)	(0.170)
SAA	-0.247*	-0.485***	0.444***	0.355	0.512***	0.449**
	(0.116)	(0.052)	(0.107)	(0.290)	(0.071)	(0.155)
CPF treatment x SAA		0.475***		0.015		0.087
		(0.105)		(0.302)		(0.206)
TPF treatment x SAA		0.113		0.254		0.081
		(0.148)		(0.367)		(0.118)
N	109	109	109	109	109	109

Notes: Columns (1)-(6) are estimated using OLS, with robust standard errors clustered on session. The dependent variable in columns (1)-(2) is a question of whether they felt the task was stressful (scaled from 0 to 3). The dependent variable in columns (3)-(4) is a question of whether they liked the work task (scaled from 0 to 4). The dependent variable in columns (5)-(6) is a question of whether they were motivated by the feedback (scaled from 0 to 4). All columns include the following control variables: gender, age, years of education, and faculty of study. Standard errors are in the parentheses.
* p<0.10, ** p<0.05, *** p<0.01.

A2. Experimental Details and Instructions

Figure A2-1: Illustration of Task Design

Figure A2-2: Illustration of Feedback Design



Experimental Instructions (translated from Norwegian)

Welcome to this experiment, and thank you for choosing to participate.

In this experiment we want you to do some work for us.

Tasks: You are going to solve math questions. You will solve tasks in 6 rounds, each round lasts for 8 minutes. A clock on the screen will tell you the time left in each round. There will be a break for half a minute between each round.

For each task you solve you get one point that will be added to your total sum for the round. As your employer we want you to solve as many tasks as possible. You must answer a task correctly in order to proceed to the next task, and you will be asked to try again if your answer is wrong.

Duration:

In total the experiment will take about 60 minutes.

Rules:

You are allowed to use your mobile phone to surf the Internet if you need a break, but you cannot use it as a calculator. There are also some newspapers available in the room. No further remedies are permitted to solve the tasks - including calculators, electronic calculators, or pen&paper. You may not leave the room until after the experiment is finished, nor communicate with other participants or with others using your mobile phone. If you do not comply with these rules, you will not be paid and you will be disqualified for participation.

Questionnaire:

In the end you will be asked to fill out a short questionnaire. These answers are anonymous.

Payment for participation:

You will earn ECU 1000 (experimental currency) for participating in the experiment. In addition, you will earn more depending on how many tasks you solve. For each correctly solved task you will earn ECU 10, which will be added to your total earnings at the end of the experiment.

ECU 10 is equivalent to 1 Norwegian krone.

Test round:

Start off by practicing the task for 1 minute to familiarize yourself with the task. Try to put in a correct answer and a wrong answer to see what happens on both occasions.

Smells Like Team Spirit: An Experiment on Relative Performance Feedback*

William Gilje Gjedrem and Ola Kvaløy[1]

Abstract: Between and within firms, work teams compete against each other and receive feedback on how well their team is performing relative to their benchmarks. In this paper we investigate experimentally how teams respond to relative performance feedback (RPF). We find that when subjects work under team incentives, then RPF on team performance increases the teams' average performance by almost 10%. The treatment effect is driven by higher top performance, as this is almost 20% higher when the teams receive RPF compared to when the teams only receive absolute performance feedback (APF). The experiment suggests that top performers are particularly motivated by the combination of team incentives and team RPF. In fact, team incentives trigger significantly higher top performance than individual incentives when the team is exposed to RPF. We also find notable gender differences. In particular, females respond negatively to individual RPF, but even more positively than males to team RPF.

*We thank Petra Nieken, Mari Rege, seminar participants at the Stavanger Workshop on Incentives and Motivation, Rady School of Management, University of Cologne, University of Oslo, ESA meeting in Bergen, and the Choice Lab at the Norwegian School of Economics for helpful comments and suggestions. Financial support from the Norwegian Research Council (227004) is gratefully acknowledged.
[1]University of Stavanger, UiS Business School, 4036 Stavanger, Norway (e-mail: ola.kvaloy@uis.no).

1. Introduction

People prefer high rank to low rank. Even when rank is independent from monetary outcomes, people are willing to take costly actions in order to climb the ladder. "….rank among our equals, is, perhaps, the strongest of all our desires" wrote Adam Smith in 1759. Modern organizations utilize this basic human insight by providing employees with feedback on their relative performance in order to motivate them to work harder.

However, although rank and relative performance feedback (RPF) is such a basic ingredient in competitive environments, it is only recently that economists have systematically studied how people respond to RPF. The early economics literature on relative performance evaluation studied the effect of connecting rank to monetary incentives (see Lazear and Rosen (1981) seminal contribution on rank order tournaments). But it has now been demonstrated, through controlled experiments in the lab and in the field, that RPF *per se* affects individual behavior. For example, Blanes i Vidal and Nossol (2011), Kuhnen and Tymula (2012) and Charness, Masclet, and Villeval (2014), find strong performance improvements in situations where RPF is provided, while Hannan, Krishnan, and Newman (2008), Gjedrem (2015), and Azmat and Iriberri (2016) find significant context specific effects of RPF.[1]

The experimental literature on RPF has so far concentrated on individual behavior and individual feedback. However, not only individuals receive RPF, but also groups of individuals, like firms, or teams within firms, who compete against each other and receive feedback about their relative performance. Sales teams or R&D teams, for instance, are benchmarked against similar teams in other firms. Moreover, firms often set up internal competitions between teams in order to sell more or innovate more (see e.g., Birkinshaw, 2001; Marino & Zábojnik, 2004; Baer, Leenders, Oldham, & Vadera, 2010). Successful teams are typically compensated by some monetary rewards, but team competitions *per se* may also be motivating.

[1]There are also studies that do not find any positive effects of RPF. Eriksson, Poulsen, and Villeval (2009), Guryan, Kroft, and Notowidigdo (2009) and Bellemare, Lepage, and Shearer (2010) find that removing RPF positively affected productivity.

Hence, in this paper we contribute to the existing literature by investigating how teams respond to relative performance feedback. We also study whether teams suffer from free-riding activities and to what extent RPF mitigates this problem. There are several reasons why people might respond differently to team feedback compared to individual feedback. The joy of winning together with a team might be different from the joy of winning alone. Similarly, the costs of losing as a team might be different from the costs of losing alone. Moreover, repeated RPF may create peer effects within the team, which again creates a different response to team RPF compared to individual RPF. We thus investigate to what extent and under which conditions teams respond to RPF. We also compare how individuals respond differently to team RPF than to individual RPF.

We do this by conducting a controlled laboratory experiment consisting of six treatments. In each treatment, subjects work on a real-effort task for six periods. We primarily vary treatments along two dimensions: team or individual incentives, and team or individual feedback. However, to establish a "baseline" of performance, we have two treatments in which subjects only receive absolute performance feedback. Under RPF, individuals (teams) are always compared with two other individuals (teams), i.e. after each period, each individual or team is ranked as either number 1, 2 or 3. Each team consists of three subjects, so each subject earns one third of total team output when provided with team incentives. The monetary outcomes are independent from feedback rankings.

Our design enables us to study systematically how performance feedback interacts with the level in which incentives and feedback are provided. Moreover, we can study heterogeneous effects: Does RPF to teams trigger higher top performance or does it boost the performances of the weakest links? And what about gender? Previous research have shown that genders tend to behave differently, and that females shy away from competition (Niederle & Vesterlund, 2007; Marianne, 2011). Moreover, it has been shown that males respond more positively than females to individual RPF (e.g., Azmat & Iriberri, 2016). Does the same apply for team RPF?

Our main results can be summarized as follows: We find that when subjects are exposed to team incentives, then RPF on how their team is doing

compared to two other teams increases the team's average performance by almost 10 percent. The treatment effect is driven by higher top performances. The average individual performance of the best performance within each team is almost 20 % higher when the teams receive RPF compared to when the teams only receive APF. These effects more or less disappear under individual incentives and/or individual RPF. Our experiment thus suggests that some subjects are particularly motivated by the combination of team incentives and team RPF. In fact, team incentives trigger significantly higher top performance than individual incentives, when subjects are exposed to team RPF.

We also find some interesting gender effects. Females respond negatively to individual RPF, but even more positively than males to team RPF. For males, team incentives have a strong negative effect compared to individual incentives, unless it is accompanied by team RPF. For females, the incentives do not matter to the same degree, and team RPF has a strong positive effect regardless of the incentive system.

Our results can contribute to explaining why team incentives are so common, despite the well-known free-rider problem. A majority of firms in the US and UK report some use of teamwork in which groups of employees share the same goals or objectives, and the incidence of team work and team incentives has been increasing over time (see Lazear and Shaw (2007) and Bandiera, Barankay, and Rasul (2013), and the references therein). Team incentives are puzzling because the individual incentive effect is quite small, and the temptation to free-ride on peers' effort is high (Holmstrom, 1982). Alchian and Demsetz (1972) note in their classic book on team production that "If one could enhance a common interest in non-shirking in the guise of team loyalty or team spirit, the team would be more efficient. The difficulty, of course, is to create economically that team spirit and loyalty". Empirical research shows, however, that team incentives do surprisingly well, and it has been hard to actually identify strong free-rider effects.[2] Theorists have also

[2]A range of studies employing different empirical approaches have identified mixed effects of team incentives. In some field studies, there is an overall performance improvement of team incentives, relative to individual incentives or relative to an absence of team incentives, see e.g., Knez and Simester (2001), Hamilton, Nickerson, and Owan (2003) and Boning, Ichniowski, and Shaw (2007). On the other

investigated more formally how firms can create the kind of team spirit that Alchain and Demsetz call for. Kandel and Lazear (1992) introduces a peer pressure function and discusses how firms can manipulate peer pressure by e.g. investing in team spirit building activities. Akerlof and Kranton (2000, 2005) incorporates identity into an otherwise standard utility function. They discuss how teams or firms can transform the workers' identity from "outsiders" to "insiders" by creating common goals that each individual shares with their team or firm. Relative performance feedback to teams can be seen as a means of creating the kind of team spirit or identity discussed by these theorists. It gives the team a common goal and potentially creates a feeling of 'us versus them'. Social psychologists have shown that intergroup comparisons can enhance the salience of the group objective – a common goal – and also how closely one identifies with the group (Brewer & Kramer, 1986). In these respects, our findings are interesting: Team incentives alone give rise to a free-rider problem also in our experiment. Without team RPF, individual incentives does significantly better than team incentives. However, introducing team RPF more than offsets the problem. Team incentives then do slightly better than individual incentives, and significantly so for top performers.[3]

hand, van Dijk, Sonnemans, and van Winden (2001), and Vandegrift and Yavas (2011), using controlled laboratory experiments to study team incentives, do not find any overall change in performance. van Dijk et al. (2001) do find that some subjects improve, but this is offset by others who free-ride. Still others find a negative effect of introducing team incentives. In an early field experiment, Erev, Bornstein, and Galili (1993) find a significant decrease in effort once team incentives were introduced. However, when also adding competition for a prize between the teams, performance levels resumed to the same level as with individual incentives. Nalbantian and Schotter (1997) find extensive shirking behavior under different types of team incentives, but competition between teams for a fixed price increases performance significantly.

[3]It should be noted that there are not only so-called behavioral or non-monetary reasons why team incentives might work. Team incentives can exploit complementarities and foster cooperation (Holmström & Milgrom, 1990; Itoh, 1991, 1992; Macho-Stadler & Pérez-Castrillo, 1993). Team incentives can also be desirable in repeated settings, as it strengthens implicit incentives, see Che and Seung-Weon (2001) and Kvaløy and Olsen (2006). However, experimental investigation of team incentives, like the one present in the paper, abstract from such technological team effects.

To the best of our knowledge, no one has yet studied the effect of relative performance feedback to teams in a laboratory experiment. However, our paper relates to recent papers in the public good literature that have investigated whether individual contribution to a group's public good might increase if the contribution is compared with the contribution in other groups. Based on Turner (1975), Böhm and Rockenbach (2013) argue that intergroup comparisons can motivate group members to increase the contribution to their own group. That is, even in the absence of monetary incentives, group members may engage in social competition to boost their group identity. Indeed, Tan and Bolle (2007), Burton-Chellew and West (2012), and Böhm and Rockenbach (2013) find support for this hypothesis. Intergroup competition increases contributions to public goods.[4]

The positive effects of intergroup comparison clearly resemble and support our findings on team RPF. There are, however, important differences in the designs. First, we present a real effort experiment: Subjects have to work on a specific task, whereas in the public goods experiments (PGEs) effort is purely an allocation of money. Our experiment does not have a defined maximum contribution to the public good, unlike PGE where contribution is restricted to the endowment amount. Second, effort costs are different. In public good games, the Nash equilibrium is to contribute nothing, while in real-effort experiments, team members should exert effort to the extent that marginal revenue equals marginal costs. PGEs typically also involve a constant marginal cost, while in real effort experiments, the marginal cost is likely to increase as effort increases (work becomes more exhaustive). Hence, the strong effects we find on relative performance feedback to teams do not replicate, but nicely complement the positive effects of intergroup comparison demonstrated in the PGE literature.

The results also contrast with a recent field study by Bandiera et al. (2013). They find that ranking teams reduces overall performance, as lower ranked teams decrease productivity. This study, however, endogenously allows subjects to select into teams, whereas we assign subjects exogenously into teams. Our lab experiment also allows us to abstract from contaminating

[4]A similar experiment by Sausgruber (2009) do not find any effect.

HRM policies or technological complementarities that may often arise in the field.

The rest of the paper is organized as follows. In section 2 we present the experimental design. In section 3 we present the results, while section 4 concludes.

2. Experimental Design

2.1 Task

Subjects work on a real-effort task of decoding numbers into letters, used in several other related experiments (e.g., Charness et al., 2014). Specifically, subjects have a list of letters each assigned with a corresponding number, and the task is to decode given sequences of four numbers into their respective letter. The experimental session consists of six working stages, each lasting five minutes. There is a break in between each stage, and during these subjects receive feedback (explained below). Participants earn a 100 NOK show-up fee. In addition, they can earn money by solving tasks, explained in the next subsection.

There are two main reasons why we have chosen this particular task. First, it requires no prior knowledge and is easy to understand. Second, we expect the task to be boring and tiresome, generating disutility of effort. To ensure disutility of effort we allow subjects to engage in alternative activities during the experiment, such as using their mobile phones for internet surfing. We require them to remain in their seat and refrain from communicating with other participants, but tell them they can freely allocate their time to whatever suits them the most. Distracting activities are typically also present in the workplaces, so if anything these activities only make it more similar to the field. The task also provides a precise measure of output, which is our productivity indicator. Each session has the same sequence of number-decoding tasks. Subjects cannot proceed to a new task before the current task is correctly solved.

2.2 Treatments

We primarily vary treatments along two dimensions: team or individual incentives, and team or individual feedback. However, to establish a

"baseline" of performance, we have two treatments in which subjects only receive absolute performance feedback. Feedback always concerns performance in the previous stage.[5] In any treatment, subjects always learn their individual absolute performance. In any team treatment, subjects always learn the total absolute performance of their team. When subjects receive RPF, individuals (teams) are always ranked relative to two other individuals (teams), and they are ranked relative to the same individuals (teams) throughout the experiment (randomly assigned). Team members work independently on the tasks, and there are no complementarities in production. Teams also remain unchanged throughout the experiment (randomly assigned).

The piece-rate for a correctly solved task is 1 NOK. In the individual incentive treatments, subjects earn the piece-rate multiplied with total number of tasks they solve. In the team incentive treatments, subjects earn the piece-rate multiplied with one third of the total number of tasks the team solved, i.e. all team members earn the same. Hence, monetary outcomes only depend on the number of tasks subjects or teams solve, not on feedback ranks.

Treatment names are structured as follows: It first denotes whether feedback is absolute (APF) or relative (RPF), then whether there are individual (ind) or team (team) incentives, and finally whether the level of feedback is on individuals (ind) or teams (team). Below we introduce treatments gradually. We start by keeping one dimension fixed and only present treatments that contain RPF first. These are displayed in Table 1.

Table 1: Summary of RPF Treatments

RELATIVE PERFORMANCE FEEDBACK	Individual RPF	Team RPF
Individual incentive	RPF-ind-ind	RPF-ind-team
Team incentives	RPF-team-ind	RPF-team-team

[5]We do not display any aggregate information based on several previous stages.

In the *RPF-ind-ind* treatment, subjects earn individual incentives and receive individual RPF. The individual RPF consists of performance information about two other participants in the session. Their performance is ranked (from 1 to 3) and they learn how many tasks the other two subjects solved. In addition to the show-up fee, subjects earn the piece-rate multiplied with the number of tasks they solve.

In *RPF-ind-team* treatment, subjects still earn individual incentives, but RPF is changed and now concerns teams rather than individuals. The team RPF consists of performance information about two other teams in the session. The team's performance is ranked (from 1 to 3) and they learn how many tasks the other two teams solved. In addition to the show-up fee, subjects earn the piece-rate multiplied with the total number of tasks they solve.

In the *RPF-team-ind* treatment, subjects still receive individual RPF, but incentives are changed and now concern team outputs rather than individual outputs. The individual RPF consists of individual performance information about the two other team members. Their performance is ranked (from 1 to 3) and they learn how many tasks the other two subjects solved. In addition to the show-up fee, subjects earn the piece-rate multiplied with one third of the total number of tasks their team solves.

RPF-ind-ind and RPF-team-ind are referred to as individual RPF treatments.

In the *RPF-team-team* treatment, subjects receive both team RPF and team incentives, rather than individual RPF and individual incentives. The team's performance is ranked (from 1 to 3) and they learn how many tasks the other two teams solved. In addition to the show-up fee, subjects earn the piece-rate multiplied with one third of the total number of tasks their team solves.

Finally, we introduce our "baseline" conditions, where we do the same variations as in the previous table, only with APF instead of RPF. These are displayed in Table 2 and explained below.

Table 2: Summary of APF Treatments

ABSOLUTE PERFORMANCE FEEDBACK	Individual APF	Team APF
Individual incentive	APF-ind-ind	
Team incentives		APF-team-team

In the *APF-ind-ind* treatment, subjects earn individual incentives and receive individual APF. Importantly, they do not learn anything about the performance of any others. In addition to the show-up fee, subjects earn the piece-rate multiplied with the total number of tasks they solve.

In the *APF-team-team* treatment, subjects earn team incentives and receive team APF, rather than individual incentives and individual APF. In addition to the show-up fee, subjects earn the piece-rate multiplied with one third of the total number of tasks their team solves.

We have not collected data for the two cells left empty in Table 2, as the primary use of APF treatments is to establish "baseline" performances. Thus, we have only included APF treatments that are of main interest to compare with RPF treatments. The empty cells are also less realistic. For example, in an *APF-team-ind* treatment, subjects would only receive individual performance feedback, but then it makes no sense to make their earnings depend on other (unknown) team members. Notice also that all treatments actually include APF, and hence RPF is an additional piece of information in the RPF treatments.

2.3 Procedures

The experiment was conducted at the University of Stavanger, Norway, in March 2015 and November 2015. We ran three sessions of each treatment over four days in March, except for the three sessions in RPF-ind-team that

we ran in November.[6] A session had up to 23 participants, and treatments with RPF or teams required a total number of participants that could be divided by three (and precisely 18 participants in RPF-ind-team and RPF-team-team). We recruited subjects through their student email accounts and posters on the University campus, and they signed up using the recruitment program Expmotor.[7] The student pool consists of a variety of students from three faculties: the faculty of Science and Technology, the faculty of Social Sciences, and the faculty of Arts and Education. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007).

We randomly seated subjects when they arrived in the computer lab. Each desk had a paper with written instructions, and we read the instructions aloud before the start of the experiment (instructions attached in the appendix). Then they worked on the task and received feedback during the breaks. Once the experiment concluded, we informed subjects about their total output and earnings. Then they completed a short questionnaire, where we asked for basic demographic details and elicited their ex post perceptions of the experiment. Specifically, we asked them how motivated they were to do the tasks, how they felt right now, and whether they thought the information in-between each stage affected them. They answered these questions on a scale from -5 to 5.

The average earnings of the 350 participants were 289 NOK (about $35), which consisted of 100 NOK show-up fee and 189 NOK performance-related pay. Each session lasted about 50 minutes. Total number of participants in each respective treatment was 68 (29 females, 39 males) in APF-ind-ind, 55 (22 females, 33 males) in RPF-ind-ind, 53 (16 females, 37 males) in RPF-ind-team, 56 (23 females, 33 males) in APF-team, 54 (27 females, 27 males) in RPF-team-team, and 63 (27 females, 36 males) in RPF-team-ind.[8]

[6]We have no reason to believe that the different month for this treatment would cause any differences per se, and predetermined characteristics of subjects participating in this treatment are very similar to the other treatments, as can be seen in the appendix Table A1-1.

[7]Developed by Erik Sørensen and Trond Halvorsen at the Norwegian School of Economics (NHH).

[8]Administrative revision found that three subjects participated twice (disregarding explicit information about this being strictly prohibited), two in RPF-ind and one in

2.4 Analysis Plan and Behavioral Expectations

Our design allows us to investigate many aspects of team performance and the effects of relative performance feedback. In this section we highlight which comparisons we want to focus on in the result section, and look into whether theory or related empirical research offers some guidance on what to expect from the experiment.

Standard economic theory offers only a few predictions on the outcome of the experiment. A straightforward prediction is that individual incentives always outperform team incentives, as free-riding is expected to occur under team incentives (Holmstrom, 1982). Another prediction is that relative performance feedback should not affect performance whatsoever, as this implies no change in the incentive structure in terms of monetary outcomes.

Moving outside standard economic theory, including behavioral related parameters into the utility maximizing function, relations become more complex. Then factors such as a person's self-confidence and social aspects in the workplace may affect the final effort decision. These parameters are often ambiguous in terms of how they affect motivation and productivity. For example, relative feedback may be intriguing in a competitive manner, but on the other hand serves to worsen the self-confidence of those who perform poorly. Given the complexity of these relations, we do not attempt to construct new theoretical models to predict outcomes of our experiment. However, we note that such parameters are needed to explain many of the empirical findings in papers that we cite.

Our main interest is whether relative performance feedback on teams influences the performance of subjects. Several papers have shown that this may be the case when provided at an individual level (Blanes i Vidal & Nossol, 2011; Kuhnen & Tymula, 2012; Charness et al., 2014); however, to our knowledge, no one has studied this explicitly on a team level. We do this by comparing how teams receiving RPF (with and without team incentives) perform relative to teams or individuals who only receive APF. Given the

RPF-ind-team. These subjects are not part of the given number of participants. In addition, these subjects could have affected their peer groups (2+2 subjects in RPF-ind and 8 subjects in RPF-ind-team). We still include these subjects, but results are robust to excluding them.

empirical literature of RPF on individual level, we may expect similar outcomes in our experiment, though we do not make any predictions ex ante.

Our design also allows us to study whether people in teams do free ride on each other, as theory strongly predicts. Surprisingly few controlled experiments have been able to identify this issue in practice.[9] We expect to see that individuals who work under individual incentives and receive APF outperform teams that work under team incentives and receive APF, as theory predicts. If we do find a free rider effect, it is also interesting to see whether the addition of RPF could mitigate this effect.

In the regression analysis, we also study the 1st and 2nd stage separately. The 1st stage is a "kick-off" stage as any treatment effect of RPF is driven by the knowledge about future feedback, and not a response to the feedback itself (as found in e.g., Blanes i Vidal & Nossol, 2011). The 2nd stage is the first working stage after any feedback is provided, and the cleanest way to identify any treatment effects of RPF.

For RPF to be beneficial, the effect on performance should persist. We check whether this is the case for any treatment effect in our experiment, by including a subsection on persistency.[10] We split data into two parts, studying the first and last half of the experiment separately. Moreover, we also do an analysis where we control for initial effort in the 1st stage, to see if any treatment effects develop differently in the "kick-off" stage compared to the remaining stages. If so, this suggests that people may respond to the initial knowledge about future RPF in one way, but when they actually receive the feedback they respond in a different way.

Our design also allows us to study whether team incentives and team RPF interact with each other in any way. Even though we might observe that adding one of these conditions affects performance in one certain way, the outcome may be different when both conditions apply. Existing empirical research offers little guidance on what to expect from this.

[9]See footnote 2 for a brief summary of the literature on this.
[10]Real and long-lasting persistence, though, is not possible to test for in a laboratory experiment.

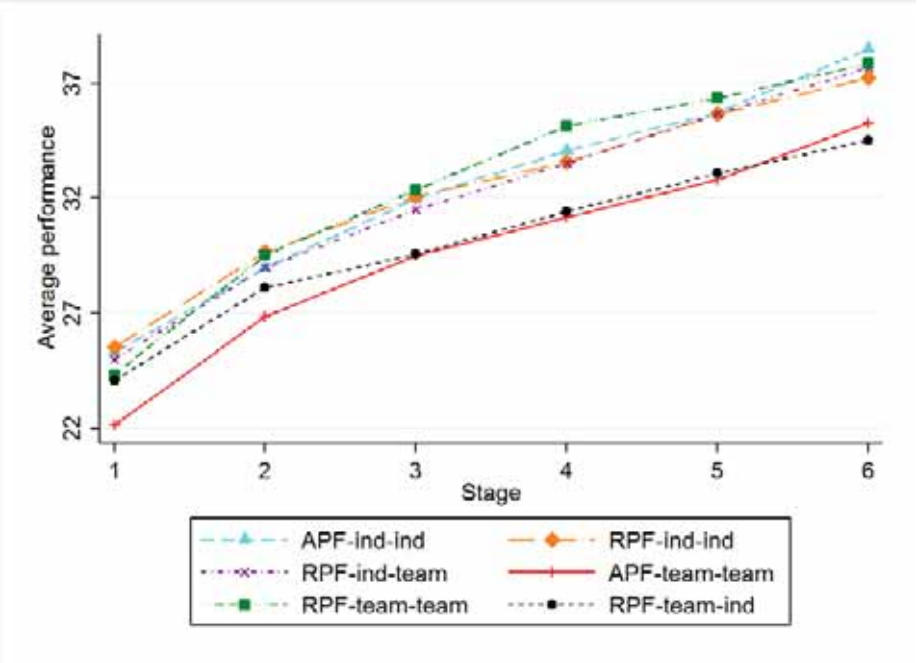
We finish the empirical analysis by looking at differential effects depending on gender and high and low performing subjects. As the empirical literature has shown that males and females tend to respond differently to competition in general (see, Marianne, 2011, for a recent review), and RPF in particular (e.g., Azmat & Iriberri, 2016), we consider it interesting to study whether this transfers into an environment that focuses on the collective performance of a team rather than individual performances. Furthermore, if there are overall treatment differences, an important aspect is whether these differences stem from higher performances at the bottom or top of the performance distribution. These differential analyses give us an idea of the sort of subjects who respond to the different feedback conditions.

3. Experimental Results

3.1 Main Observations

Figure 1 displays average performance of subjects across stages.

Figure 1: Average Performance across Stages



We first want to establish whether relative performance feedback affects individual behavior in teams. To answer this question, we compare the performance of subjects in RPF-team-team to subjects in APF-team-team. From Figure 1 we see a clear treatment effect. More formally, the average performance in RPF-team-team (32.6 tasks solved) is significantly greater (Mann Whitney U-test (MW): p=0.09, Randomization test (RT): p=0.02) than in APF-team-team (29.6), see Table 3. [11,12] The performance is about 10% better in RPF-team-team compared to APF-team-team. The effect seems to be present from the very beginning of the experiment, suggesting that

[11]We use Mann-Whitney U-test and Randomization test when comparing means throughout this section, unless otherwise specified. When based on the performance across all stages, we use each subject's average performance across all of these. The Randomization tests are based on 200.000 simulations.
[12]The difference (26.9 vs. 29.5) in the 2nd stage is also significant, MW: p=0.04 and RT: p=0.01.

knowledge about the future performance feedback *per se* is enough to induce subjects to higher effort.

Observation 1: Subjects working under *team incentives* perform better when they also receive team RPF compared to only team APF.

Table 3: Team Incentives and RPF

	Average Performance (SD)			Mann-Whitney z-Statistics (p-value)	
	APF-team-team (1)	RPF-team-team (2)	RPF-team-ind (3)	(1) vs (2)	(1) vs (3)
Stage 1	22.16 (5.77)	24.35 (6.81)	24.10 (4.76)	-1.48 (0.138)	-1.36 (0.175)
Stage 2	26.86 (4.49)	29.52 (6.23)	28.11 (5.06)	-2.06 (0.040)**	-1.22 (0.223)
All stages	29.63 (5.56)	32.60 (7.63)	30.13 (5.51)	-1.69 (0.090)*	-0.21 (0.834)
N	57	54	63	111	120

Notes: Mann-Whitney pairwise test. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Having established that team RPF works well under team incentives, one may ask whether the same applies to individual RPF under team incentives. To investigate this, we compare APF-team-team to RPF-team-ind.[13] The average performance in RPF-team-ind (30.1) is statistically the same (MW: p=0.83, RT: p=0.62) as in APF-team-team (29.6). Hence, there seems to be no difference in performance when subjects receive individual RPF under team incentives.[14,15]

Next, compare directly the two RPF treatments under team incentives. After the first stage (i.e. from stage 2 and onward) the average performance in RPF-team-team (34.3) is significantly higher (MW: p=0.09, RT: p=0.03) than in RPF-team-ind (31.3).[16] Thus, under team incentives, subjects seem to respond more positively to team RPF than individual RPF. Notice that this effect is only present from stage 2 and onwards, suggesting that the effect is driven by different reactions to the feedback. Whereas the treatment difference was immediate between APF-team-team and RPF-team-team, it is a direct response to the feedback between RPF-team-team and RPF-team-ind. This suggests that subjects in the RPF-team-ind respond negatively to knowledge about their teammate's performance, relative to how subjects in the RPF-team-team respond to knowledge about the performance of other teams.

Observation 2: Subjects working under *team incentives* do not perform better when they also receive individual RPF compared to only team APF.

[13]One could argue that the feedback information in RPF-team-ind is (at least implicitly) available in APF-team-team as well, as there are only three subjects in one team, and they know both their own performance and the total of the team.
[14]In this comparison there is only one condition that changes. As subjects in the RPF-team-ind learn everything that subjects in the APF-team-team learn, they only change is the additional individual RPF.
[15]A different approach is to compare team averages rather than subject averages. In such analysis, the difference between APF-team-team and RPF-team-team over all periods is even more significant with p=0.026 (based on 38 observations).
[16]Including stage 1 leads to an insignificant difference (MW: p=0.13, RT: p=0.05), but considering the development in performance seen in Figure 1, it is more appropriate to compare performance from stage 2 and onwards, especially if we want to capture the reactions after they observe feedback.

Moreover, subjects who receive team RPF perform better than those who receive individual RPF.

What about the "pure" effects of changing incentives, without the involvement of RPF? We may compare the effects of this by comparing the two APF treatments.[17] Here, individual incentives do better than team incentives. The average performance in APF-ind-ind (32.4) is significantly higher (MW: p=0.01, RT: p=0.01)[18] than in APF-team-team (29.6), see Table 4. These results indicate the presence of a free-rider problem: Subjects working under individual incentives solve, on average, almost 10% more tasks than those working under team incentives.

Table 4: Free-Rider Problem

	Average Performance (SD)		Mann-Whitney z-Statistics
	APF-ind-ind	APF-team-team	(p-value)
	(1)	(2)	(1) vs (2)
Stage 1	25.31 (4.90)	22.16 (5.77)	3.16 (0.002)***
Stage 2	28.97 (5.32)	26.86 (4.50)	2.34 (0.020)**
All stages	32.43 (6.06)	29.63 (5.56)	2.57 (0.010)**
N	68	57	125

Notes: Mann-Whitney pairwise test. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Observation 3: Subjects solve less task under team incentives than under individual incentives.

[17]Strictly speaking, changing from individual to team incentives and from individual to team APF is a multiple change of conditions. However, as argued in previous section, there is no realistic middle way of only changing incentives or only changing to team APF.

[18]If we only study the difference in stage 2 (29.0 vs. 26.9), the difference is also significant with MW: p=0.02 and RT: p=0.02. Using team averages, as described in footnote 15, the difference over all periods is also significant with p=0.090.

An interesting comparison, although a change of multiple conditions, is to compare the average performance of subjects in APF-ind-ind (32.4) to RPF-team-team (32.6). Statistical tests reveal no significant performance difference between them (MW: p=0.65, RT: p=0.89), see also Table A1-2. Hence, moving from APF-ind-ind to APF-team-team (step 1) revealed a free-rider problem. Moving from APF-team-team to RPF-team-team (step 2) revealed a positive team feedback effect. The net result of these two steps cancel each other out, so that the addition of team RPF (step 2) seems to offset the free-rider problem with team incentives (step 1).

Finally, we observe no average performance difference between APF and any RPF under *individual incentives.* The average performance in RPF-ind-ind (32.3) is statistically the same (MW: p=0.42, RT: p=0.91) as in APF-ind-ind (32.4), see Table 5. Moreover, the average performance in RPF-ind-team (32.1) is statistically the same (MW: p=0.58, RT: p=0.79) as in APF-ind-ind (32.4). Hence, the positive effect of *team* RPF applies only under team incentives, not under individual incentives.

Table 5: Individual Incentives and RPF

	Average Performance (SD)			Mann-Whitney z-Statistics (p-value)	
	APF-ind-ind (1)	RPF-ind-ind (2)	RPF-ind-team (3)	(1) vs (2)	(1) vs (3)
Stage 1	25.31 (4.90)	25.53 (5.91)	25.00 (4.93)	0.29 (0.771)	0.52 (0.606)
Stage 2	28.97 (5.32)	29.65 (6.11)	29.04 (5.27)	-0.17 (0.862)	-0.06 (0.954)
All stages	32.43 (6.06)	32.29 (6.95)	32.13 (5.84)	0.80 (0.423)	0.56 (0.576)
N	68	55	53	123	121

Notes: Mann-Whitney pairwise test. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Observation 4: Subjects under *individual incentives* perform equally well, independently of feedback.

3.2 Robustness of Results

Thus far, we base our results on comparing mean performances, not controlling for any other potentially important characteristics.[19] Reported in Table 6 are OLS and Random Effects GLS estimations, controlling for other factors such as age and gender.[20,21] APF-team-team is the baseline (ref.). We include a column for the 1st stage, the 2nd stage and a column of all stages.[22]

[19]In the appendix, Table A1-1, we check for randomization across treatments. Some minor differences exist, so controlling for such differences may prove important to the robustness of our findings.

[20]In the regressions, we use robust standard errors clustered on sessions. However, as the number of clusters may be too low, it could downward bias our standard errors. Therefore, we use a more conservative approach of only having (C-1) degrees of freedom when stating p-values, where C is the number of clusters.

[21]Alternatively, we could increase number of clusters by applying the second highest level of clusters. The requirement is independence across clusters. In team RPF treatments, the only interaction between subjects occur at the level of feedback, i.e. all members of all teams that interact. For the remaining treatments, there are either no interaction or only interaction between three subjects. To have a common level of clusters across all treatments, we constructed clusters that included all interacting subjects for these treatments (however, all subjects do not necessarily interact within this cluster). This approach only provided marginal differences to the results presented in the paper. The only part with notable differences is section 3.3, where significance levels drop to 5% level or 10% level. For this approach in the analysis of gender, the interaction between team RPF and team incentive no longer remain significant for males, and the other variables drop slightly in significance.

[22]The remaining stages are in the appendix, Table A1-3.

Table 6: Main Results: Treatment Effects on Productivity

Stage(s):	1st stage	2nd stage	All stages
	(1)	(2)	(3)
APF-team-team	Ref.	Ref.	Ref.
APF-ind-ind	3.149***	2.202***	2.529***
	(0.8345)	(0.2976)	(0.6027)
RPF-ind-ind	3.758**	2.887***	3.563***
	(1.3377)	(0.5186)	(0.8071)
RPF-ind-team	3.471***	2.733***	3.559***
	(1.1521)	(0.8012)	(0.7455)
RPF-team-team	2.618	2.720**	3.578**
	(1.8772)	(1.0981)	(1.2679)
RPF-team-ind	2.437*	1.510**	1.426**
	(1.3011)	(0.5784)	(0.6648)
$Stage_t$			2.366***
			(0.0757)
Constant	31.428***	35.059***	32.604***
	(2.9264)	(2.5666)	(2.6599)
Adjusted R^2	0.094	0.059	
Observations	350	350	2100

Notes: OLS coefficients reported in columns (1) – (2) and Random Effects GLS coefficients reported in column (3), with robust standard errors in parentheses, corrected for clustering across sessions. Dependent variable is number of solved tasks. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality.

$^{*}\ p < 0.10,\ ^{**}\ p < 0.05,\ ^{***}\ p < 0.01.$

First, in column (3), observe the significant effect of RPF-team-team relative to the baseline, strongly supporting our first observation that team RPF triggers motivation to exert higher effort. The effect is consistent throughout the experiment. Then consider individual RPF under team incentives. The coefficient estimate is positive and significant when we include controls to

our estimation, and the performance of subjects in this treatment is slightly higher than the baseline. But the effect weakens in the later stages, as can be seen in Table A1-3. Compare then the performance in RPF-team-team to the RPF-team-ind. Across all stages the RPF-team-team subjects outperform the RPF-team-ind subjects by about 2 tasks, but this difference is just about significant at the 10% level (p=0.101). However, if we only look at stages 2-6, allowing subjects to respond to the feedback, the difference between them is significant (p=0.036, see also Table 7 below).

Consider then the pure incentive effect, i.e. the difference between individual incentives (APF-ind-ind) and team incentives (APF-team-team). The treatment effect exists in all columns. This supports our third observation, even after controlling for other potentially important factors.

Finally, consider the effects of RPF under *individual incentives*. The difference between the coefficients of APF-ind-ind and RPF-ind-ind represents the effect of individual RPF, which is not significant (p=0.16). However, the difference between the coefficients of APF-ind-ind and RPF-ind-team is significant (p=0.07), suggesting that adding controls reveals slightly positive effect of team RPF on performance also under individual incentives.

In Table 7, we split the sample into two, concentrating on the performances of the first three stages and last three stages separately. We see that relationships between the performances across treatments are generally persistent, especially under individual incentives. Under team incentives, the effects of team RPF seem to persist and even increase in point estimates against the baseline over the final stages. It suggests that the effects of providing team RPF is robust over time. This in contrast to RPF-team-ind, where we clearly see a drop in the treatment effect over the final stages. It no longer remains significant and the improvement over the final stages is significantly less than the improvement of subjects in APF-team-team.[23] The performance of subjects with team RPF is also significantly higher (p=0.025) than the performance of subjects with individual RPF in the final three stages

[23]Based on a regression including all stages, interacting the final three stages with treatments, p<0.01.

(under team incentives). The improvement is also greater in RPF-team-team than RPF-team-ind over the final stages relative to the first stages ($p<0.01$).

Table 7: Persistence of Treatment Effects

Stages:	Stages 1-3	Stages 4-6	Stages 2-6	
	(1)	(2)	(3)	(4)
APF-team-team	Ref.	Ref.	Ref.	Ref.
APF-ind-ind	2.418***	2.641***	2.406***	-0.547
	(0.6598)	(0.5627)	(0.5582)	(0.4144)
RPF-ind-ind	3.668***	3.457***	3.398***	-0.720
	(0.9188)	(0.7410)	(0.7196)	(0.6516)
RPF-ind-team	3.176***	3.941***	3.509***	-0.062
	(0.7987)	(0.7189)	(0.6837)	(0.4891)
RPF-team-team	3.078**	4.078***	3.771***	1.320
	(1.3408)	(1.2122)	(1.1804)	(0.7782)
RPF-team-ind	1.804**	1.047	1.099*	-1.769***
	(0.7920)	(0.6119)	(0.5851)	(0.5924)
$Stage_t$	3.367***	1.856***	2.009***	2.009***
	(0.1323)	(0.1062)	(0.0777)	(0.0777)
Correct in 1st stage				0.938***
				(0.0559)
Observations	1050	1050	1750	1750

Notes: Random Effects GLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across sessions. Dependent variable is number of solved tasks. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality.

$^{*}\ p < 0.10,\ ^{**}\ p < 0.05,\ ^{***}\ p < 0.01.$

A slightly different way to study persistence is to focus explicitly on stages 2-6, and use the 1st stage performance as a control for initial effort. Column 3 in Table 7 shows the treatment effects in stages 2-6. Noteworthy is the rapid

drop in the treatment effect of RPF-team-ind, which no longer remains significantly greater than the baseline. The other relationships are qualitatively unchanged. Then, in column 4, we add the first stage performance as a control. It shows that subjects in the RPF-team-team (almost significantly) increase the performance gap to the baseline even more (p=0.108). In the RPF-team-ind, on the other hand, there is a significant performance drop relative to the baseline (p<0.01).

3.3 Interaction Effects

The RPF treatments fit into a 2 x 2 design, varying between individual incentives or team incentives and individual RPF or team RPF (see Table 1).[24] In order to study how team incentives and team RPF affect each other, we employ a regression with an interaction term between team incentives c and team RPF r. This gives the following model:

$$y_i = \alpha + \beta_1 c_i + \beta_2 r_i + \beta_3 c_i r_i + controls + \varepsilon_i,$$

where $c_i = 1$ if subject i is working under team incentives (i.e., RPF-team-team or RPF-team-ind), and 0 if subject i is paid individual incentives; $r_i = 1$ if subject i is provided with team RPF (i.e. RPF-ind-team or RPF-team-team), and 0 if subject i is provided with individual RPF. Controls are the same as indicated in Table 6. Then β_1 is the effect on performance (y_i) of team incentives without team RPF, β_2 is the effect of team RPF without team incentives, while β_3 estimates the interaction between them.

In Table 8 we can see that there is a strong negative effect of team incentives alone, whereas team RPF alone has no significant effect. However, we find a strong and positive interaction effect between the variables. This suggests that team feedback and team incentives are complements, i.e. providing team RPF positively strengthens the influence of team incentives, and vice versa.

[24]Note that the reference for comparison is not the same for subjects in the two different individual RPF treatments, as subjects in RPF-ind are compared to two other subjects in the session, whereas subjects in RPF-team-ind are compared to two other subjects within the same team.

The net effect of both team incentives and team RPF is slightly positive, although not significant.

Table 8: Changing Incentives and Feedback

Stage(s):	All stages (1)	Stages 1-3 (2)	Stages 4-6 (3)
Individual incentives and individual RPF	Ref.	Ref.	Ref.
Team incentives	-2.694***	-2.563***	-2.826***
	(0.6718)	(0.6420)	(0.7486)
Team RPF	-0.456	-1.022	0.109
	(0.5998)	(0.5914)	(0.6856)
Team incentives x Team RPF	3.533***	3.484***	3.583***
	(0.9364)	(0.9474)	(0.9909)
Stage $_t$	2.301***	3.300***	1.698***
	(0.0992)	(0.1719)	(0.1337)
Constant	32.958***	28.822***	38.110***
	(2.9674)	(2.4412)	(3.8004)
Observations	1350	675	675

Notes: Random Effects GLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across sessions. Dependent variable is number of solved tasks. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality.

$^{*}\ p < 0.10,\ ^{**}\ p < 0.05,\ ^{***}\ p < 0.01.$

Observation 5: Team incentives and team RPF are complements.

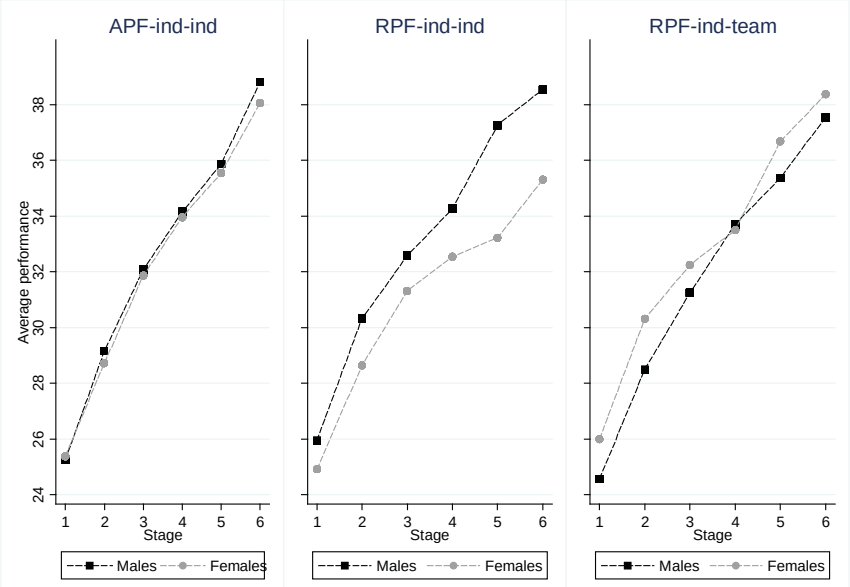
3.4 Gender Analysis

Previous research has shown that genders respond differently to competition (Niederle & Vesterlund, 2007) and that this also applies to feedback and

incentives (Azmat & Iriberri, 2016). In this section, we study gender effects in our experiment.

In Figure 2 we separately plot the performance of each gender for the three *individual incentive* treatments. In APF-ind-ind (left), there are virtually no gender difference. In the RPF-ind-ind (middle), males seem to perform better than females. And females actually perform slightly better than males in RPF-ind-team (right).[25] Finally, females in RPF-ind-team perform better than females in the other two individual incentive treatments. Males perform marginally better in RPF-ind-ind compared to males in APF-ind-ind.

Figure 2: Gender - Individual Incentive Treatments

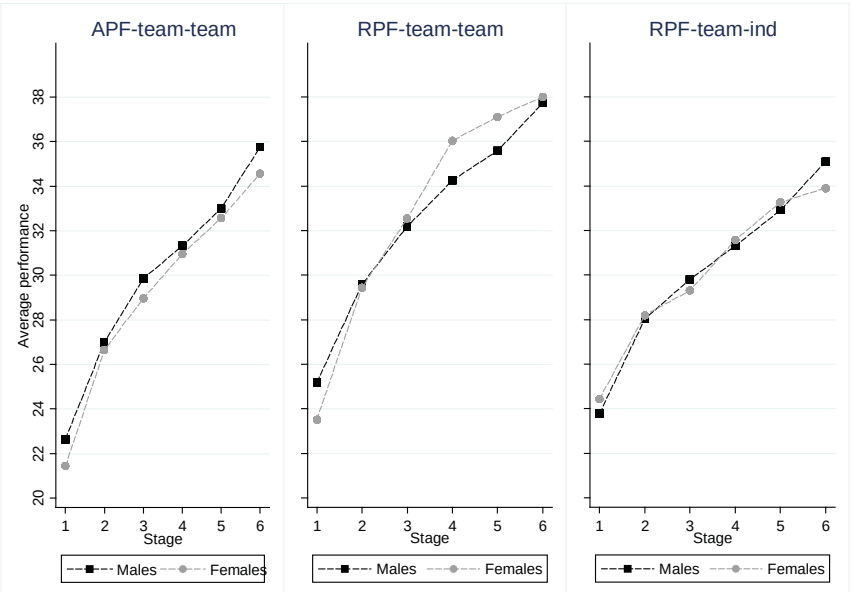


[25]This difference is not significant, based on a pooled regression version of Table 9 where treatments are interacted with gender.

Then we plot the gender specific performance for the *team incentive* treatments in Figure 3. In the APF-team-team (left), there is no visible differences in performance across gender. For the RPF-team-team (middle), females perform better than males, especially towards the end of the experiment. In the RPF-team-ind (right), there is no visible difference in performance across gender. Both genders seem to perform better in RPF-team-team relative to the other two treatments.

Notice that females perform better than males in two out of six treatments; both of them providing team feedback. When the feedback generates competition between teams, females perform better than females without this feedback, and better than males within the same treatment. Whenever the feedback generates individual competition, females respond negatively. Further analysis on this is required.

Figure 3: Gender – Team Incentive Treatments



In Table 9, we run regressions on gender. Consider males in column (1). Males do better in all treatments relative to males in the baseline. Under individual incentives, males in RPF-ind-ind outperform males in both APF-ind-ind ($p<0.01$) and RPF-ind-team ($p<0.01$). Hence, individual feedback triggers males. In column (2), females only perform better in APF-ind-ind and RPF-ind-team relative to females in the baseline. Considering individual incentives separately, females in APF-ind-ind and RPF-ind-team do better than females in RPF-ind-ind ($p<0.05$ for both), suggesting that females dislike individual RPF.

Consider then gender differences. Under individual incentives, there are no differences in the performance across gender, except that males outperform females provided with individual RPF ($p<0.001$).[26] There are no gender differences under team incentives. In columns (3) – (4), we study gender specific treatment effects in stages 2-6 only, controlling for stage 1 performance. The most notable result from these columns is the significant drop in performance of females that receive individual RPF. This suggests that females dislike such attention to individual performance. No such drop is evident in team RPF treatments.

[26]This is based on a pooled regression with treatments interacted with gender.

Table 9: Gender Analysis

Panel:	All stages		Stages 2-6	
	Males	Females	Males	Females
	(1)	(2)	(3)	(4)
APF-team-team	Ref.	Ref.	Ref.	Ref.
APF-ind-ind	1.795*	3.712***	-0.806	-0.096
	(1.0387)	(1.1628)	(0.7278)	(0.6374)
RPF-ind-ind	5.225***	0.647	0.447	-3.118***
	(0.7819)	(1.4760)	(0.9815)	(0.9262)
RPF-ind-team	2.805***	5.228**	0.018	-0.296
	(0.7744)	(2.2997)	(0.6276)	(0.7189)
RPF-team-team	4.249***	1.977	0.793	0.633
	(1.2168)	(1.7549)	(0.6221)	(1.3062)
RPF-team-ind	1.501*	0.873	-0.884	-3.382***
	(0.7733)	(1.3504)	(1.1121)	(0.6702)
Stage $_t$	2.409***	2.306***	1.011***	0.878***
	(0.0890)	(0.1258)	(0.0564)	(0.0711)
Correct 1st stage			3.826	6.818**
			(3.4298)	(2.9221)
Observations	1218	882	1015	735

Notes: Random Effects GLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across sessions. The dependent variable in columns (1) – (2) is number of solved tasks in all stages, whereas in columns (3) – (4) it is number of solved tasks in stages 2-6 only. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for economics students and a dummy for Norwegian nationality.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In sum, these observations support previous findings (as in e.g. Azmat and Iriberri (2016)):

Observation 6: Males respond positively to individual RPF, while females respond negatively. Both genders respond positively to team RPF. Individual RPF makes females produce less.

Consider now the interaction effects between feedback and incentives. In Table 10 we employ the same analysis as in Section 3.3 (Table 8), but on each gender separately. First, observe that males respond more negatively to team incentives alone than females. Second, males respond negatively to team RPF alone, whereas females respond positively. Hence, while males are triggered by individual RPF, females are triggered by team RPF. Finally, we observe that the positive interaction effect demonstrated in Table 8 is gender specific. For males there is a strong complementarity between team incentives and team RPF, although the net effect of shifting both factors is insignificant. Females, on the other hand, only need team RPF to improve performance, and do not gain additional productivity when interacting the two variables. Their net differential performance of changing to both team incentives and team RPF (the sum of all coefficients) is positive ($p=0.047$).

Table 10: Changing Incentives and Feedback – Gender Analysis

Panel:	Males (1)	Females (2)
Individual incentives and individual RPF	Ref.	Ref.
Team incentives	-3.600***	-1.326*
	(1.1046)	(0.7177)
Team RPF	-2.652***	3.956**
	(0.8195)	(1.2812)
Team incentives X Team RPF	4.312***	0.124
	(1.2671)	(1.5096)
Stage $_t$	2.355***	2.227***
	(0.1109)	(0.1869)
Constant	37.294***	30.263***
	(4.7561)	(3.1577)
Observations	780	570

Notes: Random Effects GLS coefficients reported, with robust standard errors in parentheses, corrected for clustering across sessions. Dependent variable is number of solved tasks across all stages. Both columns include the following control variables: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for economics students and a dummy for Norwegian nationality.

$^{*}\ p < 0.10,\ ^{**}\ p < 0.05,\ ^{***}\ p < 0.01.$

Observation 7: Females respond positively to team RPF, independently of incentives. Males respond negatively to both team incentives and team RPF alone, but a strong positive complementary effect between the two offsets the negative effects.

3.5. Heterogeneous Effects

We have seen that RPF affects average performances. In this section, we investigate heterogeneous effects, i.e. to what extent the treatments affect the performance distributions. We are particularly interested in comparing the top performances between treatments.

We categorize subjects within a team as either best or worst, based on their average performance over all stages. Hence, a performer categorized as best keeps this categorization in all rounds, even though some in the team may have done better in a single stage. We compare the difference between the performance of the best and the worst subject within a team, and compare this difference across treatments. Figure 4 shows a substantially larger gap between the best and the worst performers within a team in the RPF-team-team, compared to any other treatment. Notice that we have also included the RPF-ind-ind for comparison, and constructed these "teams" based on the same subjects as their comparison group of two other subjects. Who drives the difference that we see in Figure 4? This is illustrated in Figure 5; high performers in RPF-team-team perform substantially better than high performers in any other treatment, whereas there are no differences for the lowest performers.

Figure 4: Difference between High and Low Performers across Treatments

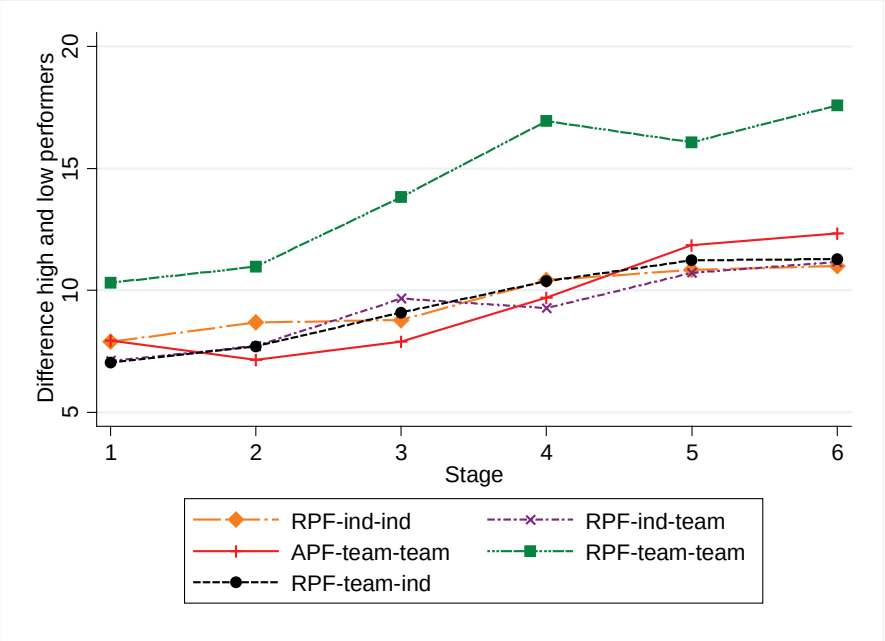
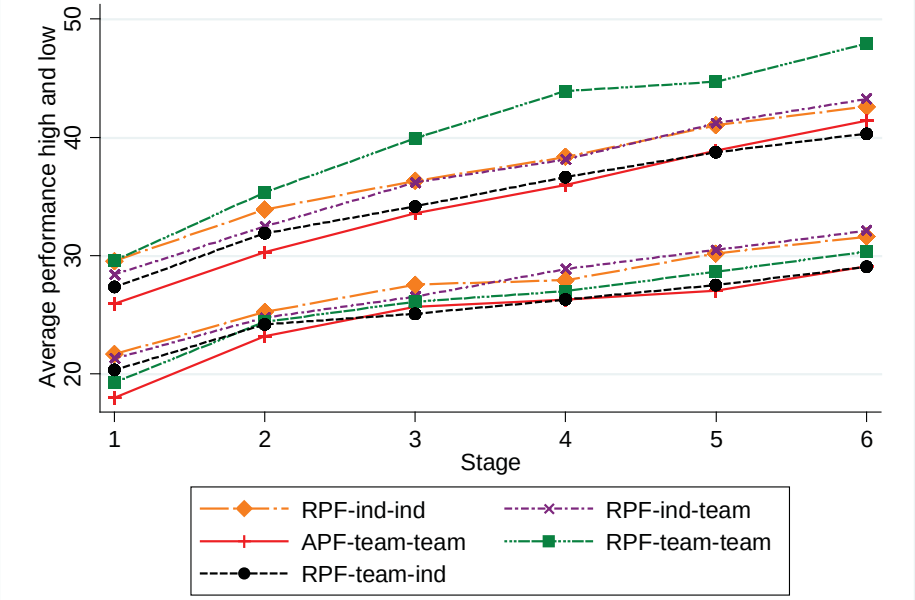


Figure 5: High and Low Performers across Treatments



Furthermore, in Table 11 we report quantile regressions across the distribution of subjects' average number of correctly solved tasks across all stages. Notice the quantile effects in both team RPF treatments. For RPF-ind-team and RPF-team-team, there is a particularly strong positive effect for those in the upper part of the performance distribution. High performers in these treatments perform better than high performers in other treatments, and particularly against the baseline APF-team-team.[27] Although this analysis does not take into account which team the subject was in, it is consistent with the findings of Figure 5, in that high performers in RPF-team-team do better than high performers in other treatments. In comparison, there seem to be no quantile effects of RPF-team-ind; subjects in this treatment perform equally well as their comparison subjects at different quantiles in the APF-team-team.

[27]Specifically, at the 90% quantile, RPF-ind-team is also significantly greater than APF-ind-ind.

Table 11: Marginal Treatment Effects across Quantiles

Quantile:	10%	25%	50%	75%	90%
	(1)	(2)	(3)	(4)	(5)
APF-team-team	Ref.	Ref.	Ref.	Ref.	Ref.
APF-ind-ind	1.396	2.410	2.369	2.711*	1.697
	(1.7573)	(1.4706)	(1.7082)	(1.3899)	(1.9684)
RPF-ind-ind	1.979	2.462	2.482	3.023	4.652*
	(2.1938)	(1.8981)	(1.5015)	(2.0786)	(2.4430)
RPF-ind-team	0.854	2.697	3.321*	4.377**	5.955**
	(1.7065)	(1.6392)	(1.7520)	(1.9019)	(2.2914)
RPF-team-team	-0.417	0.492	1.673	5.545**	7.545***
	(1.7406)	(1.6561)	(1.9975)	(1.9592)	(2.5874)
RPF-team-ind	0.396	0.977	1.470	1.051	2.106
	(2.3174)	(1.8932)	(1.6135)	(1.8456)	(2.2698)
Constant	32.896***	37.541***	36.911***	44.503***	49.333***
	(3.8628)	(4.4789)	(3.7987)	(4.3926)	(3.9050)
Observations	350	350	350	350	350

Notes: Quantile regression coefficients reported, with robust standard errors in parentheses, based on bootstrapping with 1.000 replications. Dependent variable is the average number of solved tasks across all stages. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Observation 8: High performing subjects perform better in treatments with team RPF than in any other treatment.

In Table 12, we present regressions including a dummy variable (BiT) for the subject who is the best performer within the team. This variable is also interacted with each treatment. Hence, the sum of the coefficients BiT and the treatment interacted with BiT, is the additional tasks the best performer solved relative to the other two subjects within the team. Therefore, to compare best performers within a team across treatments, say between best performers in RPF-ind-ind and APF-team-team, one has to take the difference between them. That is, for the concrete example, one has to sum the coefficients for RPF-ind-ind and RPF-ind-ind x BiT in order to find the corresponding estimated difference.[28]

Consistent with Figure 5, best performers in RPF-team-team perform significantly better than best performers in the baseline. Moreover, the best performers in RPF-team-team also perform significantly better than the best performers in both RPF-ind-team (p=0.03) and RPF-team-ind (p=0.01).[29] Notably, in Table 11 high performers seemed to perform better in both team RPF treatments. Table 12 suggests that for RPF-team-team this is driven by the best performer within each team, whereas in RPF-ind-team it must be that there are more often multiple high performers in the same team (as there is no interaction effect between BiT and RPF-ind-team).

Observation 9: The best performer within a team performs better in RPF-team-team compared to any other treatment.

[28]Similarly, to compare the best performer in RPF-team-team to RPF-team-ind, the difference between them is the sum of the coefficients (RPF-team-team + RPF-team-team x BiT) – (RPF-team-ind + RPF-team-ind x BiT), i.e. (2.05+4.30) – (1.45-0.10) = 5.

[29]Also in point estimates against the best performers in RPF-ind-ind (p=0.14).

Observation 9 implies that team incentives trigger significantly higher top performance than individual incentives, when subjects are exposed to team RPF. But note that the second and third performers within the team perform significantly worse under team incentives than under individual incentives. This can be seen directly from the coefficients to RPF-ind-ind and RPF-ind-team when compared against the baseline, but the difference is also significant for those in RPF-team-ind relative to the individual incentive treatments (column (3), both $p<0.01$).

Table 12: Best Performers across Treatments

Stages:	1st stage (1)	2nd stage (2)	All stages (3)
APF-team-team	Ref.	Ref.	Ref.
RPF-ind-ind	3.501**	2.222***	3.540***
	(1.5584)	(0.6901)	(0.7625)
RPF-ind-team	3.466**	2.525**	3.647***
	(1.5296)	(1.0740)	(0.8219)
RPF-team-team	1.875	1.458	2.052
	(1.9212)	(0.8831)	(1.2320)
RPF-team-ind	2.610	1.229	1.453*
	(1.5699)	(0.7455)	(0.6980)
BiT (Best in Team)	5.481***	5.047***	6.864***
	(1.2606)	(1.2029)	(0.8641)
RPF-ind-ind x BiT	0.416	1.356	0.007
	(1.7578)	(1.7810)	(1.5093)
RPF-ind-team x BiT	-0.322	0.045	-0.447
	(1.5786)	(1.2655)	(1.0078)
RPF-team-team x BiT	2.192	3.577*	4.297***
	(1.7282)	(1.6955)	(1.2533)
RPF-team-ind x BiT	-0.849	0.357	-0.104
	(1.4659)	(1.4605)	(1.0329)
Stage $_t$			2.328***
			(0.0835)
Constant	25.553***	29.417***	25.651***
	(2.6936)	(1.9901)	(2.2125)
Adjusted R^2	0.295	0.321	
Observations	282	282	1692

Notes: OLS coefficients reported in columns (1) – (2) and Random Effects GLS coefficients reported in column (3), with robust standard errors in parentheses, corrected for clustering across sessions. Dependent variable is number of solved tasks. BiT is a dummy variable taking value 1 if the subject is the best performer in his or her team, 0 otherwise. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4. Discussion and Conclusion

In this paper we investigate experimentally how teams respond to relative performance feedback (RPF). We find that when subjects are exposed to team incentives, then RPF on how their team is doing compared to two other teams increases the team's average performance by almost 10 percent. The treatment effect is driven by the teams' top performers. The average individual performance of the top performers within each team is almost 20 % higher when the teams receive relative performance feedback compared to when the teams only receive absolute performance feedback. Our experiment suggests that subjects, and in particular top performers, are motivated by the combination of team incentives and team RPF. In fact, team incentives trigger significantly higher top performance than individual incentives, when subjects are exposed to team RPF.

We also find some interesting gender effects. Females respond negatively to individual RPF, but even more positively than males to team RPF. For males, team incentives have a strong negative effect compared to individual incentives, unless it is accompanied by team RPF. For females, the incentives do not matter to the same degree, and team RPF has a strong positive effect regardless of the incentive system.

Although we provide evidence on the performance enhancing effects of team RPF and its interaction with team incentives, our design does not enable us to identify the exact mechanism behind the performance improvement. In fact, the heterogeneous response to team RPF that we find indicates that team RPF does not provide any general peer pressure or identity effects. Rather, it seems to provide a strong motivation for the high ability subjects, but not so for those with lower ability. This result complements the interesting and somehow puzzling findings by Hamilton et al. (2003), namely that high ability workers were more attracted to team work than low ability workers. When offering workers at a garment plant the opportunity to shift from individual piece rates to team incentives, the high-productivity workers tended to join teams first, despite a loss in earnings for many of them. Hamilton et al suggested that high-ability workers may acquire a higher social status in teams and are therefore willing to join teams even if their own

pay is reduced. Our results illuminate Hamilton et al's findings, which suggest that high-ability workers are not motivated by team incentives alone. Rather, they seem to be motivated by the chance to help the team achieve some non-monetary goals, which in our experiment is higher ranking.

For managers designing feedback interventions in their organization, there are several implications of this experiment. We find that competition between teams for higher ranks may be an efficient way to improve the productivity of employees, in particular if they are paid as a team. The competition should be between teams rather than competition within teams. Finally, with respect to gender differences, females seem to be particularly productive when they work as a team and are provided with team performance data rather than individual performance data.

References

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *Quarterly journal of Economics, 115*(3), 715-753.

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the economics of organizations. *The Journal of Economic Perspectives, 19*(1), 9-32.

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review, 62*(5), 777-795.

Azmat, G., & Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy, 25*(1), 77-110.

Baer, M., Leenders, R. T. A., Oldham, G. R., & Vadera, A. K. (2010). Win or lose the battle for creativity: The power and perils of intergroup competition. *Academy of Management Journal, 53*(4), 827-845.

Bandiera, O., Barankay, I., & Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association, 11*(5), 1079-1114.

Bellemare, C., Lepage, P., & Shearer, B. (2010). Peer pressure, incentives, and gender: An experimental analysis of motivation in the workplace. *Labour Economics, 17*(1), 276-283.

Birkinshaw, J. (2001). Why is knowledge management so difficult? *Business Strategy Review, 12*(1), 11-18.

Blanes i Vidal, J., & Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science, 57*(10), 1721-1736.

Boning, B., Ichniowski, C., & Shaw, K. (2007). Opportunity counts: Teams and the effectiveness of production incentives. *Journal of Labor Economics, 25*(4), 613-650.

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology, 50*(3), 543-549.

Burton-Chellew, M. N., & West, S. A. (2012). Pseudocompetition among groups increases human cooperation in a public-goods game. *Animal Behaviour, 84*(4), 947-952.

Böhm, R., & Rockenbach, B. (2013). The inter-group comparison–intra-group cooperation hypothesis: Comparisons between groups increase efficiency in public goods provision. *PloS one*, *8*(2), e56152.

Charness, G., Masclet, D., & Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, *60*(1), 38-55.

Che, Y.-K., & Seung-Weon, Y. (2001). Optimal incentives for teams. *The American Economic Review*, *91*(3), 525-541.

Erev, I., Bornstein, G., & Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, *29*(6), 463-478.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679-688.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171-178.

Gjedrem, W. G. (2015). *Relative performance feedback: Effective or dismaying?* Working paper. UiS Business School. University of Stavanger, Norway.

Guryan, J., Kroft, K., & Notowidigdo, M. J. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, *1*(4), 34-68.

Hamilton, Barton H., Nickerson, Jack A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, *111*(3), 465-497.

Hannan, R. L., Krishnan, R., & Newman, A. H. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review*, *83*(4), 893-913.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 324-340.

Holmström, B., & Milgrom, P. (1990). Regulating trade among agents. *Journal of Institutional and Theoretical Economics*, 85-105.

Itoh, H. (1991). Incentives to Help in Multi-Agent Situations. *Econometrica*, *59*(3), 611-636.

Itoh, H. (1992). Cooperation in Hierarchical Organizations: An Incentive Perspective. *Journal of Law, Economics, & Organization, 8*(2), 321-345.

Kandel, E., & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy, 100*(4), 801-817.

Knez, M., & Simester, D. (2001). Firm-wide incentives and mutual monitoring at Continental Airlines. *Journal of Labor Economics, 19*(4), 743-772.

Kuhnen, C. M., & Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science, 58*(1), 94-113.

Kvaløy, O., & Olsen, T. E. (2006). Team incentives in relational employment contracts. *Journal of Labor Economics, 24*(1), 139-169.

Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy, 89*(5), 841-864.

Lazear, E. P., & Shaw, K. L. (2007). Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives, 21*(4), 91-114.

Macho-Stadler, I., & Pérez-Castrillo, J. D. (1993). Moral hazard with several agents: The gains from cooperation. *International Journal of Industrial Organization, 11*(1), 73-100.

Marianne, B. (2011). Chapter 17 - New perspectives on gender. In D. Card & O. Ashenfelter (Eds.), (Vol. 4, Part B, pp. 1543 - 1590): Elsevier.

Marino, A. M., & Zábojnik, J. (2004). Internal Competition for Corporate Resources and Incentives in Teams. *The RAND Journal of Economics, 35*(4), 710-727.

Nalbantian, H. R., & Schotter, A. (1997). Productivity under group incentives: An experimental study. *The American Economic Review, 87*(3), 314-341.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics, 122*(3), 1067-1101.

Sausgruber, R. (2009). A note on peer effects between teams. *Experimental Economics, 12*(2), 193-201.

Tan, J. H., & Bolle, F. (2007). Team competition and the public goods game. *Economics Letters, 96*(1), 133-139.

Turner, J. C. (1975). Social comparison and social identity: Some prospects for intergroup behaviour. *European Journal of Social Psychology, 5*(1), 1-34.

van Dijk, F., Sonnemans, J., & van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review, 45*(2), 187-214.

Vandegrift, D., & Yavas, A. (2011). An experimental test of behavior under team production. *Managerial and Decision Economics, 32*(1), 35-51.

Appendix

A1 Tables

Table A1-1: Summary Statistics of Control Variables

	APF-ind-ind	RPF-ind-ind	RPF-ind-team	APF-team-team	RPF-team-team	RPF-team-ind	Pearson^2 / Kruskal Wallis
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Econ students	0.132	0.0364	0.0377	0.123	0.204	0.143	0.044[30]
	(0.341)	(0.189)	(0.192)	(0.331)	(0.407)	(0.353)	
Norwegian-Nationality	0.706	0.473	0.434	0.579	0.519	0.413	0.010[31]
	(0.459)	(0.504)	(0.500)	(0.498)	(0.504)	(0.496)	
Age	24.29	26.05	26.08	24.25	25.57	25.35	0.015
	(4.316)	(3.955)	(4.751)	(3.291)	(4.364)	(4.656)	
Female	0.426	0.400	0.302	0.404	0.500	0.476	0.365
	(0.498)	(0.494)	(0.463)	(0.495)	(0.505)	(0.503)	
Average grade	2.559	2.055	2.340	2.526	2.370	2.508	0.003[32]
	(0.720)	(0.780)	(0.678)	(0.782)	(0.623)	(0.592)	
Observations	68	55	53	57	54	63	350

Notes: Mean and (standard deviation). In column (7) we report p-value of Pearson^2 for binary variables and Kruskal Wallis for non-binary variables.

[30]Excluding RPF-ind leads these differences to be insignificant (p=0.150)
[31]Excluding APF-ind leads these differences to be insignificant (p=0.385)
[32]Excluding RPF-ind leads these differences to be insignificant (p=0.352)

Table A1-2: Team RPF and Free-Riding

	Average Performance (SD)			Mann-Whitney z-Statistics (p-value)	
	APF-ind-ind	RPF-ind-ind	RPF-team-team		
	(1)	(2)	(3)	(1) vs (3)	(2) vs (3)
Stage 1	25.31 (4.90)	25.53 (5.91)	24.35 (6.81)	1.20 (0.230)	0.87 (0.382)
Stage 2	28.97 (5.32)	29.65 (6.11)	29.52 (6.23)	0.01 (0.994)	0.20 (0.841)
All stages	32.43 (6.06)	32.29 (6.95)	32.60 (7.63)	0.46 (0.648)	-0.06 (0.954)
N	68	55	54	122	109

Notes: Mann-Whitney pairwise test. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A1-3: Treatment Effects across Stages

Stages:	3rd stage	4th stage	5th stage	6th stage
	(1)	(2)	(3)	(4)
APF-team-team	Ref.	Ref.	Ref.	Ref.
APF-ind-ind	2.557***	3.070***	2.980***	3.378***
	(0.5777)	(0.6835)	(0.6538)	(0.9001)
RPF-ind-ind	2.726***	2.276***	2.783***	1.778**
	(0.6547)	(0.5267)	(0.7393)	(0.6832)
RPF-ind-team	2.853***	3.539***	3.977***	3.430**
	(0.6817)	(1.1322)	(1.1992)	(1.3356)
RPF-team-team	3.022**	3.882***	3.625***	2.563
	(1.2275)	(1.1901)	(1.1188)	(1.5192)
RPF-team-ind	0.473	0.629	0.780	-0.373
	(0.8411)	(0.6798)	(0.9434)	(1.0629)
Constant	40.523***	45.213***	46.721***	49.436***
	(2.6384)	(2.7256)	(3.1974)	(3.6538)
Adjusted R^2	0.095	0.123	0.109	0.076
Observations	350	350	350	350

Notes: OLS coefficients reported, with robust standard errors in parenthesis, corrected for clustering across sessions. Dependent variable is number of solved tasks. All columns have the following control variables included: Time on the day of the session (FE in panel), age, average grades at University level, a dummy for gender, a dummy for economics students and a dummy for Norwegian nationality. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A2 Experimental Instructions

Welcome to the experiment (APF-ind-ind)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition, you will earn 1 NOK for each task you solve. In other words, your payment depends on how many tasks you solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and refrain from communicating with other participants. You may use your

mobile phone to surf the internet, but please ensure that it is in a mute state before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

***

Welcome to the experiment (RPF-ind-ind)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition, you will earn 1 NOK for each task you solve. In other words, your payment depends on how many tasks you solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

In addition your performance will be ranked relative to two other randomly selected participants in the room, and you will be informed about how many tasks they have solved. You will be ranked relative to the same participants in all of the breaks. Ranks will not affect your payment.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and refrain from communicating with other participants. You may use your mobile phone to surf the internet, but please ensure that it is in a mute state before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

***

Welcome to the experiment (RPF-ind-team)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Team:
You are part of a team consisting of a total of three randomly selected participants in the room, and you will all be working simultaneously on the same type of tasks. The team will remain unchanged throughout the experiment.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition, you will earn 1 NOK for each task you solve. In other words, your payment depends on how many tasks you solve. Your payment does not depend on how many tasks the other team members solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

In addition you will also be informed about the total output of your team in the previous stage. Also, your team performance will be ranked relative to two other teams in the room, and you will be informed about how many tasks these teams have solved. Your team will be ranked relative to the same teams in all of the breaks. Ranks will not affect your payment.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and refrain from communicating with other participants. You may use your mobile phone to surf the internet, but please ensure that it is in a mute state before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

***

Welcome to the experiment (APF-team-team)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Team:
You are part of a team consisting of a total of three randomly selected participants in the room, and you will all be working simultaneously on the same type of tasks. The team will remain unchanged throughout the experiment.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition, your team will earn 1 NOK for each task a team member solves. The total earnings of the team is then divided equally among each team member independently of actual contribution. In other words, your payment depends on how many tasks you and your other team members solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

In addition you will also be informed about the total output of your team in the previous stage.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and

refrain from communicating with other participants. You may use your mobile phone to surf the internet, but please ensure that it is in a mute state before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

***

Welcome to the experiment (RPF-team-team)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Team:
You are part of a team consisting of a total of three randomly selected participants in the room, and you will all be working simultaneously on the same type of tasks. The team will remain unchanged throughout the experiment.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition,

your team will earn 1 NOK for each task a team member solves. The total earnings of the team is then divided equally among each team member independently of actual contribution. In other words, your payment depends on how many tasks you and your other team members solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

In addition you will also be informed about the total output of your team in the previous stage. Also, your team performance will be ranked relative to two other teams in the room, and you will be informed about how many tasks these teams have solved. Your team will be ranked relative to the same teams in all of the breaks. Ranks will not affect your payment.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and refrain from communicating with other participants. You may use your mobile phone to surf the internet, but please ensure that it is in a mute state before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

***

Welcome to the experiment (RPF-team-ind)

Task description:
We ask you to decode letters into numbers. You are given a list of letters, all of which have been assigned with a corresponding number. Your task is then to decode given sequences of four letters into numbers.

Example: Given this list of letters

A	B	C	D	E	F	G
8	12	14	10	9	6	24

Task-
Decode these letters: **A | E | G | F**
Correct answer: **8 | 9 | 24 | 6**

Stages and process of the experiment:
The experiment consists of six working stages, and the duration of each stage is five minutes. There are unlimited number of tasks in each stage. A countdown in the upper right corner of the computer screen display remaining time of current stage. After the final stage, we will ask you to fill out a short questionnaire. Total duration of the experiment is estimated to be about 45 minutes.

Team:
You are part of a team consisting of a total of three randomly selected participants in the room, and you will all be working simultaneously on the same type of tasks. The team will remain unchanged throughout the experiment.

Payment:
Everyone earns 100 NOK for participating in the experiment. In addition, your team will earn 1 NOK for each task a team member solves. The total earnings of the team is then divided equally among each team member independently of actual contribution. In other words, your payment depends on how many tasks you and your other team members solve.

Breaks:
In between each stage there will be a minute break. During the breaks you will be provided with information about how many tasks you have correctly solved and how much you have earned during the previous stage.

In addition you will also be informed about the total output of your team in the previous stage. Also, your contribution to the team performance will be ranked relative to the other two team members, and you will be informed about how many tasks each team member have solved. Ranks will not affect your payment.

Rules:
You choose freely how to spend your time during the experiment. However, we do require you to remain in your seat throughout the experiment, and refrain from communicating with other participants. You may use your mobile phone to surf the internet, but please ensure that it is in a mute state

before we start. It is strictly prohibited to use the pc to anything other than the experiment, as different usage may cause technical problems with the experiment.

Thank you for participating in the experiment.

Feedback and Risk-Taking with Own and Other People's Money*

Kristoffer W. Eriksen[1],William Gilje Gjedrem and Jon Kristian Heimdal[2]

Abstract: We investigate how manipulating feedback frequency on investment outcomes affects risk-taking in an investment game, when subjects make investment decisions for both themselves and others. We use the standard investment game of Gneezy and Potters (1997), and apply a within-between experimental design. Subjects invest both their own money and other people's money (within), while the frequency of feedback varies between subjects. Our main result shows that feedback frequency affects relative investment that subjects make for themselves and others: When feedback frequency is high, subjects invest on average the same amount in the risky lottery for both themselves and others. However, when feedback frequency is low, subjects invest significantly less in the risky lottery on behalf of others compared to what they do for themselves.

*For helpful comments and discussion, we would like to thank Ola Kvaløy and Mari Rege. Financial support from Stiftelsen for Anvendt Finans (SAFI) and UiS Business School is gratefully acknowledged.
[1]University of Stavanger, 4036 Stavanger, Norway. E-mail: kristoffer.w.eriksen@uis.no.
[2]4306 Sandnes, Norway. Email: jkheimdal@yahoo.no.

1. Introduction

People often take risk on behalf of others. For example, politicians decide on behalf of the local or national population and CEOs make decisions associated with risk taking on behalf of employees and owners. In finance, investment managers trade on behalf of their customers. In 2015, U.S. registered investment companies managed assets for more than $18 trillion, and this was on behalf of more than 90 million retail investors (*Investment Company Fact Book*, 2016). Their clients' willingness to take risk is often unknown or uncertain to the investment manager, and he may also choose different investment portfolios on behalf of others than what he does with his own wealth.[1] Furthermore, their interests in the outcome of the investments do not necessarily align, as investment managers often bear limited direct consequence of the investment outcomes.[2]

Even though investments on behalf of others is extensive, research offers only limited guidance to how people choose to make such investments and is particularly scant on how feedback on investment outcomes affects these decisions. The frequency of feedback on investment outcomes has previously been shown to affect investment decisions with own money (see e.g., Gneezy & Potters, 1997), and frequent feedback is natural for investment managers who closely monitor portfolios.

People who invest and take risk with their own money are affected by the frequency of feedback on investment outcomes. Benartzi and Thaler (1995) introduced the behavioral hypothesis termed myopic loss aversion (MLA) as a possible explanation to the famous equity premium puzzle (Mehra & Prescott, 1985).[3] It suggests that investors move towards less risky investments the more

[1]For example, he may have expectations about his clients' preferences (towards risk, investment ethics, investment horizon, etc.) and therefore adjust the portfolio accordingly.
[2]Investment managers are guided by incentive schemes that typically involve a fee calculated as a fraction of asset under management, in addition to a component related to excess performance over some benchmark (Heinkel & Stoughton, 1994).
[3]The equity premium puzzle refers to the implausible high risk aversion needed to explain the magnitude of the difference in return between equity and the risk free alternative.

frequently they receive and evaluate feedback on investment outcomes. While the experimental literature over the last 20 years has shown that people respond to such feedback manipulation when investing own money (starting with Gneezy and Potters (1997); Thaler, Tversky, Kahneman, and Schwartz (1997)), private investors often delegate wealth management to investment managers. Such professionals are also found to exhibit behavior consistent with MLA in experimental settings using their own money (Haigh and List (2005); Eriksen and Kvaløy (2010a)),[4] however less is known about how and whether the bias transfers to investment decisions on behalf of others.

In this paper, we investigate whether feedback frequency affects decision making when people make investment decisions both for themselves and for others. We make use of the standard investment game first introduced by Gneezy and Potters (1997), and employ a within-between subjects design. That is, while we vary the feedback frequency between subjects (high and low frequency), the same subjects make risky decisions with both own money and other people's money.

The within-subject part of the experiment allows us to shed some light on how people adapt their investment decision when facing a situation where they make a choice both with own money and another person's money, and to what extent the manipulation of feedback frequency affects this adaption. The between part of the experiment allows us to study whether subjects exhibit MLA with own and other people's money, and the within part allows us to study how much risk they take with own and other people's money (within). Combining these dimensions, we may also study the relative investment of subjects, i.e. how much they choose to invest with own money relative to how much they choose to invest with other people's money, and whether the manipulation of feedback frequency affects this.

Our results show that when people invest on behalf of others, feedback frequency on investment outcomes matter. The amount they invest is the same across low and high feedback frequency. However, the relative investment is different across feedback frequency. When the frequency is low, subjects invest significantly less with other people's money compared to their own money.

[4]Also supporting external validity of MLA, as most other studies are conducted in the laboratory with students.

When feedback frequency is high, they invest about the same amount with own money as with other people's money. In general, people do seem to exhibit MLA when they invest own money, but not when they invest other people's money. Thus, manipulating feedback frequency does not seem to make people less afraid of risk when they invest other people's money, and therefore average risk-taking is less than with own money. Consequently, in terms of maximizing expected earnings, people who make investment choices on behalf of others may fail to perform any better than what their clients' would have done themselves.

Background and related literature

The experiment is in the intersection between two strands of literature, as it relates both to the literature on MLA and the recent literature on risk-taking on behalf of others.

The MLA hypothesis draws on prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) and mental accounting (Kahneman & Tversky, 1984).[5] A consequence of MLA (on investor behavior) is that the evaluation period is crucial to the attractiveness of risky investments in the stock market. Frequent updates may make the stock market less attractive, as disutility from short-term fluctuations is consumed frequently. Less frequent updates, however, reduces the probability of observing losses, making risky options more attractive. Hence, according to the MLA hypothesis, people who are loss averse and frequently receive updates will be less willing to make risky investments. Besides the early papers on MLA, the hypothesis is thoroughly investigated in recent studies (see Thaler (2005); Bellemare, Krause, Kröger, and Zhang (2005); Sutter (2007); Langer and Weber (2008); Fellner and Sutter (2009); Zeisberger, Langer, and Weber (2014)).

Recently, a small experimental literature on risk-taking on behalf of others has emerged. Results are mixed and they seem to be sensitive to both the experimental task, incentive structure and context. To summarize; Chakravarty,

[5]Prospect theory advocates that the disutility from experiencing a loss is disproportionally large compared to the utility from an equally large gain. Mental accounting refers to how people evaluate and organize economic outcomes, and for the case of investments, the tendency people have to evaluate investments frequently and independently.

Harrison, Haruvy, and Rutström (2011), Polman (2012), Agranov, Bisin, and Schotter (2014) and Andersson, Holm, Tyran, and Wengström (2016) all find that subjects take higher risk with other people's money. Charness and Jackson (2009), Bolton and Ockenfels (2010), Eriksen and Kvaløy (2010b), Reynolds, Joseph, and Sherwood (2011), Füllbrunn and Luhan (2015) and Eriksen, Kvaløy, and Luzuriaga (2015) obtain the opposite result. Using the same investment game as we do, Füllbrunn and Luhan (2015) find people to be more risk averse when investing on behalf of others in an experiment where incentives for the decision-maker is either fixed or perfectly aligned with those bearing the consequences. In contrast, Pollmann, Potters, and Trautmann (2014) report lower risk aversion with OPM, which is moderated by accountability. However, none of these study the effects of feedback frequency on decision making for others.

Our study serves as a robustness check to the findings of Eriksen and Kvaløy (2010b), where they concluded that subjects investing on behalf of other people also exhibit MLA. An important difference in our study is that subjects may adapt their investment decision to the fact that they make (or have made) the same investment decision with their own money.[6] This should make subjects spend more time to reflect on the investment choice at hand, and it may change how they perceive the preferences of the other person.[7] Hence, our design tests whether subjects exhibit MLA when investing on behalf of others, when they face a situation where they have to invest similarly using their own money. Another design difference is that we neutrally tell subjects that they are investing for both

[6]In Eriksen and Kvaløy (2010b) subjects only made the decision on behalf of others.
[7]For example, as they have to choose an investment amount both for themselves and the other person, they may base the investment choice for the other person in relation to what they choose for themselves.

themselves and another person in the same session, whereas Eriksen and Kvaløy (2010b) assigned subjects to the role of investors and clients.

To summarize, we add to the referred literature above on the following aspects. Related to the literature on risk-taking with other people's money, we test risk-taking in the classical investment game of Gneezy and Potters, and let subjects invest both own money and other people's money. To the best of our knowledge, this is the first paper to let the same subject invest both own money and other people's money under the standard conditions of this investment game. Moreover, no one has previously varied feedback frequency (on investment outcomes) on this within dimension. Our results are mixed; subjects invest less with other people's money when feedback is low, but equally much when feedback is high. Related to the literature on MLA, we test the robustness of the previous finding that subjects exhibit MLA also when investing other people's money. Our results indicate that it is not robust to a situation where the investor also makes the same investment choice with own money, suggesting that they adapt their decision in such situations. Finally, we show that the relative investment between own and other people's money is different and dependent on the frequency of feedback on investment outcomes.

The rest of the paper is organized as follows: In section 2, we present the experimental design and procedures. Section 3 contains the results from the experiment. Section 4 briefly discusses findings and concludes.

2. Experimental Design and Procedures

2.1 Design

We adapt the well-known experimental design of Gneezy and Potters (1997), who tested the hypothesis of MLA by manipulating the feedback frequency of investment outcomes for subjects in a simple investment game. The experiment is a repeated investment game, where subjects can invest in a risky lottery for 12 periods. In each period, the subject starts out with an endowment of 100 ECU. The probability of winning the lottery is 1/3, while the probability of losing is 2/3. A win secures a return of 2.5 times the invested amount, otherwise the invested amount is lost. The amount they choose not to invest is kept for

certain. The lottery has a positive expected return.[178] Instructions can be seen in the appendix.

We vary the frequency of feedback on lottery outcomes between subjects and whether the investment decision is made with own money or with other people's money within subjects. Thus, as can be seen in Table 1, the experimental design has both a within and between dimension. In the experiment subjects receive feedback about lottery outcomes (and make investment decisions) either after every period (High Frequency, HF) or after every third period (Low Frequency, LF). In six of the twelve periods, subjects invest their own money (OWN), and in the remaining six periods they invest other people's money (OPM). The person they invest money on behalf of is, to them, an unknown subject within the same session. To control for potential discrepancy in the results due to order effects, we reverse the order in which subjects invest own and other people's money across sessions. Thus, with the variation along feedback frequency and the reversing of OWN versus OPM, we have four experimental treatments presented in Table 1.

Table 1: Experimental Design

	OWN → OPM	OPM → OWN	Number of subjects
High frequency (HF)	*HF-OWN-OPM* (40)	*HF-OPM-OWN* (42)	(82)
Low frequency (LF)	*LF-OWN-OPM* (39)	*LF-OPM-OWN* (34)	(73)
Number of subjects	(79)	(76)	(155)

Notes: The table presents the experimental design, and corresponding sample size for each of the four treatments.

[8]Expected outcome when investing X: $\left[X * \left(\frac{1}{3}\right) * \left(\frac{5}{2}\right)\right] + X - \left[X * \left(\frac{2}{3}\right)\right] = 0.167X + X$

2.2 Procedures

In each lottery round, subjects receive ECU 100 (Experimental Currency Unit) to invest or keep as certain gain. In Norwegian kroner, this is equivalent to 16 NOK (approx. $2). In total, subjects have ECU 1200 to invest. Final earnings depend on their investment decisions, the other person's investment decision and the lottery outcomes. The average earning across all treatments was 208 NOK. A session lasted about 15 to 20 minutes. We paid subjects in cash straight after the completion of the session.

The collection of data for the experiment occurred in two periods. The first period was during the spring of 2013.[9] To increase the number of observations, a second set of sessions were conducted in January 2015. The design in the second period was identical to the first, and we ran sessions of all four treatments in both periods. The subject pool is undergraduate students from the University of Stavanger, and 79 (76) students participated in 2013 (2015). Sample size for each treatment is

presented in Table 1.[10] We recruited students using their email accounts and from official university student organizations on Facebook.

When subjects went from investing own (other's) to other's (own) money, we included a questionnaire to make it clear that conditions changed in accordance with the instructions.[11, 12] After the conclusion of all 12 lottery rounds, we asked subjects to answer a questionnaire to collect demographic variables and some background information.

We read instructions aloud before the experiment started, and subjects had a written copy available throughout the experiment. We slightly changed the

[9]The first part of data collection was conducted as part of the master thesis of J. Heimdal, see Heimdal (2013).
[10]In one session, the total number of participants was an odd number. We therefore matched one subject with the average investment outcome of all the other subjects, when he or she received money from OPMs' investments. This subject was unaware of this.
[11]This questionnaire asked them to enter the computer id-number and asked a question about whether they would be interested in participating in similar experiments. As the order of the investment decision is reversed across sessions, we are not too worried about any potential demand effects from this questionnaire.
[12]Each investment page also stated the investment condition clearly.

formulation in the instructions (from the original instructions of Gneezy and Potters (1997)) on the lottery payoff. Specifically, we did not use an equation, but instead explained payoffs in words and by using examples.[13] The experiment was programmed and conducted using z-Tree (Fischbacher, 2007).[14]

3. Results

3.1 Main Results

Table 2 presents summary statistics for all treatments. The mean investment is close to ECU 60 in all treatments, but the median investment is lower in the two treatments where subjects make decisions concerning OPM first. The experimental data, with both the within and between dimensions, are more completely presented by Table 3. This table presents mean investments under the two different feedback regimes for both OWN and OPM.

Table 2: Summary Statistics

	Mean	St.dev	Median	# obs.	# subjects
HF-OWN-OPM	63.27	38.42	75	480	40
LF-OWN-OPM	62.85	33.32	67	468	39
HF-OPM-OWN	58.91	34.16	50	504	42
LF-OPM-OWN	63.35	30.52	60	408	34

Notes: The table presents summary statistics for the four treatments.

[13]We made this change to make it even easier to understand. Only once during all sessions were we asked about how payoffs worked, suggesting that this was not causing any ambiguities.

[14]In the original study by Gneezy and Potters (1997) they used pen and paper.

Table 3: Mean Investments by Feedback Frequency and OWN/OPM

		OWN			OPM		
	Period	Mean	SD	# obs./sub.	Mean	SD	# obs./sub.
HF	1 - 3	60.23	37.07	120/40	54.94	31.88	126/42
	4 - 6	58.63	41.66	120/40	56.39	35.07	126/42
	1 - 6	59.43	39.36	240/40	55.67	33.45	252/42
	7 - 9	60.75	33.63	126/42	69.87	35.17	120/40
	10 - 12	63.57	35.66	126/42	64.36	38.96	120/40
	7 - 12	62.16	34.62	252/42	67.11	37.14	240/40
	1 - 12	60.83	36.99	492/82	61.25	35.72	492/82
LF	1 - 3	59.38	30.3	117/39	55.94	31.34	102/34
	4 - 6	72.31	32.00	117/39	56.94	30.47	102/34
	1 - 6	65.85	31.76	234/39	56.44	30.84	204/34
	7 - 9	68.76	28.83	102/34	60.38	32.46	117/39
	10 - 12	71.76	28.51	102/34	59.31	36.77	117/39
	7 - 12	70.26	26.64	204/34	59.85	64.61	234/39
	1 - 12	67.90	30.39	438/73	58.26	32.92	438/73
Total		64.16	34.21	930/155	59.84	34.44	930/155

Notes: The table presents mean investments, standard deviation and number of observation and subjects by high and low feedback frequency and by OWN and OPM.

The analysis starts by looking at the between-subject dimension of our design. Do people exhibit MLA when investing OWN? Column two in Table 3 shows investments concerning OWN under the two feedback conditions. The average investment over all twelve periods is higher under LF feedback compared to HF feedback. Subjects who receive frequent feedback and make frequent investment decisions invest about ECU 61, while subjects who receive less frequent feedback and invest more infrequently invest on average about ECU 68. This difference is not significant at the 10% level (p=0.103) using a two-sided Mann-Whitney U-test, as can be seen in the lower left corner of Table 4.[15] The size of the difference is similar to other related papers; however, the

[15]We use only one observation per subject. Also, for the tests in periods 1-6 and periods 7-12, we only use about half of the sample in each test (i.e. only subjects who actually invested OWN in the first 6 periods are included in periods 1-6, and only subjects who actually invested OWN in the last 6 periods are included in periods 7-12).

standard deviations are higher.[16] As we will see in the regression analysis that follows in section 3.2, our result is stronger when we organize data as a panel and add controls.

[16]For example, Zeisberger et al. (2014) have means equivalent to 67.2 and 58.3, but with standard deviations as low as 22.6 and 20, respectively.

Table 4: Mean Investments by Feedback Frequency and OWN versus OPM

	OWN			OPM			HF feedback			LF feedback		
	HF feedback vs. LF feedback			HF feedback vs. LF feedback			OWN vs. OPM			OWN vs. OPM		
Period	n	z-value	p-value	n	z-value	p-value	n	z-value	p-value	n	z-value	p-value
1 - 6	79	1.013	0.311	76	0.125	0.900	82	-0.538	0.590	73	-1.531	0.126
7 - 12	76	1.236	0.210	79	-0.820	0.411	82	0.744	0.457	73	-1.524	0.127
1 - 12	155	1.629	0.103	155	-0.451	0.652	164	0.064	0.949	146	-2.130	0.033

Notes: Test statistics and p-values are based on two-sided Mann-Whitney U-tests comparing investment choices by feedback frequency and OWN versus OPM.

When subjects make decisions with OPM, column five in Table 3 shows that there are only small between-subject differences in investment between low and high feedback frequency. Under HF feedback, the mean investment is ECU 61, while under LF feedback it is ECU 58. Thus, these results suggest that subjects in our experiment did not exhibit MLA when they invested OPM.[17] Moreover, although far from significant, the difference is even in the opposite direction of what the MLA hypothesis predicts.[18] This result contradicts those of Eriksen and Kvaløy (2010b), suggesting that their result is not robust to changes in the experimental context, either because the experiment is framed differently or subjects need to adapt their investment decision to concern both own and other people's money.

Result 1: *When subjects invest other people's money, they do not exhibit investment behavior consistent with MLA.*

The analysis now turns to the within-subject part of the experiment. Do subjects invest differently with own and other's money? From Table 3 we see that subjects invest less with OPM (ECU 58) than with OWN (ECU 68) when feedback frequency is low. The MW-test furthest to the right in Table 4 shows that this difference is significant (p=0.033).[19] Hence, subjects who are in the LF feedback regime choose to invest significantly less with OPM compared to OWN. On the other hand, subjects investing under HF feedback choose to invest about the same amount with OPM (ECU 61.25) as with OWN (ECU 60.83).

[17]To obtain a significant difference at the 10% significance level, with power 80% and the given number of observations in our experiment, we would need an effect size of d=0.41. As this is quite large, meaning subjects investing OPM with low frequency would have to invest about ECU 11 more for the difference to be significant, the insignificant result must be interpreted cautiously.
[18]Related MW tests are in Table 4. The use of observations are equivalent as described under investing OWN.
[19]These tests (for periods 1-12) are based on two observations per subject (one observation with OWN and one observation with OPM)

If we then combine the within and between-subject part of the design, we see that the relative investment made under LF with OWN and OPM is different from the same relative investment made under HF.[20] Thus, the manipulation of feedback frequency does not seem to make people less afraid of risk when they invest OPM. Moreover, in Table 5, we see that in three out of the four treatments, subjects invest less under OPM compared to OWN. A potential explanation for this result is that subjects perceive themselves as being more risk-seeking than others, hence acting in a more risk-averse manner when investing on behalf of others. This would be consistent with the *risk-as-value* hypothesis proposed by Brown (1965).

[20]Independent of feedback, the average investment with OPM (ECU 64) is insignificantly (p=0.15) greater than OWN (ECU 60).

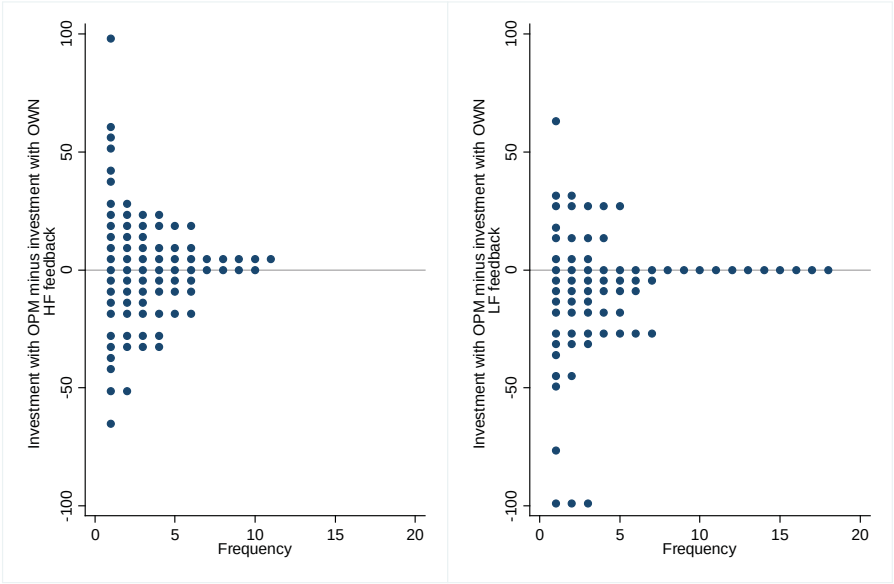
Table 5: Within Treatment Investments

Treatment	Mean	Mean OWN	Mean OPM	z-value	p-value	# subjects
HF-OWN-OPM	63.27	59.43	67.11	-1.71	0.09	40
LF-OWN-OPM	62.85	65.85	59.85	0.59	0.55	39
HF-OPM-OWN	58.91	62.16	55.67	1.86	0.06	42
LF-OPM-OWN	63.35	70.26	56.44	4.09	<0.01	34

Notes: The table presents mean investments for the four treatments, as well as Wilcoxon matched-pairs signed-ranks test comparing within subject investments.

To investigate this a bit further, in Figure 1 we plot average individual differences between investments under OPM and OWN. To the left are subjects who invest under HF feedback. Here, almost as many subjects take less risk with OPM, as there are subjects taking more risk. To the right are subjects who invest under LF feedback. The number of subjects who reduce risk with OPM is higher than subjects who increase this risk. There are also plenty more subjects who invest the same across OWN and OPM compared to HF feedback. Thus, when the feedback frequency is low, subjects seem less inclined to take risk with OPM relative to OWN.

Figure 1: Differences in Investment with OPM and OWN



Notes: The plot illustrates the frequency of subjects (X) who invest different amounts with OPM compared to OWN. HF feedback to the left. LF feedback to the right.

In Table 6, under HF feedback, 96% of all subjects (77 out of 82) make a different investment choice with OPM compared to OWN. Consistent with Figure 1, there are almost as many who take less risk with OPM as there are subjects taking more risk with OPM. Under LF feedback, the same symmetry cannot be observed. Here, 75% (55 out of 73) of all subjects make a different

choice with OPM. However, far more subjects take less risk with OPM than the opposite. From column one in Table 6, we see that 40/82 (49%) of subjects under HF feedback increase risk when investing on behalf of others, whereas only 16/73 (22%) do so under LF feedback. This difference in proportion is significant (test of equal proportions: $z=-3.475$, $p<0.01$). Further, only 6% of subjects receiving HF feedback invest the same amount under both OWN and OPM, compared to 25% of the subjects who receive LF feedback (test of equal proportions: $z=3.245$, $p<0.01$). Finally, more subjects in the LF feedback regime (39/73) reduce investments under OPM, compared to subjects from the HF feedback regime (37/82). However, this difference is not significant. We also compared the investment levels between low and high frequency feedback for those who take less (or more) risk with other people's money. Overall, the absolute change in investments when going from OWN to OPM is equal and independent of feedback frequency.

Result 2: *When feedback frequency is high, subjects invest on average the same amount in the risky lottery for both themselves and others. However, when feedback frequency is low, subjects invest significantly less in the risky lottery on behalf of others compared to what they do for themselves.*

Table 6: Investing More, Equally, or Less with OPM

	N	OPM (1)	OWN (2)	OPM-OWN (3)
Panel A: HF feedback				
Less risk OPM	37	53.79	73.33	-19.54
Same risk	5	96.67	96.67	0.00
More risk OPM	40	63.73	44.78	18.95
	82			
Panel B: LF feedback				
Less risk OPM	39	46.46	73.05	-26.59
Same risk	18	80.92	80.92	0.00
More risk OPM	16	61.53	40.72	20.81
	73			

Notes: Numbers displayed are averages in ECU.

3.2 Robustness of Results

By constructing the data as a panel, we run Tobit regressions without and with Random Effects, displayed in Table 7. As the differences between columns are small, we only comment on results with Random Effects. Observations are censored at the lowest (0) and highest (100) possible investment amount. We use four observations per subject, so that for subjects in HF feedback (who actually make 12 decisions) we use the average of three choices as one observation.

Table 7 generally portrays the same results as in the previous section. First, investors exhibit MLA when investing OWN, as the coefficient for HF feedback is significantly lower than the reference category (LF feedback, OWN). Second, people do not seem to exhibit MLA when investing OPM, as there is no difference between investments across feedback conditions (difference is ECU 2.45= ECU 15.25 – ECU 12.80, p=0.655). Third, notice that the OPM coefficient is significantly negative, meaning that under LF feedback, subjects invest less with OPM than with OWN. For subjects under HF feedback, the OPM coefficient and the interaction between OPM and HF

feedback cancel each other out, so there is no difference in the investment with OPM compared to OWN. Finally, as subjects under LF feedback lower their investment with OPM (- ECU 14.62), and subjects under HF feedback do not (ECU 0.63 = - ECU 14.62 + ECU 15.25), the change in relative investment is different under LF feedback and HF feedback.

Result 3: *When subjects invest their own money, they show behavior consistent with MLA.*

Table 7: Robustness of Results

	Tobit	RE Tobit
	(1)	(2)
LF feedback	Ref.	Ref.
OPM	-14.13***	-14.62***
	(4.581)	(3.582)
HF feedback	-11.71**	-12.80**
	(4.570)	(5.535)
OPM x HF feedback	14.44**	15.25***
	(6.239)	(4.847)
Observations	620	620
Log-likelihood	-2410	-2353
Left-censored obs.	22	22
Right-censored obs.	153	153

Notes: Column (1) is a Tobit regression without random effects. Column (2) is a Random Effects Tobit regression. Dependent variable is the individual investment. LF feedback investing OWN is the reference category. OPM is an indicator equal to 1 if the investment is on behalf of other people's money, 0 otherwise. HF feedback is an indicator equal to 1 if subjects get high frequency feedback, 0 otherwise. Both columns include controls for gender, age, marital status, study program, grades, order of investment condition, whether the subject regularly invests in stocks or mutual funds, and participation in similar experiments. Standard errors are in parentheses.

*$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.3 Differential Analysis: Gender

There may be gender differences in investment decisions, and in particular, whether subjects exhibit MLA and their willingness to take risk. In Table 8 we run subsamples of females and males, respectively. Both genders seem to exhibit MLA when investing own money, and in particular males.[21] As in the main analysis, neither gender exhibits MLA when investing OPM. Under LF feedback, both genders invest significantly less with OPM compared to OWN, and in particular males.[22] Under HF feedback, males invest slightly more with OPM (ECU 8.96 = - ECU 27.89 + ECU 36.85), whereas females invest slightly less with OPM (- ECU 6.65 = - ECU 9.46 + ECU 2.81). In a pooled regression, males invest significantly more with OPM (p=0.066) than females do under HF feedback.

[21]The difference between females and males is not significant in a pooled regression.
[22]The difference between females and males is not significant in a pooled regression.

Table 8: Differential Analysis: Gender

Sample:	Females	Males
	(1)	(2)
LF feedback, OWN	Ref.	Ref.
OPM	-9.46***	-27.89***
	(3.615)	(7.961)
HF feedback	-9.89*	-21.35**
	(5.916)	(10.775)
OPM x HF feedback	2.81	36.85***
	(5.246)	(9.844)
Observations	352	268
Log-likelihood	-1405	-915
Left-censored obs.	5	17
Right-censored obs.	64	89

Notes: Columns are estimated using Random Effects Tobit regressions, and represent subsamples of each gender separately. Dependent variable is the individual investment. LF feedback with OWN is the reference category. OPM is an indicator equal to 1 if the investment is on behalf of other people's money, 0 otherwise. HF feedback is an indicator equal to 1 if subjects get high frequency, 0 otherwise. Both columns include controls for age, marital status, study program, grades, order of investment condition, whether the subject regularly invests in stocks or mutual funds, and participation in similar experiments. Standard errors are in parentheses.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

4. Discussion and Concluding Remarks

In this paper, we have experimentally investigated how people's investment choices are affected by variations in feedback frequency on outcomes, when they invest both own and other people's money. We find that the manipulation matters in several aspects. First, the relative investment people make with their own money and other people's money is different depending on feedback

frequency. Specifically, when the feedback frequency is low, they invest significantly less with other people's money, whereas when feedback frequency is high, they invest about the same amount with own and other people's money. Second, when subjects invest their own money, they invest less when feedback frequency is high relative to when feedback frequency is low. This supports the MLA hypothesis. When investments are on behalf of someone else, on the other hand, we do not find any difference in investment between high and low feedback frequency. Average investments are higher when subjects invest own money compared to other people's money.

Our results show that the findings of Eriksen and Kvaløy (2010b) on people exhibiting MLA with other people's money is not robust to a change in the experimental context. Specifically, our subjects had to adapt their investment decision to a situation where they had to consider investments with both own and other people's money in a neutrally framed experiment. In contrast to them, we find no evidence suggesting that subjects exhibit MLA when investing on behalf of others. Both these studies find that subjects choose lower investments with other people's money.

MLA is a well-established and plausible theoretical explanation for the equity premium puzzle. However, many people outsource the investment of their wealth to some professional investor, and knowledge about how people choose to invest on behalf of others is important; it may improve our understanding of the extent of the MLA hypothesis and more generally about risk attitudes when making decisions on behalf of others. Our study suggests that feedback on investment outcomes is important, as it influence how people make decisions with their own money relative to how they invest on behalf of others. Unfortunately, it does not provide an answer to why this is so. We can only provide speculative reasons, such as choosing to invest in accordance with their beliefs about the other person's preferences or to invest a "middle way" when investing on behalf of others. More research is required to give us a broader understanding of these issue.

References

Agranov, M., Bisin, A., & Schotter, A. (2014). An experimental study of the impact of competition for other people's money: The portfolio manager market. *Experimental Economics, 17*(4), 564-585.

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2016). Deciding for others reduces loss aversion. *Management Science, 62*(1), 29-36.

Bellemare, C., Krause, M., Kröger, S., & Zhang, C. (2005). Myopic loss aversion: Information feedback vs. investment flexibility. *Economics Letters, 87*(3), 319-324.

Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics, 110*(1), 73-92.

Bolton, G. E., & Ockenfels, A. (2010). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment. *The American Economic Review, 100*(1), 628-633.

Chakravarty, S., Harrison, G. W., Haruvy, E. E., & Rutström, E. E. (2011). Are you risk averse over other people's money? *Southern Economic Journal, 77*(4), 901-913.

Charness, G., & Jackson, M. O. (2009). The role of responsibility in strategic risk-taking. *Journal of Economic Behavior & Organization, 69*(3), 241-247.

Eriksen, K. W., & Kvaløy, O. (2010a). Do financial advisors exhibit myopic loss aversion? *Financial Markets and Portfolio Management, 24*(2), 159-170.

Eriksen, K. W., & Kvaløy, O. (2010b). Myopic investment management*. *Review of Finance, 14*(3), 521-542.

Eriksen, K. W., Kvaløy, O., & Luzuriaga, M. (2015). *Risk-taking with other people's money*. University of Stavanger.

Fellner, G., & Sutter, M. (2009). Causes, consequences, and cures of myopic loss aversion – An experimental investigation*. *The Economic Journal, 119*(537), 900-916.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, Springer, 10(2)*( ), 171-178, June.

Füllbrunn, S., & Luhan, W. J. (2015). *Am I my peer's keeper? Social responsibility in financial decision making*. Ruhr Economic Papers.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics, 112*(2), 631-645.

Haigh, M. S., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis. *The Journal of Finance, 60*(1), 523-534.

Heimdal, J. K. (2013). *Myopic loss aversion and the equity premium puzzle.* (Master Thesis), University of Stavanger, Norway. (UIS-SV-HH/2013)

Heinkel, R., & Stoughton, N. M. (1994). The dynamics of portfolio management contracts. *Review of Financial Studies, 7*(2), 351-387.

Investment Company Fact Book. (2016). www.icifactbook.org: Investment Company Institute (ICI).

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American psychologist, 39*(4), 341-350.

Langer, T., & Weber, M. (2008). Does commitment or feedback influence myopic loss aversion?: An experimental analysis. *Journal of Economic Behavior & Organization, 67*(3–4), 810-819.

Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics, 15*(2), 145-161.

Pollmann, M. M. H., Potters, J., & Trautmann, S. T. (2014). Risk taking by agents: The role of ex-ante and ex-post accountability. *Economics Letters, 123*(3), 387-390.

Polman, E. (2012). Self–other decision making and loss aversion. *Organizational Behavior and Human Decision Processes, 119*(2), 141-150.

Reynolds, D. B., Joseph, J., & Sherwood, R. (2011). Risky shift versus cautious shift: Determining differences in risk taking between private and public management decision-making. *Journal of Business & Economics Research (JBER), 7*(1).

Sutter, M. (2007). Are teams prone to myopic loss aversion? An experimental study on individual versus team investment behavior. *Economics Letters, 97*(2), 128-132.

Thaler, R. H. (2005). *Advances in behavioral finance* (Vol. 2): Princeton University Press.

Thaler, R. H., Tversky, A., Kahneman, D., & Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics*, *112*(2), 647-661.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323.

Zeisberger, S., Langer, T., & Weber, M. (2014). *Do changes in reporting frequency really influence investors' risk taking behavior? Myopic loss aversion revisited*. SSRN. Retrieved from https://ssrn.com/abstract=1786360

Appendix

Experimental Instructions (LF-OWN-OPM, translated from Norwegian)[23]

The experiment consists of 2 games with 6 rounds. After game 1 there will be a short questionnaire, while after game 2 there will be a longer questionnaire.

1. In the first six rounds (game 1) you will be making investment choices on behalf of yourself, while in the last six rounds (game 2) you will make decisions on behalf of another person.
2. You will start with 100 experimental units (EK) in every round.

How does the Investment Choices Work?

- You make a decision for three rounds at a time when you decide how much to invest in the lottery (from 0 to 100)
- With 1/3 chance the lottery will give you 2,5 times the investment in return, and with 2/3 chance you will lose the investment. (Example: If you invest 100, then 1 out of 3 times you will receive 350, and 2 out of 3 times you will get 0 in return)
- The amount you choose not to invest will be yours for certain (Example: If you bet 0 in the lottery you get 100 for certain in that round)
- Participants will randomly be divided into type 1, 2 or 3 each round, and one of these types will randomly be drawn as the winner
- 100 EK is equal to 16 Norwegian kroner (NOK)

[23]We only include the instructions from one version of LF feedback and HF feedback, the first is an example where they started investing OWN before OPM and the second is the opposite.

Game 1

In this game you will make investment choices on behalf of yourself. Here you will influence your own payment. After round 3 you will get the results from rounds 1 to 3. Thereafter you must decide how much you want to invest in the lottery in rounds 4 to 6.

Game 2

In this game you will make investment choices on behalf of another person in the same way as game 1. The choices you make will influence another participant's payment. One randomly chosen participant will do the same for you.

Payment

The amount of Norwegian kroner (NOK) you have after the first 6 rounds will be paid out after both game 1 and game 2 are finished. In game 2 another participant will have made choices that affect your payment in game 2.

Practical Information

Follow the instructions on the screen, enter (0 to 100) and press the OK-button along the way. Take into account that there might be some waiting during the experiment. You are not allowed to talk and/or have contact with other participants. If you have any questions, raise your hand and we will answer the question. After the experiment is completed, you will write your name on the receipt sheet.

Experimental Instructions (HF-OPM-OWN, translated from Norwegian)

1. The experiment consists of 2 games with 6 rounds. After game 1 there will be a short questionnaire, while after game 2 there will be a longer questionnaire.
2. You will make investment choices on behalf of another person the first six rounds (game 1), while in the last six rounds (game 2) you will make decisions on behalf of yourself.
3. You will start with 100 experimental money (EK) in every round.

How does the Investment Choices Work?

- You make a decision for each round when you decide how much you want to invest in the lottery (from 0 to 100)
- With 1/3 chance the lottery will give you 2,5 times the investment in return, and with 2/3 chance you will lose the investment. (Example: If you invest 100 then 1 out of 3 times you will receive 350 and 2 out of 3 times you will get 0 in return)
- The amount you choose not to invest will be yours for certain (Example: If you bet 0 in the lottery you get 100 for certain in that round)
- Participants will randomly be divided into type 1, 2 or 3 each round, and one of these types will randomly be drawn as the winner
- 100 EK is equal to 16 Norwegian kroner (NOK) every round.

Game 1

In this game you will make investment choices on behalf of another person. The choices you make will influence another participant's payment. One randomly chosen participant will do the same for you. After round 1 you will get the result from this round. Thereafter you must decide how much you want to invest in the lottery in round 2. After the draw you get the results from this round. This procedure will also be the same for rounds 3 to 6.

Game 2

In this game you will make investment choices on behalf of yourself in the same way as in game 1. Here you will influence your own payment.

Payment

The amount of Norwegian kroner (NOK) you have after the first 6 rounds will be paid out after both game 1 and game 2 are finished. In game 1 another participant will have made choices that affect your payment in game 1.

Practical Information

Follow the instructions on the screen, enter (0 to 100) and press the OK-button along the way. Take into account that there might be some waiting during the experiment. You are not allowed to talk and/or have contact with other participants. If you have any questions, raise your hand and we will answer the question. After the experiment is completed, you will write your name on the receipt sheet.