Essays in Organizational Economics

Universität Hamburg
Fakultät Wirtschafts- und Sozialwissenschaften

Dissertation
Zur Erlangung der Würde eines Doktors
der Wirtschafts-und Sozialwissenschaften
„Dr. rer. pol.“
(gemäß der PromO vom 24. August 2010)

vorgelegt von
Niklas Wallmeier
aus Münster (Westf.), Deutschland

Hamburg, 04. Juli 2019

Thesis Committee

Chairman: Prof. Dr. Andreas Lange

First Examiner: Prof. Dr. Gerd Mühlheußer

Second Examiner: Prof. Thomas Siedler, (PhD)

Third Examiner: Prof. Dr. Jan Marcus

The disputation was held on August 28, 2019.

Wie man's macht, macht man's richtig.

Für Brigitte und Martin und Wim.
Für Vivian.

Acknowledgements

First and foremost I am indebted to Gerd Mühlheußer for his encouragement, loyalty and patience. I learned about the importance of purposeful, diligent and dedicated research which I will greatly benefit from in my future challenges.

Wim Kösters and John Haisken-DeNew deserve my deepest thanks for enthusing me with economics at my first internship at the RWI and accompanying me through my entire academic life. Also, I was fortunate to work with my co-authors Andreas Roider, Sandra Schneemann, Dirk Sliwka and Kathrin Thiemann, who offered a tremendous support and gave me the opportunity to learn from their qualities. A number of colleagues, for example Berno Büchel, Leonie Gerhards, Christos Litsios, Pamela Mertens, Achim Voß and many others, contributed to an inspiring and amicable work environment.

Seminar and conference participants in Cologne, Essen, Hamburg and Melbourne enriched my research with critical comments and valuable suggestions. Leo Kahane, Roberto Serrano, Ralph Bayer and various anonymous referees provided thorough review services which helped to improve my papers exceptionally.

Financial support of the WiSo Graduate School and the support by Olaf Bock (representative of the team of the WiSo research laboratory) enabled a large part of my projects and are gratefully acknowledged.

Finally, I thank my family, most of all my partner Vivian Pasquet and my parents Brigitte Wallmeier and Martin Freitag, my friends and colleagues who told me my research was interesting and cool when I was in doubt and, of course, for challenging me by asking '*what do you want to do with this in the future*?' which helped me figuring out that *this* is what I want to do in the future.

Contents

Chapter 1.

Introduction

Gibbons and Roberts (2012, p. 1) describe organizational economics as applying "economic logic and methods to understand the existence, nature, design and performance of organizations, especially managed ones." This dissertation addresses key questions of this field in four empirical studies.

Economists have long suggested that variations in management practices drive productivity differences (Syverson, 2011). Building on this, two studies center on the role of management personnel. The first discusses the impact of managers for organizational success and its implications for personnel management (Chapter 2, co-authored by Gerd Mühlheußer, Dirk Sliwka and Sandra Schneemann). The second analyzes the effect of legal whistleblower protection on the cooperative climate between managers and employees, and its implications for the design of such a law (Chapter 3).

The two other studies focus on the role of groups in organizations, as "[a]ll organizations share the need for collective action and the allocation of resources through non-market methods,"(Arrow, 1974, p. 33). The first investigates the relation between group composition and unethical behavior in collective decision-making (Chapter 4, co-authored by Gerd Mühlheußer and Andreas Roider). The second asks whether group composition can be designed optimally with respect to individual characteristics of the group members and the effect of relative feedback on performance (Chapter 5, co-authored by Kahtrin Thiemann). In the following, I describe each chapter and its primary contributions to the economics literature.

Chapter 2 investigates a manager's share in the success of an organization. That managers affect productivity has been proposed by Walker (1887) and was emphasized by Leibenstein (1966) and Lucas (1978). The ability of the person at the top affects an organization through a number of channels, should trickle down through the hierarchy, and thus have a strong effect on organizational performance (Rosen, 1982). But how big are these effects? What difference does the quality of the single person at the top make for the overall performance of the organization?

The empirical literature, which aims at measuring the contribution of individual managers to the performance of their organization (see e.g., Bertrand and Schoar, 2003;

Lazear, Shaw, and Stanton, 2015; Graham, Li, and Qiu, 2012), exploits the variation which arises from the fact that, in the course of the careers, some managers are active in several organizations or functions which allows to disentangle their contribution from other factors. This is a difficult endeavor as CEOs, for instance, typically stay at the top of a specific firm for longer time periods and work as CEOs only for a very small number of different firms in their lifetime – which limits the scope to measure their contribution to organizational success.

We consider this issue in the context of professional sports which has advantages for the question at hand: (i) team performance is publicly observable on a weekly basis and (ii) managers move very frequently between teams – much more frequently than managers in firms. Also, observing the same manager in different organizations, thus using different sets of resources and working with different people, is crucial to measure a manager's contribution to overall success. We use this information to estimate the impact of managers on team success, thereby also addressing the practical debate on this issue.

This is the first study to apply this idea to the professional sports sector. It contributes to the growing literature empirically analyzing the impact of managers on different economic measures, such as corporate behavior (Bertrand and Schoar, 2003), corporate tax avoidance (Dyreng, Hanlon, and Maydew, 2010), managerial compensation (Graham, Li, and Qiu, 2012), or disclosure choices (Bamber, Jiang, and Wang, 2010). In addition, all managers in our study operate in the same industry, and this industry attracts a huge amount of public attention. As a result, most of these managers are very well-known to the interested public, so that the estimated individual effects are of interest in their own right. We find a considerable variation in performance contributions. Moreover, we document an impact of managers on teams' style of playing, and we show that once famous and successful players do not necessarily make good managers later on in their careers. Furthermore, we show that the estimated effects are useful to predict performance later in the managers' careers. Hence, our results can be helpful in identifying "under-valued" managers.

Another key question of organizational economics concerns the norms regarding behavior toward other members in the organization. How do these norms affect behavior and organizational performance and how does it depend on the social, legal and regulatory environment (Gibbons and Roberts, 2012, p. 3)? In Chapter 3, we focus on the

norms regarding unethical behavior and loyalty within the organization. Specifically, we want to analyze how whistleblower protection, as an instrument to fight corporate fraud, affects the cooperation between a manager and an employee.

The extensive and widespread economic damage of corporate fraud is well documented (Economist Intelligence Unit, 2015). Accordingly, detection and deterrence of corporate fraud has become a major target for policy makers. There are intuitive arguments for the use of insider knowledge for law enforcement. First, using insider knowledge increases the share of fraud cases that can be detected. Second, whistleblowing might not only facilitate law enforcement, but the mere threat of insiders reporting could deter wrongdoing ex-ante. Correspondingly, evidence in favor of whistleblowing as an instrument for crime deterrence is prominent in economic literature (Dyck, Morse, and Zingales, 2010). Yet, becoming a whistleblower comprises a non-negligible trade-off for an organization member. She potentially faces costs from a breach of loyalty and career risks in the form of a dismissal or a denied promotion which hamper whistleblowing (see, e.g. Near and Miceli, 1986; Alford, 2001; Rehg, Miceli, Near, and Van Scotter, 2008; Cassematis and Wortley, 2013). As a consequence, whistleblowers might be encouraged to come forward by legally protecting them from retaliation. To this end, international organizations as the G20 group, or the OECD requested protection for whistleblowers (OECD, 2016) and legislators made an effort to increase the legal certainty (According to Thüsing and Forst, 2016, 15 out of 23 surveyed countries have implemented a specific whistleblowing law).

However, these legal approaches are discussed controversially. Whistleblower laws often contain the low barrier of a 'reasonable belief' which deters obviously unfounded complaints, but may nevertheless result in an increase in false claims (Callahan and Dworkin, 1992; Howse and Daniels, 1995; Givati, 2016). Furthermore, efficiency does not solely rely on lawful behavior, but also on trust as a foundation for productive cooperation to take place. This trust between co-workers might be reduced if employees use sensitive information to file a complaint. Therefore, if a whistleblower protection law encourages more reporting, it may hinder beneficial cooperation (Dworkin and Near, 1997).

This study focuses on these two costs of whistleblower protection and evaluates them against the benefits of detecting and deterring misbehavior due to protection laws in an experimental setting. Our main goal is to investigate the influence of whistleblower protection on reporting behavior, compliance and cooperation. We consider the two

most frequent instruments of whistleblower protection, namely the provision of incentives and anonymity for reporting. In the context of whistleblowing, a laboratory approach has two major advantages compared to the field. First, only detected fraud is observable in actual organizations, such that the true amount of misbehavior remains unknown. Second, we only observe reporting behavior given the state of compliance.

Our results suggest that whistleblower laws offer a rich potential for fighting the damage of corporate fraud through both increased deterrence and detection, but also provoke adverse effects. Since employment protection provides an incentive for reporting in general, it does not only increase truthful reporting, but also triggers false whistleblowing by the employees. A novel finding of this paper relates to further costs associated with false reports in the form of decreased cooperation - especially in the presence of false whistleblowing. The results of our study suggest that a legislator could pass different tailor-made laws for different sectors, since the importance of cooperation may well vary between the industries of an economy. For example, laws that apply to organizations where efficiency is driven rather by compliance than by cooperation, could be designed similar to the one presented in this chapter. In contrast, if a company's success heavily depends on productive cooperation, the policy could acknowledge this by avoiding an excessive amount of false claims at the cost of non-maximal deterrence.

Chapter 4 also considers unethical behavior, but turns the focus away from individual and towards collective decision-making. A number of experimental studies have analyzed lying as one prominent type of unethical behavior. There is strong evidence for lying, but often not to the maximal extent possible; suggesting that there are private costs associated with such unethical behavior (see e.g., Abeler, Raymond, and Nosenzo, 2018). With respect to gender differences, it seems that males are somewhat more prone to lying than females, but often the effect is small or not statistically significant (Dreber and Johannesson, 2008; Childs, 2012; Erat and Gneezy, 2012; Houser, Vetter, and Winter, 2012; Conrads, Irlenbusch, Rilke, and Walkowitz, 2013; Conrads, Irlenbusch, Rilke, Schielke, and Walkowitz, 2014; Abeler, Becker, and Falk, 2014).

The literature on lying behavior has mainly analyzed decisions by *individuals*; possibly in strategic interaction with other individuals as in tournaments (see e.g., Conrads, Irlenbusch, Rilke, Schielke, and Walkowitz, 2014). However, in important economic, social, or political decisions, a *group* of individuals must reach a decision *jointly*. In

fact, there is growing evidence from contexts other than lying that groups often decide markedly different than individuals (for surveys, see Charness and Sutter, 2012; Kugler, Kausel, and Kocher, 2012). On the one hand, groups are better at solving cognitive tasks and act more selfishly (see e.g., Maciejovsky, Sutter, Budescu, and Bernau, 2013; Bornstein, Kugler, and Ziegelmeyer, 2004; Falk and Szech, 2013). That suggests that groups might be more willing to realize the potential monetary gains from lying. On the other hand, there is evidence that "moral reminders" reduce dishonesty (Pruckner and Sausgruber, 2013). Hence, discussions within groups might lead them to lie less. Taken together, it seems a priori unclear whether lying is more prevalent in groups compared to individuals. Moreover, for the lying behavior of groups their gender composition might matter (see e.g., Dufwenberg and Muren, 2006). Consequently, this chapter aims at providing insights on the unethical behavior of groups and individuals, and the role of gender in this context.

In line with the previous literature, we find no clear evidence for gender differences under individual decision-making on lying. In contrast, in the case of group decision-making, more pronounced gender effects arise; resulting in more (less) aggregated unethical behavior in male (female) groups. Moreover, male groups seem to have a greater tendency towards exploiting the full gains from lying than female groups. Finally, mixed groups with equal shares of males and females behave similarly to male groups. Hence, organizational designers might want to pay particular attention to decisions that are taken by purely male (or male-dominated) groups.

Organizational economics also addresses the question how rewards affect behavior and performance. With respect to this, Chapter 5 investigates the additional motivational effect of relative performance feedback. Performance feedback is frequently leveraged to induce a change in behavior. For instance, 60% of manufacturing firms reveal performance data to their employees (Bloom and Van Reenen, 2007). If individuals have reference-dependent preferences (see e.g. Tversky and Kahneman, 1979; Kőszegi and Rabin, 2006), the effect of recognition can depend on the kind of relative performance feedback and on the performance distribution in the reference group. We consider different kinds of performance feedback since it may vary with an organization's philosophy. Some firms actively highlight only the top performers (e.g. the "employee of the month", Kosfeld and Neckermann, 2011) in order to motivate employees to perform better. The reference point can also vary with culture as acquired by groups of

people that share a religion or ethnic origin. In more competitive cultures, individuals may be expected to compete for the top positions. Opposite, in less competitive cultures, social comparison may play a less emphasized role and individuals are expected to compare themselves to the average. Therefore, the question arises whether group composition can be optimized for a given performance feedback in order to maximize group performance.

Thiemann (2017) addresses this issue theoretically, focusing on the question whether ability-segregated classes (also referred to as *ability tracking* or *ability grouping*) or classes with students of heterogeneous ability are preferable. Theory predicts that it depends on the culture of competitiveness of the student body, that is, on the kind of the reference point and the importance of social comparison. In an laboratory experiment, we test these predictions in environments where subjects perform a real-effort task while they are evaluated either against the *average* or the *best* performance of their reference group. To affect the ability distribution within the reference group, the members are drawn randomly either from the *entire* pool of participants or only from those of the same ability category (*high* or *low*). In addition, we test the role of gender concerning the optimal performance feedback and grouping policy. Gender might be of importance, since women and men have been found to differ to a huge extent in their preferences for competitiveness (see e.g. Gneezy, Niederle, Rustichini, et al., 2003; Niederle and Vesterlund, 2007; Niederle, 2016). In our experiment, high reference points and pressure for social comparison will create a competitive environment and might cause different effort choices of male and female subjects.

Our study contributes to two fields of economic literature. The first is the empirical literature on peer effects (for an overview see Herbst and Mas, 2015). While these studies focus on a single performance feedback, we contrast the effects of different relative performance feedback: the *average* peer achievement and the *best* peer performance. Second, our study contributes to the literature that addresses the effect of grouping individuals according to their ability. (e.g. see surveys by Slavin, 1990; Meier and Schütz, 2008).

We find support only for male subjects behaving according to theory-derived optimal performance. On the other hand, women even reduce output in response to being told to have performed worse than the best in their group, underlying that women behave contrary to our theoretical predictions. With respect to the grouping treatments, we find that female mean performance is significantly lower under *random grouping* than

under *ability grouping*, while men perform significantly better under *random grouping*.

Our findings have implications for the design of feedback technologies and grouping procedures. Based on our results, copying successful designs may not be a promising strategy when the characteristics of the target group are substantially different. Instead, a decision maker may acknowledge the individuals' background.

The main contributions of this dissertation concern the design of organizations and the evaluation of performance. Chapter 2 shows that managerial talent is important for organizational success and, even though the observation of talent may be difficult due to confounding factors, identification is possible if certain conditions are fulfilled. Also with respect to the performance of an organization, Chapter 5 stresses the important role of the heterogeneity of a group for the motivational effect of performance feedback. Chapter 4 considers the composition of a group from a different perspective. The findings suggest that groups, which are at risk of making unethical decisions, might tend rather to honest decisions if they are gender-balanced. Finally, Chapter 3 also relates to unethical decisions and demonstrates that there may be a trade-off between encouraging the reporting of unethical behavior on the one hand, and fostering cooperation on the other hand.

Chapter 2.

The Contribution of Managers to Organizational Success: Evidence from German Soccer[1]

Abstract

We study the impact of managers on the success of professional soccer teams using data from the German *Bundesliga*, where we are exploiting the high turnover of managers between teams to disentangle the managers' contributions. Teams employing a manager from the top of the ability distribution gain on average considerably more points than those employing a manager from the bottom. Moreover, estimated abilities have significant predictive power for future performance. Also, managers also affect teams' playing style. Finally, teams whose manager has been a former professional player perform worse on average compared to managers without a professional player career.

JEL-Codes: J44, J63

Keywords: Managerial Skills, Human Capital, Empirical, Fixed Effects, Professional Sports

2.1 Introduction

It is widely believed that managers have a huge impact on the success of organizations. The ability of the person at the top affects an organization through a number of channels and should trickle down through the hierarchy and thus have a strong effect on organizational performance (Rosen, 1982). But how big are these effects? What difference does the quality of the single person at the top make for the overall performance of the organization? There is a recent empirical literature which aims at measuring the contribution of individual managers to the performance of their organization (see e.g., Bertrand and Schoar, 2003; Lazear, Shaw, and Stanton, 2015; Graham, Li, and

[1]This chapter is co-authored by Gerd Mühlheußer, Sandra Schneemann and Dirk Sliwka and has been published as Muehlheusser et al. (2018) in the *Journal of Sports Economics*.

Qiu, 2012) exploiting the variation which arises from the fact that, in the course of the careers, some managers are active in several organizations or functions which allows to disentangle their contribution from other factors. However, this is a difficult endeavor as CEOs, for instance, typically stay at the top of a specific firm for longer time periods and work as CEOs only for a very small number of different firms (very often only one) in their lifetime – which limits the scope to measure their contribution to organizational success.

In this paper, we consider this issue in the context of professional sports which, apart from being of interest in its own right, has further advantages for the question at hand: (i) team performance is publicly observable on a weekly basis and (ii) managers move very frequently between teams – much more frequently than managers in firms. And observing the same manager in different organizations thus using different sets of resources and working with different people is crucial to measure a manager's contribution to overall success. We use this information to estimate the impact of managers on team success, thereby also addressing the practical debate on this issue. For instance, in a popular book in the context of English soccer, Kuper and Szymanski (2009) are rather skeptical about the importance of managers, arguing that "[i]n a typical soccer talk, the importance of managers is vastly overestimated." (p. 123). The aim of our paper is to address this issue by disentangling econometrically the impact of individual managers from the overall strength of their respective team.

From a methodological point of view, we thereby follow the approach applied by Abowd, Kramarz, and Margolis (1999) (who use wages of employees working for different employers) and Bertrand and Schoar (2003) (who study CEOs working for different firms) and evaluate the impact of individual managers by estimating OLS regressions that include both team and manager fixed effects using data from the last 21 seasons of the *Bundesliga*, Germany's major soccer league. We then investigate the obtained manager fixed effects further and our results point to considerable heterogeneity: For instance, teams employing a manager at the 80% ability percentile gain on average 0.30 points per game more than those employing a manager at the 20% ability percentile. This corresponds to a difference of 18% of the average number of points won per game. We also conduct a cross validation exercise by estimating manager fixed effects using the data only up to a certain season and then investigate whether these fixed effects are useful to predict future performance. We find that this indeed is the case: these measures of managerial ability have a substantial predictive power for future perfor-

mance of the teams employing the respective manager. Furthermore, apart from team performance, we show that managers also have a significant effect on teams' playing style in terms of how offensively they play. We also find a negative correlation between the fixed effects for team performance and offensive style, supporting the view that successful managers are not necessarily the ones whose teams please crowds through their offensive play. Last, but not least, we investigate whether observable manager characteristics (in particular, whether they have been a former professional or even national team player and if so, on which position) also affects team performance. We find that if anything, the teams of managers who were former professionals perform worse on average than their less prominent counterparts.

Our paper contributes to the growing literature empirically analyzing the impact of managers on different economic measures, such as corporate behavior (Bertrand and Schoar, 2003), corporate tax avoidance (Dyreng, Hanlon, and Maydew, 2010), managerial compensation (Graham, Li, and Qiu, 2012), or disclosure choices (Bamber, Jiang, and Wang, 2010). In a prominent study, Bertrand and Schoar (2003) assess the impact of managers on firm performance, analyzing to what extent manager fixed effects can explain the observed heterogeneity in corporate behavior. They use a manager-firm matched panel data set that comprises different CEOs in different firms and focus only on those firms that have employed at least one *mover* manager, i.e. a manager who can be observed in at least two firms. The results show that manager fixed effects are important determinants in explaining corporate behavior. More recently, Lazear, Shaw, and Stanton (2015) study data from a large call center where supervisors move between teams (and team composition varies over time) which allows to disentangle the effect of different supervisors on performance. To the best of our knowledge, our paper is the first to apply this idea to the professional sports sector. Moreover, all managers in our study operate in the same industry, and this industry attracts a huge amount of public attention. As a result, most of these managers are very well-known to the interested public, so that the estimated individual fixed effects are of interest in their own right. Furthermore, we show that the estimated effects are useful to predict performance later in the managers' careers. Hence, our results can be helpful in identifying "under-valued" managers.

A further strand of literature has followed a different methodological route in order to measure managerial quality in professional sports: In a first step, a (stochastic) efficiency frontier is estimated for each team, and then in a second step, the quality of

a manager is assessed in terms of the team's proximity to this frontier during his term (see e.g., Carmichael and Thomas, 1995; Fizel and D'Itry, 1997; Dawson, Dobson, and Gerrard, 2000a,b; Dawson and Dobson, 2002; Kahane, 2005; Hofler and Payne, 2006). Frick and Simmons (2008) also use stochastic frontier analysis to show (also for the Bundesliga) that relative coach salaries have a significant impact on team efficiency.

The remainder of the paper is structured as follows: We first describe the data and the empirical framework in Section 2.2. In Section 2.3 we present our results with respect to the estimated manager fixed effects and the resulting heterogeneity of managers. Section 2.4 provides robustness checks along two dimensions: Firstly, we cross-validate our results by estimating first manager and team fixed effects for a restricted sample, and then use these estimates to predict team performance for the remaining seasons in our data set (Section 2.4.1). Secondly, we relax the assumption that all team-specific information is captured by a (time-invariant) team fixed effect, and consider (relative) team budgets as additional (time-variant) covariates (Section 2.4.2). In Section 2.5 we analyze the impact of managers on the offensive style of their teams. Section 2.6 investigates the impact of managers' background as professional players on team performance. Finally, Section 2.7 discusses possible caveats of our framework and concludes.

2.2 Empirical Framework

2.2.1 Data

The German Bundesliga – one of the strongest and economically most viable soccer leagues in the world – consists of 18 teams, and in each season, each team plays twice against each other team (one home match for each team), resulting in two half-seasons with 17 match days each. In each match, a winning (losing) team is awarded 3 (0) points, a draw results in 1 point for each team, and teams are ranked according to their accumulated points.[2] Our data set contains all Bundesliga matches played in the 21 seasons from 1993/94 until 2013/14 (9 matches played on each of 714 match days

[2]When several teams have accumulated the same number of points, the goal difference is used as the tie-breaking rule. In the first two season covered 1993/94 and 1994/95 the Bundesliga still applied a "two-point rule" where the winner of a game was awarded two points instead of three. We converted the data from these two seasons to the three-point rule.

leading to a total of 6426 matches).[3]

In our analysis, the unit of observation is the performance of a manager-team pair during a half-season (that is, match days 1 through 17 and 18 through 34, respectively). Therefore our dependent variable (*Points*) is the average number of points per game gained in the course of a half-season. Considering half-seasons has the advantage that a team faces each other team exactly once during that time, so that distortions due to different sets of opponents are reduced.

Throughout we refer to a *spell* as a non-interrupted relationship between a manager-team pair.[4] To be considered in the subsequent analysis, we require that a spell must last for at least 17 consecutive matches in the Bundesliga, and throughout the paper we refer to this as the *Footprint* condition (F).[5] This condition excludes observations from managers who are responsible for a team only for a small number of games.[6] While such short-term managers might have an impact on the team's short-term performance, they are unlikely to "leave a footprint". Out of the 176 managers in our data set, 116 remain after condition F is applied. The 60 managers and corresponding 109 spells which do not satisfy condition F are excluded from the further analysis. On average these spells lasted for a mere 6 matches only (see Appendix 2.B for more details).

Spells satisfying condition F often stretch over several half-seasons (thereby leading to multiple observations for our dependent variable), but the time interval of a spell does typically not divide evenly into half-seasons. The reason is that managers are frequently hired and dismissed within (half-) seasons.[7] In these cases, we consider the

[3]A large part of the data was kindly provided by deltatre AG, but it is also publicly available, e.g., from the website www.weltfussball.de.

[4]In a small number of cases, the same manager-team pair has multiple spells, that is, a team has hired the same manager again after several years, e.g., Ottmar Hitzfeld (Bayern Munich) or Felix Magath (Wolfsburg). We count each of such periods as separate spells.

[5]In a similar vein, Bertrand and Schoar (2003) require at least three joint years for a manager-firm pair to considered in the analysis. We have chosen 17 matches to limit the scope of distortions due to the strength of the opponent teams.

[6]For instance, there are interim coaches who are hired only for a small number of matches after a coach has been fired and before a permanent successor is found. In our sample, the average spell of such interim managers lasts for 2.35 matches only. But there are also some managers who are dismissed because of weak performance after being in office only for a small number of matches. Note that condition F gives rise to the possibility that teams feature an uneven number of half-season observations.

[7]Within-season dismissals are a very typical feature in European professional sports. On average, about 35-40% of the teams dismiss their manager within a given season at least once (see e.g. Muehlheusser, Schneemann, and Sliwka, 2016; De Paola and Scoppa, 2012; Tena and Forrest, 2007; Audas, Dobson, and Goddard, 2002). In the 21 seasons of our sample, we observe in total 192 such within-season dismissals.

performance in all half-seasons of the spell, weighted with the number of matches in the respective half-season.[8]

2.2.2 Identification of Manager Fixed Effects

We consider the following empirical model to explain the performance of team i under manager k in half season t

$$Points_{itk} = \gamma_i + \lambda_k + \alpha_t + \epsilon_{itk}, \tag{2.1}$$

where the dependent variable measures the average number of points per game won by team i during the half-season $t = 1, ...42$.

We start by applying a parsimonious approach and include only fixed effects for teams (γ_i), managers (λ_k), and half seasons (α_t) as explanatory variables. In a later robustness check, we also capture time-variant variation at the team level by including a proxy for their relative budgets in a given season (see Section 2.4.2). However, our preferred approach does not include budgets as a team's budget will also depend on recent performance and thus will typically be influenced by the current manager.[9] Obviously, γ_i and λ_k cannot be identified separately when the respective teams and managers are only jointly observed (that is, team i is only observed with manager k, and manager k is only observed with team i) since both variables are perfectly collinear in this case. Hence, to identify the different fixed effects, (at least some) managers and teams must be observed with multiple partners (see e.g., Abowd, Kramarz, and Margolis, 1999; Bertrand and Schoar, 2003).

In the context of European professional soccer, the rate of manager turnover is quite high. One reason is the high frequency of within-season managerial change as discussed above, but replacing managers between seasons is also quite common.[10] As a result, our

[8]For example, when a manager is hired at match day 5, and fired after match day 30 of the same season, this spell satisfies condition F, and there are two observations for this manager-team pair (one for the first half-season encompassing match days 5 to 17 and one for the second with match days 18 to 30, respectively). To take into account that the spell covers none of these two half-seasons in full, the average points won in each half-season are weighed using analytic weights which are inversely proportional to the variance of an observation (Stata command *aweights*).

[9]For instance, the top 5 teams at the end of a season are allowed to participate in the UEFA competitions *Champions League* or *Europe League* in the following season, both of which are financially very attractive.

[10]In the 21 seasons in our data set, in addition to the 192 within-season dismissals, there are 59 cases of managerial change between seasons.

data contains a large number of managers which are observed with many different teams (up to seven), and many teams which are observed under many different managers (up to 13) which creates a large amount of variation in observed manager-team matches. From a methodological point of view, this renders this industry particularly suitable for the identification of manager fixed effect .

Throughout, we distinguish between two types of managers: *movers* and *non-movers*. We refer to a manager as a (non-)mover when he is observed with at least two different (only one) team(s). Out of the 116 managers satisfying the footprint condition F, 44 (38%) managers are movers, while 72 (62%) are non-movers. As already explained, for all teams employing only non-mover managers, it is not possible to disentangle team and manager fixed effects, and therefore to identify a separate manager fixed effect. In contrast, for all teams observed with at least one mover manager, manager fixed effects can be estimated also for the non-mover managers. In line with Bertrand and Schoar (2003), we require that teams are observed with at least one mover, and refer to this as the mover-team (MT) condition. This condition is satisfied by 29 out of the 37 teams in our data set. The remaining 8 teams are excluded from the analysis.[11] The same is true for the 13 managers (none of them eliminated by condition F, all non-movers) who have been employed by these teams, leading to 13 excluded spells in addition to those already excluded due to condition F as explained above.[12] Our final data set covers 103 managers (44 movers, and 59 non-movers), 29 teams, 206 spells, and 764 observations for the dependent variable *Points*.

Table 2.1 gives an overview of all 103 managers in our final sample. As can be seen from the table, more than 80% of the 44 movers in our sample are either observed with two or three different teams. But we also observe managers who have worked for many more teams (up to seven as in the case of Felix Magath, for instance).

Moreover, Table 2.2 shows descriptive information for the 29 teams in our final data set, which illustrates again the frequency of managerial changes: For example, almost 60% of these teams have employed at least five (non-interim) managers. And 20% of the teams have even had at least ten managers during the last 21 seasons. Finally, Figure 2.1 and Table 2.3 give further descriptive information concerning the dependent

[11]Typically, these teams are small and enter the Bundesliga occasionally by promotion, and are relegated to the second division again after a small number of seasons. See Table 2.17 in Appendix 2.C for more information on these teams and their managers.

[12]Note that we first apply condition F and then condition MT, thus excluding those (three) managers who did work for two different teams, but where one of the spells is eliminated by condition F, see Table 2.17 in Appendix 2.C.

	Manager	No. of teams	No. of obs		Manager	No. of teams	No. of obs
1	Advocaat, Dick	1	2	53	Löw, Joachim	1	4
2	Augenthaler, Klaus	3	13	54	Magath, Felix	7	34
3	Babbel, Markus	3	6	55	Marwijk, Bert van	1	5
4	Berger, Jörg	3	11	56	Maslo, Uli	1	4
5	Bommer, Rudi	1	2	57	McClaren, Steve	1	2
6	Bongartz, Hannes	3	6	58	Meyer, Hans	3	13
7	Bonhof, Rainer	1	2	59	Middendorp, Ernst	1	6
8	Brehme, Andreas	1	5	60	Mos, Aad de	1	1
9	Daum, Christoph	3	13	61	Möhlmann, Benno	2	7
10	Demuth, Dietmar	1	2	62	Neubarth, Frank	1	2
11	Doll, Thomas	2	9	63	Neururer, Peter	3	13
12	Dutt, Robin	3	8	64	Oenning, Michael	1	1
13	Dörner, Hans-Jürgen	1	4	65	Olsen, Morten	1	5
14	Engels, Stephan	1	2	66	Pacult, Peter	1	4
15	Fach, Holger	2	4	67	Pagelsdorf, Frank	2	15
16	Favre, Lucien	2	12	68	Pezzaiuoli, Marco	1	1
17	Fink, Thorsten	1	5	69	Rangnick, Ralf	4	17
18	Finke, Volker	1	20	70	Rapolder, Uwe	2	3
19	Fringer, Rolf	1	2	71	Rausch, Friedel	2	8
20	Frontzeck, Michael	3	9	72	Rehhagel, Otto	3	13
21	Funkel, Friedhelm	6	27	73	Reimann, Willi	2	4
22	Gaal, Louis van	1	4	74	Ribbeck, Erich	2	5
23	Gerets, Eric	2	7	75	Rutten, Fred	1	2
24	Gerland, Hermann	1	2	76	Röber, Jürgen	3	16
25	Gisdol, Markus	1	3	77	Sammer, Matthias	2	10
26	Gross, Christian	1	3	78	Scala, Nevio	1	2
27	Guardiola, Pep	1	2	79	Schaaf, Thomas	1	29
28	Götz, Falko	2	9	80	Schaefer, Frank	1	2
29	Hecking, Dieter	3	16	81	Schlünz, Juri	1	3
30	Heesen, Thomas von	1	4	82	Schneider, Thomas	1	2
31	Herrlich, Heiko	1	2	83	Sidka, Wolfgang	1	3
32	Heynckes, Jupp	5	15	84	Skibbe, Michael	3	14
33	Hitzfeld, Ottmar	2	23	85	Slomka, Mirko	2	13
34	Hyypiä, Sami	1	2	86	Solbakken, Stale	1	2
35	Jara, Kurt	2	8	87	Soldo, Zvonimir	1	3
36	Jol, Martin	1	2	88	Sorg, Marcus	1	1
37	Keller, Jens	1	3	89	Stanislawski, Holger	2	4
38	Klinsmann, Jürgen	1	2	90	Stepanovic, Dragoslav	1	4
39	Klopp, Jürgen	2	18	91	Stevens, Huub	3	21
40	Koller, Marcel	2	9	92	Streich, Christian	1	5
41	Korkut, Tayfun	1	1	93	Toppmöller, Klaus	4	17
42	Krauss, Bernd	1	7	94	Trapattoni, Giovanni	2	8
43	Kurz, Marco	1	4	95	Tuchel, Thomas	1	10
44	Köppel, Horst	1	3	96	Veh, Armin	5	18
45	Körbel, Karl-Heinz	1	3	97	Verbeek, Gertjan	1	2
46	Köstner, Lorenz-Günther	2	6	98	Vogts, Berti	1	2
47	Labbadia, Bruno	3	11	99	Weinzierl, Markus	1	4
48	Latour, Hanspeter	1	1	100	Wiesinger, Michael	1	2
49	Lewandowski, Sascha	1	3	101	Wolf, Wolfgang	3	17
50	Lienen, Ewald	5	17	102	Zachhuber, Andreas	1	4
51	Lorant, Werner	1	15	103	Zumdick, Ralf	1	2
52	Luhukay, Jos	3	6		Total	∅2.62	∑764

Only managers after application of conditions F and MT.
Unit of observation: Half-season
Time period: The 21 seasons from 1993/94 - 2013/14.

Table 2.1: The Bundesliga Managers in the Final Data Set

	Team	No. of managers	No. of movers	No. of non-movers	No. of obs
1	1860 Munich	3	1	2	22
2	Aachen	1	1	0	2
3	Augsburg	2	1	1	6
4	Bayern Munich	9	6	3	43
5	Bielefeld	6	3	3	19
6	Bochum	5	3	2	25
7	Bremen	7	3	4	45
8	Cologne	13	7	6	32
9	Dortmund	7	5	2	42
10	Duisburg	4	3	1	15
11	Frankfurt	9	8	1	30
12	Freiburg	4	1	3	30
13	Hamburg	11	9	2	46
14	Hannover	6	5	1	27
15	Hertha Berlin	8	8	0	30
16	Hoffenheim	5	3	2	13
17	Kaiserslautern	7	5	2	31
18	Leverkusen	12	8	4	47
19	Mainz	2	1	1	16
20	Mönchengladbach	13	9	4	44
21	Nürnberg	7	4	3	25
22	Rostock	7	5	2	26
23	Schalke	9	6	3	44
24	St. Pauli	3	1	2	8
25	Stuttgart	13	9	4	48
26	Uerdingen	1	1	0	4
27	Unterhaching	1	1	0	4
28	Wattenscheid	1	1	0	2
29	Wolfsburg	10	9	1	38
	Total	∅6.41	∅4.38	∅2.03	∑764

Only teams after application of conditions F and MT.
Unit of observation: Half-season.
Time period: The 21 seasons from 1993/94 - 2013/14.

Table 2.2: The Bundesliga Teams in the Final Data Set

variable *Points* and the spells in our final data. Figure 2.1 shows the distribution of team performance measured by the average number of points per game in the relevant half-season.

Note that manager-team pairs win on average 1.41 points per game. On average, a spell lasts for slightly less than 60 matches, and the 103 managers in the final data set are observed with about two spells on average, but this number can be as large as eight.

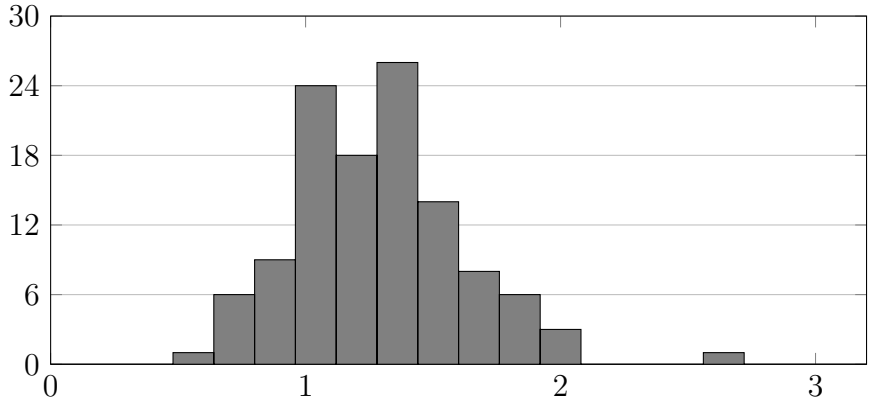


Figure 2.1: Histogram of Dependent Variable *Points* (all managers, weighted)

Variable		**Obs.**	**Mean**	**Std. Dev.**	**Min**	**Max**
Points	all managers	764	1.410	0.452	0	3
	only movers	533	1.435	0.452	0	3
Matches per spell	all managers	206	58.903	53.483	17	479
	only movers	133	59.872	40.639	17	204
Half-seasons per spell	all managers	206	3.93	3.154	1	29
	only movers	133	4.008	2.404	1	12
Number of spells	all managers	103	1.981	1.350	1	8
	only movers	44	3.159	1.293	2	8

Only teams after application of conditions F and MT.
Points refer to the average number of points per game per half-season, weighted by the number of games of the respective manager-team pair in a half-season.

Table 2.3: Descriptive Statistics

2.3 Empirical Analysis

We now investigate whether the identity of the managers indeed has a significant impact on the team's performance. In a first step, we follow Bertrand and Schoar (2003) and start with analyzing the joint effect of managers and teams on the outcome variable and whether and to what extent the explanatory power of the regressions increases once manager fixed effects are included (Section 2.3.1). In a next step, we analyze the coefficients of the individual manager fixed effects in more detail (Section 2.3.2).

2.3.1 The (Joint) Impact of Managers on Team Performance

Table 2.4 shows the results of three different models which differ with respect to the set of independent variables used. Model 1 contains only half-season fixed effects, Model 2 contains both half-season and team fixed effects, while in Model 3 manager fixed effects are included in addition. As can be seen, the explanatory power sharply increases once team fixed effects are included (Model 2). When comparing Models 2 and 3, the inclusion of manager fixed effects leads to an increase of the R^2 by 11.4 percentage points (or 32.1%), and the adjusted R^2 increases by 2.5 percentage points (or 8.6%). Moreover, the F-Test for the joint significance of the manager fixed effects is highly significant ($p < 0.01$).

	Model 1	**Model 2**	**Model 3**
Half-Season FE	Yes	Yes	Yes
Team FE	No	Yes	Yes
Manager FE	No	No	Yes
N	764	764	764
R^2	0.007	0.355	0.469
adj. R^2	-0.049	0.291	0.316
F-test Manager FE			8.633
p-value			0.000

Dependent variable: Average points per game per half-season. Clustered on half-season level, weighted with the number of matches per manager-team pair in half-season

Table 2.4: The Joint Impact of Managers on Team Performance

2.3.2 Estimation of Manager Fixed Effects: Comparing the Performance Contributions of Managers

We now analyze the individual manager fixed effects in more detail. As explained above and analogous to the argument by Abowd, Kramarz, and Margolis (1999), manager fixed effects can be estimated not only for the 44 movers in our sample, but also for the 59 non-movers (such as Pep Guardiola, Luis van Gaal) as long as their only team is also observed with at least one mover, i.e., satisfies condition MT. Note however, that the identification of the fixed effect of non-movers must come from disentangling it from the fixed effect of their (only) team. This might be problematic if this team is

only observed with a few other managers. In contrast, for movers we can exploit the larger variation since several teams and their respective team fixed effects are involved. Consequently, we first focus our discussion on the fixed effects for the mover managers.

Table 2.5 presents the estimated fixed effects for the 44 mover managers in our final sample, ranked by the size of the coefficient which, for each manager, measures his deviation from a reference category. In general, which of the coefficients for the fixed effects are statistically significant depends on the choice of the reference category, and in Table 2.5 the median manager (Bruno Labbadia) is chosen. In this case, the (statistically significant) coefficient for Jürgen Klopp (rank 1 on left part of Table 2.5) implies that *ceteris paribus* his teams have won on average 0.46 points per match more than a team coached by a manager of median ability.[13] This performance increase corresponds to 33% of the 1.41 points awarded on average per game during a half-season (see Table 2.3), and hence would on average lead to an additional $17 \cdot 0.46 = 7.82$ points per half-season for the respective team.[14] For the season 2012/13, for example, this amount of additional points won would have pushed a team from rank 13 to rank 4, which would have allowed the team to participate in the highly prestigious and financially attractive UEFA Champions League.

For the sake of comparison, the right part of Table 2.5 ranks the managers simply with respect to the average number of points won with their respective teams in the considered spells. As is evident, this procedure favors those managers who have worked for the big teams such as Bayern Munich, Borussia Dortmund or Schalke 04, which have more financial resources to hire the best players. Comparing these two rankings leads to remarkable differences: For example, Giovanni Trapattoni is ranked second using this simple procedure, while our empirical analysis suggests that his quality is below average (rank 36). On the other hand, we find a strongly positive value for Dieter Hecking (rank 4), who has less experience with top teams, and hence is only listed at position 21 in the ranking purely based on points won. Overall, the correlation between the two measures of ability is not too high ($\rho = 0.5$).

[13] Of course, each individual fixed effect is estimated with some noise. When comparing the estimates for the individual fixed effects to the median manager only the effect of Jürgen Klopp is statistically different from the median manager. When moving the reference category downwards the number of significant coefficients at the top increases. For example, when compared to a manager at the lower 25% percentile, the coefficients of the top four managers are significant. Furthermore, a large number of pairwise comparisons of managers also exhibit statistically significant differences.

[14] The top rank for Klopp (currently manager of the Premier League team FC Liverpool) seems reasonable, as he was very successful with his first Bundesliga team (the underdog Mainz), and has then led Dortmund to two national championships and to the final of the UEFA Champions League.

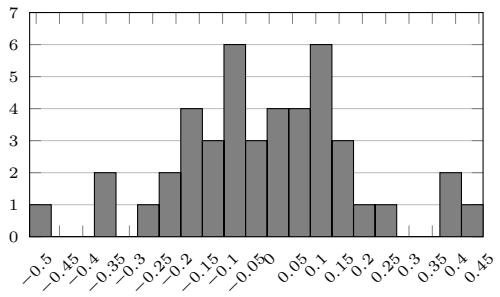
Estimated Fixed Effect			*Average Points Won Per Match*°		
Rank	**Manager**	**Coeff.**	**Rank**	**Manager**	**Avg. Points**
1	Klopp, Jürgen	0.459**	1	Hitzfeld, Ottmar	2.008
2	Favre, Lucien	0.411	2	Trapattoni, Giovanni	1.820
3	Slomka, Mirko	0.378	3	Heynckes, Jupp	1.788
4	Hecking, Dieter	0.264	4	Sammer, Matthias	1.759
5	Rehhagel, Otto	0.202	5	Rehhagel, Otto	1.729
6	Sammer, Matthias	0.164	6	Klopp, Jürgen	1.712
7	Götz, Falko	0.148	7	Daum, Christoph	1.687
8	Heynckes, Jupp	0.146	8	Magath, Felix	1.644
9	Röber, Jürgen	0.127	9	Slomka, Mirko	1.556
10	Magath, Felix	0.121	10	Favre, Lucien	1.545
11	Rangnick, Ralf	0.114	11	Stevens, Huub	1.530
12	Meyer, Hans	0.112	12	Doll, Thomas	1.508
13	Neururer, Peter	0.098	13	Röber, Jürgen	1.496
14	Hitzfeld, Ottmar	0.097	14	Rausch, Friedel	1.481
15	Daum, Christoph	0.078	15	Skibbe, Michael	1.473
16	Veh, Armin	0.073	16	Labbadia, Bruno	1.439
17	Stevens, Huub	0.067	17	Ribbeck, Erich	1.431
18	Lienen, Ewald	0.053	18	Rangnick, Ralf	1.425
19	Köstner, Lorenz-Günther	0.040	19	Jara, Kurt	1.384
20	Babbel, Markus	0.035	20	Veh, Armin	1.367
21	Rausch, Friedel	0.018	21	Hecking, Dieter	1.362
22	Labbadia, Bruno	0 (Ref)	22	Toppmöller, Klaus	1.360
23	Bongartz, Hannes	-0.009	23	Götz, Falko	1.356
24	Doll, Thomas	-0.014	24	Babbel, Markus	1.321
25	Stanislawski, Holger	-0.042	25	Augenthaler, Klaus	1.317
26	Pagelsdorf, Frank	-0.051	26	Pagelsdorf, Frank	1.303
27	Funkel, Friedhelm	-0.058	27	Berger, Jörg	1.299
28	Skibbe, Michael	-0.066	28	Gerets, Eric	1.289
29	Toppmöller, Klaus	-0.073	29	Neururer, Peter	1.287
30	Wolf, Wolfgang	-0.079	30	Wolf, Wolfgang	1.284
31	Jara, Kurt	-0.084	31	Meyer, Hans	1.240
32	Koller, Marcel	-0.119	32	Dutt, Robin	1.215
33	Augenthaler, Klaus	-0.127	33	Lienen, Ewald	1.203
34	Fach, Holger	-0.136	34	Möhlmann, Benno	1.164
35	Gerets, Eric	-0.148	35	Köstner, Lorenz-Günther	1.149
36	Trapattoni, Giovanni	-0.170	36	Fach, Holger	1.127
37	Dutt, Robin	-0.171	37	Bongartz, Hannes	1.113
38	Berger, Jörg	-0.174	38	Funkel, Friedhelm	1.087
39	Rapolder, Uwe	-0.217	39	Koller, Marcel	1.053
40	Frontzeck, Michael	-0.225	40	Rapolder, Uwe	1.041
41	Luhukay, Jos	-0.240	41	Luhukay, Jos	1.022
42	Möhlmann, Benno	-0.333	42	Reimann, Willi	1.017
43	Reimann, Willi	-0.342	43	Stanislawski, Holger	0.981
44	Ribbeck, Erich	-0.514*	44	Frontzeck, Michael	0.942

° *Average Points Won Per Match* refers to the average number of points gained in spells satisfying conditions F and MT.
Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

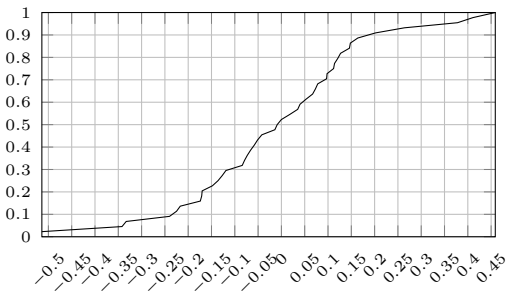
Table 2.5: Ranking of Mover Managers. Fixed Effects Versus Average Points Won

Figure 2.2 shows the distribution of fixed effects as reported in the left column of Table 2.5. The histogram depicted in panel a) suggests that Bundesliga managers are quite heterogeneous with respect to their abilities, giving rise to a difference of up to 1 point per match between the managers at the top and the bottom of the ranking. Moreover, as can be seen from the cumulative distribution depicted in panel

b), managers around the 80% ability percentile (Jupp Heynckes or Jürgen Röber) gain on average 0.30 points per game more than those at the 20% percentile (Giovanni Trapattoni or Eric Gerets). This corresponds to a difference of 18% of the average number of 1.41 points won per game (see Table 2.3). In general, many (but not all) of these fixed effects are statistically different from each other in a pairwise comparison.



(a) Frequency of Coefficients



(b) Cumulative Distribution of the Coefficients

Figure 2.2: Frequency and Distribution of Manager Fixed Effects

In summary, our results are in line with previous results from other industries such as Bertrand and Schoar (2003) and Graham et al. (2012) who find that executives are an important factor determining organizational performance. Moreover, the degree of heterogeneity between individuals with respect to this ability seems remarkable, in particular as we take into account only the top segment of the labor market for football managers, i.e. our sample of managers already contains a selected group of the most able ones as each single year, only 24 new managers complete a mandatory training program for head coaches organized by the German Football Association (DFB). All in all, our results do not support the argument that such mandatory training programs would make the population of Bundesliga managers quite homogenous (see e.g., Breuer and Singer, 1996).

Furthermore, our results indicate that the sporting and financial implications of decisions concerning the hiring of managers can be substantial: for example, 33 out of the 63 teams which were either directly relegated to the second division or had to play an additional relegation round to avoid relegation, would have been saved from relegation respectively the relegation round if they had won 5 additional points in the course of the season.[15] According to our analysis, this corresponds to the difference

[15]From 1993/94 to 2007/08 the last three teams were relegated directly to the second division. As

between a manager at the 20%- and 50%-percentile.

Table 2.13 in Appendix 2.A reports also the fixed effects estimates for non-mover managers (in grey), i.e. those that we observe only with a single team (and where this team satisfies condition MT). As argued by Abowd et al. (1999), these fixed effects are also identified, but the estimates rely on a precise estimation of the respective team fixed effects. This seems a strong requirement for those teams who are observed with only a few other managers (mostly non-movers themselves). Given the few sources of variation and the small number of observations in such cases, the disentangling of the two fixed effects does not always seem convincing and leads to implausible results. Two cases in point here are Thomas Tuchel (Mainz) and Peter Pacult (1860 Munich) whose manager fixed effects seem excessively high (rank 1 and 3, respectively, in Table 2.13) in the light of their accomplishments.[16] In contrast, as can be seen from Table 2.14 (also in Appendix 2.A), the estimated team fixed effects for their teams Mainz and 1860 Munich (left column) appear to be excessively low (rank 29 and 26, respectively) compared to the performance of these teams measured in terms of points won (rank 11 and 13, respectively, right column). Hence, the estimates for such non-mover managers that were employed by teams that did not employ many movers have to be interpreted with caution.

2.4 Robustness

In this section, we check the robustness of our results. First, we cross-validate our estimates of the managers' abilities, by analyzing whether the estimated fixed effects are able to predict future performance (Section 2.4.1). Second, we also consider time-variant proxies for the teams' budgets in the regressions (Section 2.4.2).[17]

2.4.1 Cross Validation: Predicting Future Performance

As a first robustness check, we check whether our estimates of manager fixed effects are useful in predicting future team performance. In particular, we ask the following

of season 2008/09, the team ranked third to last and the team ranked third in the second division compete in two extra matches for the final Bundesliga slot for the next season.

[16] As of season 2015/16, Thomas Tuchel is the manager of Borussia Dortmund and hence by now a mover, but this season is not contained in our data set.

[17] Instead of half-seasons, we have also used full seasons as the time horizon for which team performance is measured, and the results are almost identical.

question: if we use our approach to obtain estimates of managers' abilities using all the data up to a certain date t which corresponds to the beginning of a season – to what extent do these estimates help to predict performance of the teams employing these managers in the season that follows? In order to do so, we proceed in several steps: First, starting with the beginning of season 2004/05 (which corresponds to half-season 23 in our data set) we estimate manager and team fixed effects restricting the data set to all outcomes prior to the season we want to predict. Hence, for each manager k and team i and date $t \in \{23, 25, 27, ..41\}$, we obtain a moving time series of fixed effects $\hat{\lambda}_k^{t-1}$ and $\hat{\gamma}_i^{t-1}$ up to date $t-1$. We then run a simple OLS regression with the average number of points obtained by a team in a half-season $t \geq 23$ as the dependent variable and the fixed effects for managers and teams (evaluated at the end of the previous full season) as independent variables.

	Model P1	**Model P2**		**Model P3**	**Model P4**
Team FE	0.660***	0.782***	*Team Points*	0.962***	0.933***
	(0.0983)	(0.100)		(0.103)	(0.119)
Manager FE		0.354***	*Manager Points*		0.0554
		(0.0891)			(0.114)
Constant	1.354***	1.364***	Constant	0.0861	0.0460
	(0.0301)	(0.0294)		(0.148)	(0.169)
Obs.	262	262	Obs.	262	262
R^2	0.148	0.197	R^2	0.250	0.251
adj. R^2	0.144	0.191	adj. R^2	0.247	0.245

Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01. Dependent variable: Average points per game per half-season for the seasons 2004/05 to 2013/14.
In models P1 and P2 the fixed effects for teams and managers are the estimates obtained from season 1993/94 up to the end of the full season preceding the half-season under consideration. Similarly, in model P3 and P4, the average points won by teams and managers are obtained up to the end of the full season preceding the half-season under consideration.

Table 2.6: Using Fixed Effects to Predict Future Performance

The key question is whether these estimated manager fixed effects have predictive power for the team's performance in the subsequent year. Table 2.6 shows the regression results, where model P1 includes only our estimates for team strength while in model P2, we add our estimates for managers' abilities. We find indeed that both our measures of team strength and managers' abilities are helpful in predicting subsequent performance. Including our proxies for the managers' abilities raises the adjusted R^2 by 33% from 0.144 to 0.191 and the coefficient of managerial ability is significant at the 1% level. Following Angrist and Pischke (2008) in interpreting regressions as approximations to the conditional expectation function, we thus conclude that our estimates

of managerial ability indeed substantially affect conditional expectations and are thus valuable predictors of future performance.

We also compare these predictive regressions to an alternative way of predicting team performance on the basis of the average number of points won by a team (with all its previous managers) and its current manager (with all his previous teams) in the past. While the average number of points won by teams in the past is indeed a valuable predictor for future performance (Model P3), the average number of points won by its manager in the past has no additional explanatory power at all (Model P4). Hence, if we want to disentangle the contribution of a manager from the underlying strength of a team to predict the team's performance, our "purged" measure of ability is more valuable than measures which are simply based on past performance outcomes.[18] Last, but not least, it is interesting to note that the slope of the manager rank (0.354) attains a value of about 45% of the slope of the team strength (0.782). Given that it seems much easier to replace a manager with a better one than to replace a whole team, picking a better manager indeed seems to be a key lever to increase team performance.

2.4.2 Testing the Impact of Further Time-Variant Variables

The model specification used in Section 2.3 was very parsimonious in the sense that it included only various (time-independent) fixed effects, but not time-variant variables such as a team's wage bill or its (relative) budget, both of which have been shown to also be crucial determinants of team performance (see e.g., Szymanski and Smith, 1997; Hall et al., 2002; Kahane, 2005). As explained above, the main reason for excluding such variables in our basic model was our concern that in the context of determining the value of managers, a team's budget in a given season will also depend on its performance in previous seasons, and hence be influenced by its manager (in case he was already in charge of the team then), so that it is not an independent control variable. For example, a top-5 team in season t is allowed to compete in the UEFA competitions (Champions League and EURO League) in season $t+1$, which typically comes with a considerable increase in revenues.[19]

[18]The results are robust when replacing the estimated fixed effects of managers and teams as estimated up to date $t-1$ with their respective percentile scores (i.e. the manager with the highest fixed effect at date $t-1$ has a percentile score of 1 and the median manager a percentile score of 0.5).

[19]For instance, according to the publicly available Deloitte report "Commercial breaks. Football Money League", Bayern Munich received 44.6 million Euro from the UEFA alone for its Champions League participation in the season 2013/14 (excluding additional gate revenues of approximately 22 million euro), while the average budget of a Bundesliga team was 41.5 million euro.

But of course the drawback of this parsimonious approach is that idiosyncratic variations in a team's financial strength over the time horizon considered are not accounted for. Hence, managers who are hired in a phase where a team has less financial resources may be disadvantaged and those that are hired in a phase where the team has more resources may benefit as variation in financial strength may be captured by the estimated manager fixed effects. To check the robustness of our results, we now also include a proxy for the (relative) budgets of teams in a given season as a time-variant variable.[20] In contrast to the English Premier League where many teams are publicly listed companies, this is not the case for the Bundesliga. Hence, they are not obliged to publish any hard financial information such as budgets or even wage bills. As a consequence, when including (relative) team budgets in the regressions, we must rely on estimates compiled by public sources such as newspapers and specific reports from banks and consulting firms. These are based on core parts of a team's income such as TV revenues, revenues from participation in the UEFA Leagues, ticket sales, and sponsoring which are in large parts publicly available. Hence, while being noisy they do reflect the relative financial strengths of the teams in a given season.[21]

From this information, we have constructed a new variable (*Budget*) which measures a team's relative budget in a given season as the ratio between its absolute budget and the average budget of all teams in that season. Table 2.7 provides some descriptive statistics on this new variable. As can be seen, Bundesliga teams are quite heterogeneous with respect to their financial possibilities, and some teams such as Bayern Munich (Freiburg) have consistently high (low) budgets and even the minimum (maximum) value is above (below) average. Moreover, the fact that several teams such as Wolfsburg, Leverkusen or Mönchengladbach exhibit minimum values smaller than one and maximum values larger than one suggests that their relative strength also has changed over time.

In Table 2.8, we report again two model specifications, where a team's relative budget

[20]We also investigated the role of further time-variant variables such as a manager's age and tenure but when including them as additional control variables in the regressions, the respective coefficients are virtually zero and statistically insignificant.

[21]In the *Bundesanzeiger*, Germany's official federal gazette regarding all public financial and legal statements made by firms, we found some 25 data points on wage bills (entire staff), and the correlation between these official numbers and our estimates is 0.979. Alternatively, one could use the market value of team rosters based on the estimates on the web page www.transfermarkt.de. While this information is only available for a subset of seasons (from 2005/06 - 2013/14), the correlation with our team budget proxies is 0.87. We are grateful to an anonymous referee for suggesting this alternative measure.

		Relative budget					Relative budget		
	Team	**Min.**	**Max.**	**Av.**		**Team**	**Min.**	**Max.**	**Av.**
1	1860 Munich	0.69	1.08	0.91	16	Hoffenheim	0.72	1.11	0.87
2	Aachen	0.43	0.43	0.43	17	Kaiserlautern	0.39	1.55	0.89
3	Augsburg	0.41	0.46	0.43	18	Leverkusen	0.76	1.38	1.08
4	Bayern Munich	1.24	3.37	2.01	19	Mainz	0.4	0.73	0.57
5	Bielefeld	0.42	0.81	0.63	20	Mönchengladbach	0.69	1.33	0.91
6	Bochum	0.47	0.81	0.64	21	Nürnberg	0.33	1.33	0.66
7	Bremen	0.84	1.44	1.12	22	Rostock	0.53	0.96	0.71
8	Cologne	0.63	1.45	1.05	23	Schalke	1.03	2.21	1.44
9	Dortmund	0.8	1.64	1.23	24	St. Pauli	0.42	0.7	0.58
10	Duisburg	0.55	0.85	0.71	25	Stuttgart	0.92	1.41	1.15
11	Frankfurt	0.64	1.21	0.91	26	Uerdingen	0.41	0.59	0.5
12	Freiburg	0.37	0.75	0.58	27	Unterhaching	0.39	0.48	0.44
13	Hamburg	0.69	1.96	1.14	28	Wattenscheid	0.49	0.49	0.49
14	Hannover	0.62	0.88	0.75	29	Wolfsburg	0.59	1.85	1.22
15	Hertha Berlin	0.55	1.64	1.04					

Only teams after application of conditions F and MT. Sources: Estimates for the 21 seasons from 1993/94 - 2013/14 from the German daily newspapers *Die Welt* (1993/94 to 1998/1999 and 2002/2003 to 2008/2009) and *Rheinische Post* (2007/2008 to 2013/2014) and study "FC Euro AG" (1997/1998 to 2004/2005) published in 2004 by *KPMG* and *WGZ-Bank*.

Table 2.7: Summary Information for Relative Budgets of Bundesliga Teams

is used in addition to (Model 4) and instead of (Model 5) team fixed effects, respectively. For the sake of comparison, the left column reports again the respective result from the basic analysis without the relative budget proxies (see right column of Table 2.4). As can be seen, the manager fixed effects remain also jointly significant at very high significance levels when the budgets are included. Moreover, also the budgets alone have a significant impact, but the adjusted R^2 is higher when team fixed effects are included in addition. Overall, compared to the baseline specification of Model 3, Model 4 leads to a slight increase of the explanatory power, while it decreases under Model 5. This suggests that budget proxies and team fixed effects provide to some extent complementary information. For instance, budgets indeed capture time variation in financial strength, but team fixed effects rather the more stable properties of teams and their management.

We investigate next whether also our estimates for the individual manager fixed effects are robust when we include the budget proxies in addition to team fixed effects (Model 4). The resulting ranking of manager fixed effects is shown in the right column of Table 2.9. Again, for the sake of comparison, the left column repeats the ranking from the basic model (see left column of Table 2.5 above). As can be seen, the ranking of managers is not altered substantially: The ranks of the top managers are virtually unchanged, and also their coefficients are very similar. Overall, Spearman's rank correlation coefficient between the ranking with and without budget proxies is

	Model 3	**Model 4**	**Model 5**
Half-Season FE	Yes	Yes	Yes
Team FE	Yes	Yes	No
Manager FE	Yes	Yes	Yes
Budget	No	Yes	Yes
N	764	764	764
R^2	0.469	0.474	0.402
adj. R^2	0.316	0.321	0.263
F-test Manager FE	8.633	6.181	11.75
p-value	0.000	0.000	0.000
F-test Team FE	22.86	11.81	
p-value	0.000	0.000	

Dependent variable: Average points per game per half-season. Clustered on half-season level, weighted with the number of matches per manager-team pair in half-season.

Table 2.8: The Joint Impact of Managers on Team Performance With Team Budgets Included

	Model 3		**Model 4**			**Model 3**		**Model 4**	
Manager	**R.**	**Coeff.**	**R.**	**Coeff.**	**Manager**	**R.**	**Coeff.**	**R.**	**Coeff.**
Klopp, Jürgen	1	0.459	1	0.542	Bongartz, Hannes	23	-0.009	26	-0.049
Favre, Lucien	2	0.411	2	0.442	Doll, Thomas	24	-0.015	14	0.080
Slomka, Mirko	3	0.378	3	0.383	Stanislawski, Holger	25	-0.042	22	0 (Ref)
Hecking, Dieter	4	0.264	4	0.307	Pagelsdorf, Frank	26	-0.051	28	-0.073
Rehhagel, Otto	5	0.202	6	0.146	Funkel, Friedhelm	27	-0.058	25	-0.047
Sammer, Matthias	6	0.164	5	0.188	Skibbe, Michael	28	-0.066	27	-0.051
Götz, Falko	7	0.148	16	0.070	Toppmöller, Klaus	29	-0.073	32	-0.087
Heynckes, Jupp	8	0.146	15	0.073	Wolf, Wolfgang	30	-0.079	24	-0.016
Röber, Jürgen	9	0.127	8	0.115	Jara, Kurt	31	-0.084	33	-0.098
Magath, Felix	10	0.121	10	0.109	Koller, Marcel	32	-0.119	29	-0.08
Rangnick, Ralf	11	0.114	7	0.126	Augenthaler, Klaus	33	-0.127	31	-0.087
Meyer, Hans	12	0.112	8	0.115	Fach, Holger	34	-0.136	36	-0.143
Neururer, Peter	13	0.098	13	0.084	Gerets, Eric	35	-0.148	37	-0.148
Hitzfeld, Ottmar	14	0.097	12	0.099	Trapattoni, Giovanni	36	-0.17	34	-0.1
Daum, Christoph	15	0.078	20	0.025	Dutt, Robin	37	-0.171	35	-0.133
Veh, Armin	16	0.073	11	0.100	Berger, Jörg	38	-0.174	39	-0.189
Stevens, Huub	17	0.067	19	0.033	Rapolder, Uwe	39	-0.217	41	-0.23
Lienen, Ewald	18	0.053	17	0.067	Frontzeck, Michael	40	-0.225	40	-0.204
Köstner, Lorenz-Günther	19	0.040	21	0.023	Luhukay, Jos	41	-0.24	38	-0.16
Babbel, Markus	20	0.035	18	0.061	Möhlmann, Benno	42	-0.333	43	-0.326
Rausch, Friedel	21	0.018	30	-0.081	Reimann, Willi	43	-0.342	42	-0.274
Labbadia, Bruno	22	0 (Ref)	23	-0.005	Ribbeck, Erich	44	-0.514	44	-0.544

In Model 4, the coefficient of the variable *Budget* is 0.167^{**} ($p < 0.058$).

Table 2.9: Ranking of Fixed Effects of Mover Managers Without and With Team Budgets

$\rho = 0.97$, suggesting that our results are indeed robust in this respect. In contrast to the above-mentioned skepticism by Kuper and Szymanski (2009) concerning the contribution of managers in determining team performance on top of teams' financial power, our results suggests that there is indeed a role for managers (at least in the Bundesliga), even after controlling for the (time-variant) financial strength of teams.

2.5 Manager Fixed Effects and Team Style

Apart from team performance, managers might also have an impact on other team variables such as a team's playing style, in particular whether it is playing rather offensively or defensively.[22] Consequently, we can apply the same method as in the above in order to analyze to what extent the identity of the manager in office has predictive power to explain a team's playing style. To this end, we start by defining the following measure of "offensiveness" of team i under manager k in half-season t:

$$\text{Offensive}_{ikt} = \frac{\text{average goals scored per match}}{\text{average points won per match}} \tag{2.2}$$

Under this measure, a team is considered to play more offensively when it scores more goals for a given average number of points won.[23] Analogously to the analysis of team performance, we first investigate whether the manager fixed effects are jointly significant in determining the playing style of teams, and the results are reported in Table 2.10.

As before, the goodness of fit increases by a large amount when adding team - and manager fixed effects (comparing Models S1 and S2 versus Model S3), respectively. Moreover, the increase is particularly large when manager fixed effects are added, while the addition of team fixed effects alone has only a small impact. This suggests that the degree to which teams are playing offensively is strongly influenced by their current managers rather than "team DNA".[24]

[22]Further dimensions of interest would be how aggressively teams play (as for example measured by the number of yellow and red cards conceded), or their physical activity level in the pitch (as for example measured by the average number of kilometers which players run during a match). Unfortunately, our data set does not contain the respective information.

[23]For example, when a match ends in a $3:3$ tie, both teams would be considered to play more offensively than under a $0:0$ tie (both outcomes resulting in one point won for each team). Note that for the league table at the end of the season, the crucial variable is the number of points won, while the difference between the numbers of goals scored and goals conceded is used as a tie-breaking rule. Given the large number of 34 match days, however, ties of this type occur only very rarely.

[24]Again, managers can also be ranked with respect to their estimated fixed effects with respect to

	Model S1	**Model S2**	**Model S3**
Half-Season FE	Yes	Yes	Yes
Team FE	No	Yes	Yes
Manager FE	No	No	Yes
N	753	753	753
R^2	0.045	0.106	0.302
adj. R^2	-0.010	0.015	0.097
F-test Manager FE			14.53
p-value			0.000

Dependent variable: Offensive rating per game per half-season. Clustered on half-season level, weighted with the number of matches per manager-team pair in half-season.

Table 2.10: The Joint Impact of Managers on Team Style

In a next step, we can compare these manager fixed effects with those based on team performance (see Table 2.4 above). Interestingly, better managers (i.e. those with larger manager fixed effects in our performance regressions) are those who prefer their team to play defensively. Figure 2.3 depicts the manager fixed effects along these two dimensions, and it reveals a negative correlation between offensive style and performance ($\rho = -0.375$). At an anecdotal level, this is consistent with the frequently heard claim that a good offense is what pleases the audience, while a good defence is what wins titles. Or, as has been concisely put by American Football coach Bear Bryant: "Offense sells tickets, defense wins championships".

the offensive style of their teams. This ranking is available from the authors upon request.

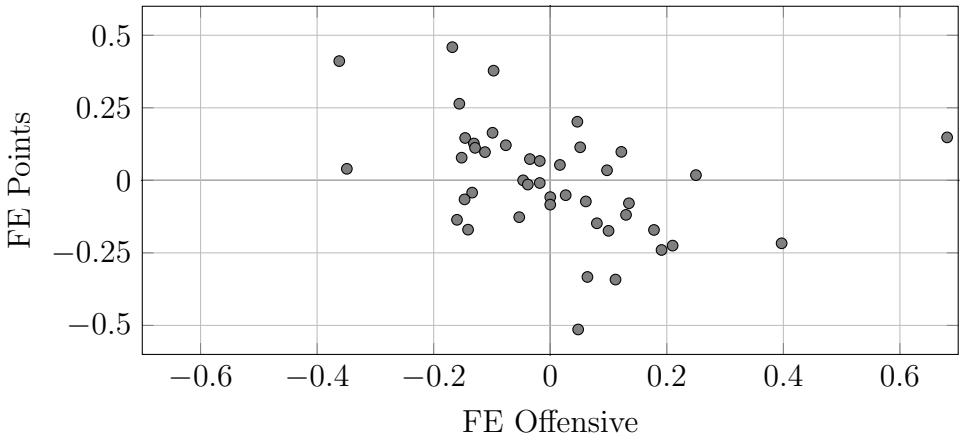


Figure 2.3: Relation Between Managerial Impact on Performance and Team Style

2.6 The Impact of Managers' Background as Professional Players

While the previous analysis was based on the impact of (unobservable) fixed effect of managers, we follow Bertrand and Schoar (2003) and also analyze the impact of *observable* characteristics of the managers on team performance. In particular, we focus on characteristics which are related to a manager's previous career as a professional player before becoming a manager. For example, in professional basketball (NBA) Goodall et al. (2011) find evidence that former NBA top players indeed make better coaches. For the case of soccer, to the best of our knowledge this issue has not yet been addressed in previous academic work. But there is a current public debate about whether or not a good manager needs the right "pedigree" (such as being a former star player or even winner of the World Cup) or whether what really counts is a thorough understanding of the game beyond own playing experience (e.g., in terms of tactics, team leadership and motivation, up-to-date expert support staff).[25] Since anecdotal evidence exists on either side, it is interesting to take a more detailed look at this

[25]For example, Mehmet Scholl, a former star player of Bayern Munich and the German national team, and now an influential TV sports commentator, claims that actual experience as a player matters for being a successful coach. In a recent interview with the leading German weekly magazine *Der Spiegel* he complains about managers who have not been successful players themselves as "... they have never played at the top level, and they have no clue how players at this level operate [...] It is all about tactics, these are mere laptop managers." (see Der Spiegel, Issue 37/2015, pp. 100).

issue.[26] In particular, the following information is available for the managers in our data set (summarized in Table 2.11: i) whether a manager was a former professional player (*Professional*, ii) whether he was formerly playing in his respective national team (*National*) and iii) a dummy variable whether he played on an offensive position (*Off-position*).[27]

Manager type	**Total**	***Professional***	***National***	***Off-position***
All managers	103	89	41	41
Only movers	44	39	17	22

Table 2.11: Managers' Background as a Professional Player

The results for the different categories are reported in Table 2.12 (note that none of the regressions includes manager fixed effects):[28]

As can be seen, the teams of managers who were former professional players do worse than those of managers who were not. This holds irrespective of whether teams are approximated by team fixed effects only (Model O1) or when budget are included in addition (Model O2). Overall, the results provide evidence for a potential overrating of prominent names in the hiring process of managers. Another interpretation is that managers who have not been former star players themselves need to be substantially better coaches in order to secure a job as a head coach in the top leagues. The latter must start their manager career in low divisions and hence, when such managers are promoted to top-tier teams, they have already proven to possess some manager quality beforehand; otherwise they would not have made to a top division team. In contrast, former professionals often start their manager careers directly in the Bundesliga or second division without any significant prior manager experience, where prominent example include Franz Beckenbauer, Jürgen Klinsmann (both Bayern Munich) and Matthias Sammer (Dortmund). In these cases, inferior manager quality only shows up *after* they have taken over a top division team (thereby entering our data set).[29]

[26]For example, while protagonists such as Franz Beckenbauer, Jupp Heynckes or Matthias Sammer were quite successful as both players and managers, in our ranking reported in Table 2.4, four out of the five top managers never made it to the Bundesliga or to some other top league.

[27]As in the regressions of Section 2.4.2, there is no significant effect of manager tenure and/or age on the results when including them as additional control variables.

[28]We have also investigated the impact of these manager characteristics on the offensive style of their teams, and there is no effect. The results are available from the authors on request.

[29]Of course, teams may nevertheless have an incentive to hire big names, because there might be other benefits (e.g., increased media attention or higher match attendance) associated with it.

	Model O1	**Model O2**	**Model O3**	**Model O4**
Professional	-0.107**	-0.100**		
	(0.047)	(0.048)		
National			-0.010	
			(0.038)	
Off-position				-0.015
				(0.030)
Budget	-	0.180***	-	-
		(0.059)		
Constant	1.395***	1.224***	1.294***	1.303***
	(0.075)	(0.102)	(0.066)	(0.066)
N	764	764	764	764
R^2	0.359	0.371	0.355	0.355
adj. R^2	0.294	0.306	0.290	0.290
F-test Team FE	37.46	17.11	30.18	33.55
p-value	0.000	0.000	0.000	0.000

Dependent variable: Average points per game per half-season. Fixed effects for half-seasons and teams included in all regressions. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01. Clustered on half-season level, weighted with the number of matches per manager-team pair in half-season.

Table 2.12: Impact of Managers' Background as Professional Players on Team Performance

Such a mechanism might also explain why our findings are qualitatively different than those of Goodall et al. (2011). In the NBA, it seems that the typical career path of former star players involves first a lower-level position such as assistant coach, and only the successful ones become eventually promoted to head coach.[30] Finally, we find no effect on performance for managers being a former member of a national team, or the position in which they used to play (see Models O3 and O4).

2.7 Conclusion

We have analyzed the impact of managers on the performance of their teams in the context of professional sports. In particular, we have estimated average additional performance contributions for individual managers by making use of the high turnover rates in the Bundesliga which allows to disentangle manager effects from the strength

[30]For example, for the upcoming NBA season 2016/17, 13 out of 30 head coaches have been former NBA players, and 11 out of these held other coaching positions in basketball before taking over their first position as a NBA head coach.

of their respective teams. We found a considerable variation in these performance contributions. Moreover, we have also documented an impact of managers on teams' style of playing, and we show that once famous and successful players do not necessarily make good managers later on in their careers.

Of course the approach also has potential limitations. For example, one could argue that the estimate for managers in top teams like Bayern Munich are computed comparing them only with other top managers while managers in bad teams are compared only with lower qualified managers. However, we observe a substantial number (26) of managers who have worked in teams of very different strengths. For instance, one manager (Felix Magath) has worked in 7 different teams (including Bayern Munich, but also substantially weaker ones such as Nürnberg or Frankfurt). These high frequency movers connect managers across different skill levels and facilitates the identification of their individual effects (see also the argument in Graham et al., 2012). But of course, the individual ability estimates have to be treated with caution for those managers who have worked only in teams which have employed only a few other managers.

A potentially more problematic assumption is the stability of the (relative) strengths of teams across the considered time period which may vary over time due to changes in the financial strength of teams. But as we have shown, the estimated manager fixed effects remain rather stable when we include time-varying information such as the relative budgets of the teams in a given season. A further possibility to address the issue of time-invariance would be to divide the 21 seasons of our data set into shorter time intervals (for example, by including team/season fixed effect vs covering, say, five seasons). However, apart from the fact that any such division of our data set into 5-year periods would appear arbitrary to some degree, this also raises collinearity issues due to a larger congruence of the time periods in which manager-team pairs are observed. For example, when a manager is observed with a team for a whole five-year period, then part of his impact will be picked up by the respective team/season fixed effect and vice versa.

Moreover, we have shown that our ability estimates have predictive power. Using past data to estimate abilities disentangling manager's contributions helps to form better expectations about future performance. In turn, it can help teams to spot talent and to detect undervalued managers on the market.

2.A Estimated Fixed Effect for All Managers (Movers and Non-movers)

The subsequent table provides a ranking of all (mover and non-mover) managers in the final data set.

Estimated Fixed Effects						*Avg. Points per Game*					
Rank	**Manager**	**Coeff.**	**Rank**	**Manager**	**Coeff.**	**Rank**	**Manager**	**⌀Points**	**Rank**	**Manager**	**⌀Points**
1	Tuchel, Thomas	0.829	53	Sidka, Wolfgang	-0.019	1	Guardiola, Pep	2.647	53	Berger, Jörg	1.299
2	Guardiola, Pep	0.694	54	Solbakken, Stale	-0.028	2	Hitzfeld, Ottmar	2.008	54	Lorant, Werner	1.291
3	Pacult, Peter	0.552	55	Gaal, Louis van	-0.028	3	Lewandowski, Sascha	1.975	55	Gerets, Eric	1.289
4	Klopp, Jürgen	0.459	56	Stanislawski, Holger	-0.042	4	Gaal, Louis van	1.937	56	Neururer, Peter	1.287
5	Keller, Jens	0.428	57	Pagelsdorf, Frank	-0.051	5	Klinsmann, Jürgen	1.862	57	Wolf, Wolfgang	1.284
6	Favre, Lucien	0.411	58	Funkel, Friedhelm	-0.058	6	Keller, Jens	1.843	58	Fink, Thorsten	1.266
7	Gross, Christian	0.401	59	Skibbe, Michael	-0.066	7	Trapattoni, Giovanni	1.820	59	Fringer, Rolf	1.265
8	Jol, Martin	0.379	60	Weinzierl, Markus	-0.070	8	Jol, Martin	1.794	60	Scala, Nevio	1.265
9	Korkut, Tayfun	0.378	61	Toppmöller, Klaus	-0.073	9	Heynckes, Jupp	1.788	61	Weinzierl, Markus	1.250
10	Slomka, Mirko	0.378	62	Wolf, Wolfgang	-0.079	10	Gross, Christian	1.769	62	Meyer, Hans	1.240
11	Lewandowski, Sascha	0.361	63	Jara, Kurt	-0.084	11	Sammer, Matthias	1.759	63	Köppel, Horst	1.231
12	Lorant, Werner	0.338	64	Klinsmann, Jürgen	-0.100	12	Rehhagel, Otto	1.729	64	Körbel, Karl-Heinz	1.229
13	Schaefer, Frank	0.333	65	Fink, Thorsten	-0.114	13	Klopp, Jürgen	1.712	65	Dutt, Robin	1.215
14	Schaaf, Thomas	0.314	66	Koller, Marcel	-0.119	14	Daum, Christoph	1.687	66	Schlünz, Juri	1.205
15	Hecking, Dieter	0.264	67	Augenthaler, Klaus	-0.127	15	Löw, Joachim	1.662	67	Lienen, Ewald	1.203
16	Krauss, Bernd	0.257	68	Verbeek, Gertjan	-0.133	16	Hyypiä, Sami	1.655	68	Zachhuber, Andreas	1.180
17	Rehhagel, Otto	0.202	69	Fach, Holger	-0.136	17	Magath, Felix	1.644	69	Möhlmann, Benno	1.164
18	Löw, Joachim	0.168	70	Vogts, Berti	-0.139	18	Schaaf, Thomas	1.618	70	Finke, Volker	1.162
19	Wiesinger, Michael	0.164	71	Engels, Stephan	-0.140	19	Vogts, Berti	1.591	71	Wiesinger, Michael	1.160
20	Sammer, Matthias	0.164	72	Körbel, Karl-Heinz	-0.143	20	Slomka, Mirko	1.556	72	Köstner, Lorenz-Gün..	1.149
21	Olsen, Morten	0.148	73	Gerets, Eric	-0.148	21	Brehme, Andreas	1.547	73	Fach, Holger	1.127
22	Götz, Falko	0.148	74	Trapattoni, Giovanni	-0.170	22	Favre, Lucien	1.545	74	Bongartz, Hannes	1.113
23	Heynckes, Jupp	0.146	75	Dutt, Robin	-0.171	23	Stevens, Huub	1.530	75	Middendorp, Ernst	1.108
24	Brehme, Andreas	0.133	76	Berger, Jörg	-0.174	24	Doll, Thomas	1.508	76	Kurz, Marco	1.100
25	Röber, Jürgen	0.127	77	Marwijk, Bert van	-0.185	25	Neubarth, Frank	1.500	77	McClaren, Steve	1.095
26	Magath, Felix	0.121	78	Heesen, Thomas von	-0.193	26	Röber, Jürgen	1.496	78	Heesen, Thomas von	1.091
27	Gisdol, Markus	0.116	79	Advocaat, Dick	-0.195	27	Krauss, Bernd	1.487	79	Funkel, Friedhelm	1.087
28	Rangnick, Ralf	0.114	80	Middendorp, Ernst	-0.206	28	Rausch, Friedel	1.481	80	Latour, Hanspeter	1.059
29	Meyer, Hans	0.112	81	Rapolder, Uwe	-0.217	29	Rutten, Fred	1.480	81	Pezzaiuoli, Marco	1.059
30	Neururer, Peter	0.098	82	Pezzaiuoli, Marco	-0.217	30	Skibbe, Michael	1.473	82	Koller, Marcel	1.053
31	Hitzfeld,Ottmar	0.097	83	Frontzeck, Michael	-0.225	31	Pacult, Peter	1.469	83	Maslo, Uli	1.048
32	Daum, Christoph	0.078	84	Herrlich, Heiko	-0.234	32	Marwijk, Bert van	1.447	84	Rapolder, Uwe	1.041
33	Veh, Armin	0.073	85	Luhukay, Jos	-0.240	33	Labbadia, Bruno	1.439	85	Luhukay, Jos	1.022
34	Stevens, Huub	0.067	86	Bommer, Rudi	-0.242	34	Ribbeck, Erich	1.431	86	Reimann, Willi	1.017
35	Maslo, Uli	0.066	87	Fringer, Rolf	-0.251	35	Dörner, Hans-Jürgen	1.426	87	Advocaat, Dick	1.000
36	Köppel, Horst	0.064	88	Finke, Volker	-0.268	36	Rangnick, Ralf	1.425	88	Engels, Stephan	1.000
37	Lienen, Ewald	0.053	89	Kurz, Marco	-0.275	37	Stepanovic, Dragoslav	1.414	89	Mos, Aad de	1.000
38	Dörner, Hans-Jürgen	0.049	90	Demuth, Dietmar	-0.276	38	Korkut, Tayfun	1.412	90	Soldo, Zvonimir	1.000
39	Streich, Christian	0.044	91	McClaren, Steve	-0.303	39	Tuchel, Thomas	1.406	91	Stanislawski, Holger	0.981
40	Köstner, Lorenz-Günther	0.040	92	Oenning, Michael	-0.332	40	Jara, Kurt	1.384	92	Solbakken, Stale	0.967
41	Latour, Hanspeter	0.039	93	Möhlmann, Benno	-0.333	41	Veh, Armin	1.367	93	Schneider, Thomas	0.952
42	Schlünz, Juri	0.038	94	Reimann, Willi	-0.342	42	Schaefer, Frank	1.364	94	Frontzeck, Michael	0.942
43	Rutten, Fred	0.035	95	Zumdick, Ralf	-0.358	43	Hecking, Dieter	1.362	95	Herrlich, Heiko	0.909
44	Babbel, Markus	0.035	96	Scala, Nevio	-0.379	44	Toppmöller, Klaus	1.360	96	Verbeek, Gertjan	0.909
45	Zachhuber, Andreas	0.032	97	Gerland, Hermann	-0.401	45	Götz, Falko	1.356	97	Gerland, Hermann	0.882
46	Rausch, Friedel	0.018	98	Schneider, Thomas	-0.421	46	Gisdol, Markus	1.341	98	Zumdick, Ralf	0.857
47	Soldo, Zvonimir	0.017	99	Stepanovic, Dragoslav	-0.424	47	Streich, Christian	1.341	99	Bommer, Rudi	0.853
48	Labbadia, Bruno	0 (Ref)	100	Mos, Aad de	-0.452	48	Sidka, Wolfgang	1.333	100	Sorg, Marcus	0.765
49	Neubarth, Frank	-0.003	101	Bonhof, Rainer	-0.466	49	Babbel, Markus	1.321	101	Oenning, Michael	0.706
50	Hyypiä, Sami	-0.003	102	Ribbeck, Erich	-0.514	50	Augenthaler, Klaus	1.317	102	Bonhof, Rainer	0.696
51	Bongartz, Hannes	-0.009	103	Sorg, Marcus	-0.562	51	Olsen, Morten	1.314	103	Demuth, Dietmar	0.647
52	Doll, Thomas	-0.014				52	Pagelsdorf, Frank	1.303			

All managers not eliminated by conditions F and MT.
Non-mover managers are highlighted in gray.

Table 2.13: Ranking of Mover and Non-Mover Managers by Size of Fixed Effect

Estimated Fixed Effects			*Average Points per Game*		
Rank	**Team**	**Coeff**	**Rank**	**Team**	**Points**
1	Bayern Munich	0.751	1	Bayern Munich	2.082
2	Leverkusen	0.460	2	Dortmund	1.755
3	Dortmund	0.347	3	Leverkusen	1.677
4	Schalke	0.230	4	Schalke	1.604
5	Hamburg	0.207	5	Bremen	1.546
6	Stuttgart	0.177	6	Stuttgart	1.510
7	Augsburg	0.147	7	Hamburg	1.444
8	Wolfsburg	0.147	8	Kaiserslautern	1.444
9	Kaiserslautern	0.143	9	Hertha Berlin	1.418
10	Freiburg	0.117	10	Wolfsburg	1.383
11	Bremen	0.058	11	Mainz	1.301
12	Hertha Berlin	0.033	12	Hannover	1.296
13	Hoffenheim	0.032	13	1860 Munich	1.293
14	Bielefeld	0.015	14	Hoffenheim	1.292
15	Frankfurt	0 (Ref)	15	Mönchengladbach	1.239
16	Bochum	-0.034	16	Frankfurt	1.212
17	Aachen	-0.046	17	Augsburg	1.206
18	Duisburg	-0.116	18	Freiburg	1.178
19	Rostock	-0.124	19	Bochum	1.175
20	Mönchengladbach	-0.128	20	Unterhaching	1.162
21	Unterhaching	-0.142	21	Rostock	1.160
22	Nürnberg	-0.165	22	Duisburg	1.135
23	Hannover	-0.187	23	Nürnberg	1.127
24	Cologne	-0.216	24	Cologne	1.114
25	St. Pauli	-0.353	25	Bielefeld	1.044
26	1860 Munich	-0.354	26	Aachen	1.000
27	Uerdingen	-0.477	27	St. Pauli	0.892
28	Wattenscheid	-0.583	28	Wattenscheid	0.826
29	Mainz	-0.621	29	Uerdingen	0.821

Table 2.14: Ranking of Teams. Fixed Effects (left) and Average Points per Game (right)

2.B Managers and Spells Eliminated by Condition F

	Manager		Manager
1	Achterberg, Eddy	31	Krautzun, Eckhard
2	Adrion, Rainer	32	Lattek, Udo
3	Arnesen, Frank	33	Lieberwirth, Dieter
4	Balakov, Krassimir	34	Lippert, Bernhard
5	Beckenbauer, Franz	35	Littbarski, Pierre
6	Bergmann, Andreas	36	Minge, Ralf
7	Brunner, Thomas	37	Moniz, Ricardo
8	Cardoso, Rudolfo	38	Moser, Hans-Werner
9	Dammeier, Detlev	39	Nemet, Klaus-Peter
10	Dohmen, Rolf	40	Neu, Hubert
11	Ehrmantraut, Horst	41	Preis, Ludwig
12	Eichkorn, Josef	42	Prinzen, Roger
13	Entenmann, Willi	43	Reck, Oliver
14	Erkenbrecher, Uwe	44	Renner, Dieter
15	Fanz, Reinhold	45	Reutershahn, Armin
16	Geideck, Frank	46	Rolff, Wolfgang
17	Gelsdorf, Jürgen	47	Schafstall, Rolf
18	Halata, Damian	48	Schehr, Ralf
19	Hartmann, Frank	49	Scholz, Heiko
20	Heine, Karsten	50	Schulte, Helmut
21	Heinemann, Frank	51	Sundermann, Jürgen
22	Henke, Michael	52	Thom, Andreas
23	Hermann, Peter	53	Tretschok, Rene
24	Hieronymus, Holger	54	Vanenburg, Gerald
25	Hrubesch, Horst	55	Völler, Rudi
26	Hörster, Thomas	56	Weber, Heiko
27	John, Christoph	57	Wilmots, Marc
28	Jonker, Andries	58	Wosz, Dariusz
29	Kohler, Jürgen	59	Ziege, Christian
30	Kramer, Frank	60	Zobel, Rainer

Table 2.15: Managers Without a Spell Satisfying Condition F

	Manager	**Team**	**Matches** (in Spell)	**Year**
1	Adrion, Rainer	Stuttgart	11	1998
2	Beckenbauer, Franz	Bayern Munich	14	1993
3	Bergmann, Andreas	Hannover	16	2009
4	Ehrmantraut, Horst	Frankfurt	16	1998
5	Entenmann, Willi	Nürnberg	15	1993
6	Gelsdorf, Jürgen	Bochum	12	1994
7	Götz, Falko	Hertha Berlin	13	2001
8	Hartmann, Frank	Wattenscheid 09	11	1993
9	Heesen, Thomas von	Nürnberg	15	2007
10	Henke, Michael	Kaiserslautern	13	2005
11	Hörster, Thomas	Leverkusen	11	2002
12	Kohler, Jürgen	Duisburg	11	2005
13	Köstner, Lorenz-Günther	Wolfsburg	15	2009
14	Krauss, Bernd	Dortmund	11	1999
15	Krautzun, Eckhard	Kaiserslautern	11	1995
16	Kurz, Marco	Hoffenheim	10	2012
17	Marwijk, Bert van	Hamburg	15	2013
18	Meier, Norbert	Mönchengladbach	11	1997
19	Meier, Norbert	Duisburg	15	2005
20	Minge, Ralf	Dresden	15	1994
21	Oenning, Michael	Hamburg	14	2010
22	Rangnick, Ralf	Schalke	13	2011
23	Rausch, Friedel	Nürnberg	16	1998
24	Rehhagel, Otto	Hertha Berlin	12	2011
25	Reimann, Willi	Nürnberg	15	1998
26	Schäfer, Winfried	Stuttgart	15	1998
27	Schafstall, Rolf	Bochum	13	2000
28	Schulte, Helmut	Schalke	11	1993
29	Slomka, Mirko	Hamburg	13	2013
30	Stevens, Huub	Stuttgart	10	2013
31	Zobel, Rainer	Nürnberg	14	1993

Table 2.16: Eliminated Spells With at Least 10, but Less Then 17 Matches

2.C Teams Eliminated by Condition MT

	Team	No. of managers	No. of obs	Managers	No. of obs
1	Braunschweig*	1	2	Lieberknecht, Torsten	2
2	Cottbus	3	13	Geyer, Eduard	6
				Prasnikar, Bojan	4
				Sander, Petrik	3
3	Dresden	1	3	Held, Siegfried	3
4	Düsseldorf*	3	7	Meier, Norbert**	2
				Ristic, Aleksandar	3
				Wojtowicz, Rudolf	2
5	Fürth*	1	2	Büskens, Michael**	2
6	Karlsruhe	2	14	Becker, Edmund	4
				Schäfer, Winfried**	10
7	Leipzig	1	2	Stange, Bernd	2
8	Ulm	1	2	Andermatt, Martin	2
		$\sum$13	$\sum$45		$\sum$45

Unit of observation: Half-season
* Some of team's managers are observed with other teams, but these spells do not satisfy condition F.
** Manager observed with several teams, but only one spell satisfies condition F so that manager is not a mover.

Table 2.17: Teams Eliminated by Condition MT and Their Managers

2.D Ranking of Manager-Fixed Effects With Respect to Team Style

	Model 3					*Model 3*			
	Performance		*Team Style*			*Performance*		*Team Style*	
Manager	**R.**	**Coeff.**	**R.**	**Coeff.**	**Manager**	**R.**	**Coeff.**	**R.**	**Coeff.**
Klopp, Jürgen	1	0.459	42	-0.168	Bongartz, Hannes	23	-0.009	23	-0.018
Favre, Lucien	2	0.411	44	-0.362	Doll, Thomas	24	-0.015	26	-0.039
Slomka, Mirko	3	0.378	30	-0.097	Stanislawski, Holger	25	-0.042	35	-0.134
Hecking, Dieter	4	0.264	40	-0.156	Pagelsdorf, Frank	26	-0.051	19	0.027
Rehhagel, Otto	5	0.202	18	0.047	Funkel, Friedhelm	27	-0.058	21	0.000
Sammer, Matthias	6	0.164	31	-0.099	Skibbe, Michael	28	-0.066	38	-0.147
Götz, Falko	7	0.148	1	0.681	Toppmöller, Klaus	29	-0.073	15	0.061
Heynckes, Jupp	8	0.146	37	-0.146	Wolf, Wolfgang	30	-0.079	7	0.135
Röber, Jürgen	9	0.127	34	-0.131	Jara, Kurt	31	-0.084	22	0 (Ref)
Magath, Felix	10	0.121	29	-0.076	Koller, Marcel	32	-0.119	8	0.13
Rangnick, Ralf	11	0.114	16	0.051	Augenthaler, Klaus	33	-0.127	28	-0.053
Meyer, Hans	12	0.112	33	-0.129	Fach, Holger	34	-0.136	41	-0.16
Neururer, Peter	13	0.098	9	0.122	Gerets, Eric	35	-0.148	13	0.08
Hitzfeld, Ottmar	14	0.097	32	-0.112	Trapattoni, Giovanni	36	-0.17	36	-0.141
Daum, Christoph	15	0.078	39	-0.152	Dutt, Robin	37	-0.171	6	0.178
Veh, Armin	16	0.0732	25	-0.035	Berger, Jörg	38	-0.174	11	0.1
Stevens, Huub	17	0.067	23	-0.018	Rapolder, Uwe	39	-0.217	2	0.397
Lienen, Ewald	18	0.053	20	0.017	Frontzeck, Michael	40	-0.225	4	0.21
Köstner, Lorenz-Günther	19	0.040	43	-0.349	Luhukay, Jos	41	-0.24	5	0.191
Babbel, Markus	20	0.035	12	0.097	Möhlmann, Benno	42	-0.333	14	0.064
Rausch, Friedel	21	0.018	3	0.25	Reimann, Willi	43	-0.342	10	0.112
Labbadia, Bruno	22	0 (Ref)	27	-0.047	Ribbeck, Erich	44	-0.514	17	0.048

Table 2.18: Ranking of Mover Managers. Performance Versus Team Style

Chapter 3.

The Hidden Costs of Whistleblower Protection

Abstract

We conduct a laboratory experiment to analyze cooperative behavior between a manager and an employee in the presence of misbehavior and protected whistleblowing. Before taking part in a trust game with her employee, a manager has the opportunity to embezzle money at the expense of a third party. Her behavior is observed by the unaffected employee who may trigger an investigation by a report. We vary the framework with respect to monetary incentives and anonymity in case of a report and compare misbehavior, reporting and cooperative behavior across treatments. Our results suggest that a whistleblower law could deter wrongdoing, but could also have a detrimental effect on cooperation in organizations when it increases the probability for false whistleblowing.

JEL-Codes: C91, D73, K42, M51

Keywords: corporate fraud, corruption, laboratory experiment, business ethics, whistleblowing.

3.1 Introduction

Motivation In an era of corporate fraud causing severe damages, whistleblowing is found to be a major source of fraud detection. Consequently, whenever a large corporate scandal is unveiled by insiders, public discussions emerge how to support employee whistleblowers in coming forward by providing legal protection.[1] This paper investigates experimentally the behavioral effects of protection in the form of incentivized and anonymous whistleblowing in two dimensions. First, we are interested how the reporting behavior of employees and the compliance of managers change after

[1]We focus on whistleblowing as organization members' disclosure of illegitimate practices under the control of their employers, to organizations may be able to effect action as defined by Near and Miceli (1985).

whistleblower protection is introduced. Second, this is the first paper that analyzes how the cooperative climate between employer and employee is affected by changes in the legal framework. The results suggest that an institutional change increasing *expected* whistleblowing not only drives down managerial wrongdoing, but also leads to a decline in productive cooperation.

The extensive and widespread economic damage of corporate fraud is well documented. In a survey by the Economist Intelligence Unit (2015), 75% of surveyed companies reported they had become a fraud victim in the previous year, which is an increase of 14 percentage points from 2012 to 2015, while the Association of Certified Fraud Examiners (2014) find that the average loss caused by fraud amounts to 5% of annual revenues. In a long-term study, Dyck, Morse, and Zingales (2017) estimate the average yearly damage of the U.S. economy due to detected and undetected fraud up to $ 360 billion. Accordingly, detection and deterrence of corporate fraud has become a major target for policy makers.

Our study analyzes the cost and benefits of legal protection for employee whistleblowers which is one relevant instrument to fight corporate fraud. There are intuitive arguments for the use of insider knowledge for law enforcement. First of all, a share of fraud cases cannot be detected by external actors due to their lag of necessary insider knowledge. Therefore, whistleblowers increase the share of fraud cases that can be detected. Second, whistleblowing might not only facilitate law enforcement, but the mere threat of insiders reporting could deter wrongdoing ex-ante.

Correspondingly, evidence in favor of whistleblowing as an instrument for crime deterrence is prominent in economic literature. For example, Dyck, Morse, and Zingales (2010) provide evidence on the general importance of non-traditional governance actors for fraud detection. Investigating fraud in the U.S. economy between 1996 and 2004, they find employee whistleblowers involved in detection in 17% of the cases having a larger share than the SEC, auditors, or the media. Furthermore, the fraction of cases detected with the help of whistleblowers has increased over the past years. The Annual Global Fraud Survey finds for 2015 that 41% of the detected fraud cases were exposed by whistleblowers (Economist Intelligence Unit, 2015). According to the Association of Certified Fraud Examiners (2014), employees were the source in 49% of tips leading to the detection of fraud. These numbers strongly suggest that whistleblowing has already become a major resource for crime detection.

Yet, becoming a whistleblower comprises a non-negligible trade-off for an organiza-

tion member, since she potentially faces costs from a breach of loyalty and career risks. Academic research in business ethics particularly identifies the fear of retaliation, e.g., a dismissal or a denied promotion, as a major obstacle for whistleblowers that has to be overcome, or eventually thwarts whistleblowing (see, e.g. Near and Miceli, 1986; Alford, 2001; Rehg, Miceli, Near, and Van Scotter, 2008; Cassematis and Wortley, 2013). As a consequence, whistleblowers might be encouraged to come forward by legally protecting them from retaliation.

To this end, international organizations as the G20 group, or the OECD requested protection for whistleblowers (OECD, 2016) and legislators made an effort to increase the legal certainty. Prominent examples can be found in the United States with the Sarbanes-Oxley Act - passed in reaction to the whistleblowing-induced collapses of Enron and WorldCom (Healy and Palepu, 2003), the Dodd-Frank Act, and the Public Interest Disclosure Act in the UK.[2]

The most-frequent features of whistleblower protection are protection of employment, i.e. guaranteeing income (see e.g., Kohn, Kohn, and Colapinto, 2004, pp. 97) and allowing for anonymous reporting.[3] These schemes should increase the willingness to report misbehavior, thereby help to uncover a larger share of fraud, but also - the organization anticipating this increase - deter the misbehavior in the first place.

However, these legal approaches are discussed controversially, since these benefits might come at a cost. To increase the legal certainty for the whistleblower reporting may not only be protected (or provided with incentives), but this protection must also be obtainable at a sufficiently low cost. Consequently, legislation often do not condition the protection grant on a successful investigation, instead a wide spread content of whistleblowing laws is the low barrier of a 'reasonable belief' to obtain the protection (Kohn, Kohn, and Colapinto, 2004, pp. 92).[4] While obviously unfounded complaints are deterred with this standard[5], one resulting adverse effect may be nevertheless an increase in false claims. That means blowing the whistle although no underlying fraud has happened to reap the benefits from protection and thereby inducing a damage for the organization and the regulatory agency (Callahan and Dworkin, 1992; Howse and

[2]15 out of 23 surveyed countries have implemented a specific whistleblowing law (see Thüsing and Forst, 2016).

[3]9 of 15 countries allow for anonymous reporting (Thüsing and Forst, 2016).

[4]Cases where protection remains intact even if it turned out that there was no misbehavior are discussed in (Anechiarico and Jacobs, 1996, pp. 67).

[5]Buccirossi, Immordino, and Spagnolo (2017) shows theoretically how to deter unfounded reports by sufficiently high fines.

Daniels, 1995; Givati, 2016). On the one hand, these false claims would cause damage for the organization due to loss in reputation form being investigated. On the other hand, the effort for screening claims for their adequacy would drive down the efficiency of the authorities.

Furthermore, in a corporate context, efficiency does not solely rely on lawful behavior, but also on productive cooperation. For such cooperation to flourish, employers and employees need to share resources and confidential information, which requires a sufficient level of trust. This trust between co-workers might be reduced if employees use sensitive information to file a complaint. For example, the employer's motivation for dismissal may not be punishment, but reputational concerns which make it unbearable to retain a whistleblower - false or honest - in the organization and continue the collaboration. These concerns may mot only occur for actual observed behavior, but also for expected reactions to the legislation. Given anonymity, the manager may now expect a larger share of her employees to blow the whistle on her, leading her to decrease cooperation.[6] Therefore, if a whistleblower protection law encourages more reporting, or even if it increases the *expected* frequency of whistleblowing, it may cause an atmosphere of distrust within an organization, which has detrimental effects on beneficial cooperation (Dworkin and Near, 1997).

This study focuses on these two costs of whistleblower protection and evaluates them against the benefits of detecting and deterring misbehavior due to protection laws in an experimental setting.

Research Question, Framework and Results Our main goal is to investigate the influence of whistleblower protection on reporting behavior, compliance and cooperation. Therefore, we create a workplace setup in the lab in which a manager and an employee share information and cooperate productively. At the beginning of a period, the manager has the opportunity to embezzle money and increase her payoff at the expense of a real third party. Her choice is always observed by the employee. While the employee's payoff is neither negatively nor positively affected by the embezzlement, she can become a whistleblower and trigger an investigation by reporting misbehavior to an authority, independent of the actual decision of the manager. In contrast to other studies, we model the authority to respond perfectly to a report reflecting the standard of a reasonable belief. In consequence, the manager can tell from an investigation that

[6]Bigoni, Fridolfsson, Le Coq, and Spagnolo (2015) find cooperation in cartels decreasing due to distrust induced from potential rewards for reporting.

the whistle was blown.[7] If a report is filed and an investigation happens, reputational cost from the investigation for the manager arise and - if embezzlement is detected - she has also to pay a fine that partly reinstates the third party.

At the end of a period, the manager and the employee interact in a modified trust game (Berg, Dickhaut, and McCabe, 1995). As the sender, the manager decides first which share of her endowment to trust to her employee or to take from the employee's endowment.[8] If this amount is positive, i.e. productive cooperation takes place, it is tripled, and sent to the employee who can in turn decide which fraction of the amount received she wants to return. If the manager choose to take money, that means beneficial cooperation does not take place, the amount is simply transferred and the period ends.

We alter the legal framework in two ways: Compared to a baseline treatment, representing the status quo legislation without any protection, we consider the two most frequent instruments of whistleblower protection, namely the provision of incentives and anonymity for reporting. The incentives for reporting in this study are modeled in the form of protection from monetary losses. While some laws provide bounties as a reward for the whistleblower we consider a rather mild form that relates to guaranteed income, i.e. guaranteeing the employee that the manager cannot take any of her endowment if she files a report.[9] In this way, we test the lower bound of incentives for whistleblowing.[10] Anonymity allows the employee to report without revealing her action prior to the trust game or the investigation to the manager. Therefore, she can be assured of the manager not condition her cooperation on the decision to blow the whistle. These variations result four treatments which allow to identify two possible ways in which whistleblower protection might affect trust of managers towards their employees. A direct influence may result from changed observed behavior of the employee. If incentives leads to more frequent reporting - truthful or false - and this is perceived as unkind behavior, trust, and therefore cooperation, might go down. This is

[7] See Mechtenberg, Muehlheusser, and Roider (2017) on the informativeness of a whistleblower report if the authority has to evaluate a complaint and Chassang and Miquel (2018) on the informational content of an investigation for the employer.

[8] In this respect, the game is similar to the moonlighting game (Abbink, Irlenbusch, and Renner, 2000).

[9] This is similar to Falk and Kosfeld (2006), where the sender in a trust game can require a minimum return from the recipient. In this study it is the recipient who can restrict the sender's choice set.

[10] For studies on bounties as whistleblower rewards see e.g.,Schmolke and Utikal (2016); Buccirossi, Immordino, and Spagnolo (2017); Butler, Serra, and Spagnolo (2018).

captured by the non-anonymous settings. A second possible way would be an indirect effect caused by expected behavior. In the anonymous settings, the manager cannot observe the employee's actual behavior, but may form a belief about the whistleblowing probability. Again, if she expects an unkind behavior, the willingness to cooperate would go down. In this case, the distrust would be caused by institutional framework itself.

By introducing incentivized and anonymous reporting one by one, we change the environment stepwise towards a stronger protection for the whistleblower. This setup allows to track the influence of the protection mechanisms on the employee's willingness to blow the whistle truthfully and falsely, as well as the compliance behavior and the manager's willingness to cooperate. In the context of whistleblowing, a laboratory approach has two major advantages compared to the field. First, only detected fraud is observable in actual organizations, such that the true amount of misbehavior remains unknown. Second, we only observe reporting behavior given the state of compliance. That means, we can account for truthful reporting when fraud was conducted and for false reporting in the case of compliance, but not for the hypothetical behavior in the state that has not been realized. Choosing a laboratory approach solves both of these informational and counterfactual issues. In addition, a number of studies show a high out-of lab correlation in unethical behavior (see for discussion Abeler, Raymond, and Nosenzo, 2018).

Our results show that both incentivized and anonymous reporting increase honest reporting and, in turn, increase compliance. That means whistleblower protection affects the behavior of all parties in the intended way. At the same time, both instruments induce adverse incentives for the employees and lead to an increase in false whistleblowing. For the managers' willingness to cooperate, we find an inverse relation to the frequency of investigations they experience. This phenomenon can be explained best by the perception of (false) whistleblowing as an unkind behavior, which negatively affects the manager's trust in her employee. The joint use of incentives and the provision of anonymity for reporting leads to a peak of investigations and drives down cooperation significantly.

The following section will review related work from business, sociology and economics literature. Sections 3.3 and 3.4 present the experimental design and the behavioral predictions, which will be analyzed in Section 3.5, before Section 3.6 discusses the results.

3.2 Related Literature

This study is the first to investigate the relation of whistleblowing and cooperation considering i) an unaffected whistleblower and ii) an affected real third party. Furthermore, we are not aware of studies that apply the concept of anonymous reporting to the whistleblowing context.

The study closest to our paper is by Mechtenberg, Muehlheusser, and Roider (2017) on the protection of whistleblowers. In a theory-guided lab experiment, they investigate the influence of different whistleblower protection laws on compliance, reporting behavior and retaliation against the whistleblower. In addition, they analyze the investigation decision of the regulatory agencies given the different legal frameworks. They find the desired increasing effect of whistleblower protection on reporting. However, when the legal protection also fosters false reporting, whistleblowing becomes a less informative signal to the regulatory agency such that a higher number of reports does not necessarily materialize in a higher number of investigations. In a framework where the employees are heterogeneous with respect to their productivity the dismissal decision of the manager could be either driven by externally given efficiency concerns or by preferences for retaliation. We complement this study by internalizing the productivity of the collaboration. Still the manager could retaliate against the whistleblower motivated only by punishment. But if the reputational damage is not too high, it depends on whether she trusts her employee enough for the collaboration to be profitable, rather than on externally given productivity. Therefore, this framework applies to broader range of employee whistleblowers who are endangered by a dismissal and are inclined to conduct a false claim.

Furthermore, our work is related and contributes to three strands of the literature on whistleblowing. Recent experimental studies cover the effect of (monetary) incentives on the willingness to blow the whistle. Bartuli, Djawadi, and Fahr (2016) analyze whistleblowing in a context where the employee faces a conflict between ethical considerations and monetary interests. They find employees who are more altruistic and more aware of ethical issues are more likely to refrain from supporting fraud and report wrongdoing. Schmolke and Utikal (2016) measure the effectiveness of incentives on the willingness to report. Fines for non-reporting, rewards and also commands increase the probability of whistleblowing in their setup. If whistleblowers are affected by the misconduct themselves, reporting is more likely if the enforcement authority is

negatively affected as well. Butler, Serra, and Spagnolo (2018) investigate the effect of monetary rewards on whistleblowing in the presence of potential crowding out of intrinsic motivation (see Benabou and Tirole, 2003; Gneezy, Meier, and Rey-Biel, 2011, for theoretical foundation and overview about crowding out). They find an enhancing effect of monetary rewards on the willingness to report and no substantial crowding out of non-monetary motivations.

Another strand considers the effects of whistleblower protection schemes on efficiency. Heyes and Kapur (2009) develop a model which allows to operationalize several behavioral motivations for whistleblowing and they show that the optimal whistleblower protection regime depends on which motivation is the driving force. Friebel and Guriev (2012) show that the possibility of whistleblower protection might harm a firm's productive efficiency if wrongdoers within the hierarchy "bribe" other members of the organization. They show that whistleblower protection might reduce effort incentives. Felli and Hortala-Vallve (2016) provide a model in which incentivized whistleblowing can prevent opportunistic behavior that takes the form of collusion or blackmail between supervisors and employees within an organization.

Our study investigates also the influence of incentives on reporting and the efficiency of whistleblower protection. It contributes to the existing literature by testing the effect of rather mild financial incentives, that is the guarantee to keep the endowment instead of an additional reward. Furthermore, it extends the relationship with the organization the employee could blow the whistle on by adding a productive collaboration.

As a third strand, a large number of studies identify the fear of retaliation as an obstacle to reporting (see, e.g. Near and Miceli, 1986; Alford, 2001; Rehg et al., 2008; Cassematis and Wortley, 2013). Highlighting the role of retaliation, Chassang and Miquel (2018) employ a cheap-talk approach in which an employee can send a report to an monitor, who in turn can then decide whether to intervene. They show that, in environments where anonymous reporting is not feasible, the optimal intervention policy must garble the whistleblower's message, because a very responsive policy would lead to retaliation and prevent reporting in the first place. They assume that retaliation is costly for the manager such that she has commit to a retaliation strategy conditional on whistleblowing ex-ante. In contrast, we allow the manager to evaluate beneficial retaliation against potentially even more profitable cooperation.

Reuben and Stephenson (2013) focus on the willingness to report lies of other members of the organization and highlight the retaliation associated with whistleblowing.

They find that former whistleblowers are less likely to be chosen by an organization, even it complies to the rules and would not have to fear an investigation. This career risk for a whistleblower is also included in our framework. Furthermore, we investigate whether managers continue to ostracize whistleblowing even if this means to pass on an opportunity for profitable cooperation.

In addition, the context of whistleblower protection is related to leniency programs in cartel prosecution. In such programs, cartel members who report their activities to the authorities are rewarded by a fine reduction. The most prominent studies analyze the effects of the reduction of fines or rewards in case of reporting a cartel. They find that a leniency program provided better outcomes than a pure fine regime or the introduction of rewards (Apesteguia, Dufwenberg, and Selten, 2007; Hinloopen and Soetevent, 2008; Feltovich and Hamaguchi, 2018), but also a stabilizing effect for the remaining cartels (Bigoni, Fridolfsson, Le Coq, and Spagnolo, 2012).[11]

As in the case of whistleblower protection, the analysis of such programs is concerned with the reporting of illegal activities from within the respective entity, but there is a crucial difference, since cartel members who report are wrongdoers themselves, while the typical whistleblower is an innocent bystander.

Besides the specific context of collusion, experimental studies have systematically investigated truth-telling in general (see e.g., Gneezy, 2005; Mazar, Amir, and Ariely, 2008; Fischbacher and Föllmi-Heusi, 2013; Muehlheusser, Roider, and Wallmeier, 2015; Abeler, Raymond, and Nosenzo, 2018). Findings demonstrate that participants cheat for their own monetary advantage, but less than predicted by standard economic theory, already when there are no negative externalities to a third party. This feature of an affected third party is added in experiments about corruption. The usual setup models an opportunity for two players to collude beneficially at the cost of a not involved third party (see e.g., Abbink, Irlenbusch, and Renner, 2002; Barr and Serra, 2009). Experimental studies have also investigated the punishment behavior of unaffected third parties. It has been shown that people are willing to punish violation of norms or unethical behavior in the lab, both if punishment is incentivized or not. This behavior can even be found if punishment is costly (see e.g., Fehr and Gächter, 2002; Fehr, Fischbacher, and Gächter, 2002; Fehr and Fischbacher, 2004). Our study contributes to this literature by combining elements from all three strands in a unified setting.

[11] See also the surveys by Spagnolo (2008) and Marvão and Spagnolo (2014).

3.3 Experimental Design

3.3.1 The Game

The game played in each period To investigate the influence of (protected) whistleblowing on misbehavior, reporting and cooperative behavior, we combine a whistleblowing game with a modified trust game. In this experiment, the subjects are assigned with the role of a manager, an employee, or a third party. Those who become a manager maintain their role throughout the experiment, whereas both the other two roles are reshuffled after each period. Before a period starts, groups of three are randomly formed with one subject of each role, such that subjects face a stranger matching and cannot infer any information about their group members from previous periods.

While the third party is completely passive, both the other roles have to make two decisions. In the whistleblowing game, the manager decides about misbehavior in the first stage. That is to comply with the law ($e = 0$) or embezzle money ($e = 1$) which generates an exclusive revenue for her and induces a cost for the third party. In stage two, the employee needs to decide whether to stay silent ($r = 0$) or to file a complaint ($r = 1$). This decision is made conditional on the compliance decision of the manager, i.e. the employee decides about reporting truthfully (r^t) in case the manager misbehaves, and about reporting falsely (r^f) in case the manager complies. Note that the employee is able to report an illegal action, regardless whether it has happened or not. Using the strategy method (Selten, 1967) at this point allows to keep track of reporting behavior independent of the compliance behavior which solves the counterfactual problem. Since the employee is not directly affected by the embezzlement, Brandts and Charness (2011) suggest that using the strategy method on reporting should not yield different results in this context. The respective reporting decision of the employee is disclosed to the group.

The trust game starts in stage three. The manager decides by choosing the level of the investment $c \in [-30, 60]$ whether beneficial cooperation takes place, and if so, to which extent. She can choose a negative amount, which would only mean a transfer from the employee's endowment to her own payoff. This represents the opportunity to retaliate against the employee by taking some of her income. If she chooses to trust, i.e. c is positive, this amount is multiplied by three and transferred to the employee. This

positive multiplier captures the social benefit of productive cooperation. While taking endowment from the employee leaves the aggregated payoff for the group unchanged, trusting a positive amount increases this payoff. Finally in stage four, depending on the received investment, the employee can return an amount t back to her manager.

At the end of a period, if a complaint has been filed, an investigation takes place and causes a cost for the manager. If thereby embezzlement is detected, the manager has to pay a fine in addition and compensation is paid to the third party which partly covers the damage. The general timing of a period is presented in Figure 3.1.

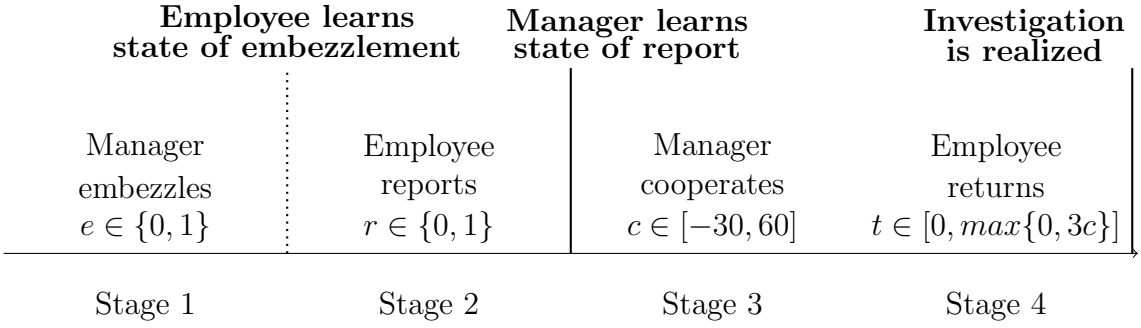


Figure 3.1: General Timing

Cost and reward parameters There are four possible combinations of the decisions on embezzlement and reporting (e, r) which can be ranked straightforwardly in terms of social welfare $\pi_S(e, r)$ if three assumptions hold: (i) the manager complying to the law is always better than non-compliance, (ii) detected fraud is better than undetected fraud, and (iii) in case of compliance, reports should not arise. In this case the order is given by

$$\pi_S(e = 0, r = 0) > \pi_S(e = 0, r = 1) > \pi_S(e = 1, r = 1) > \pi_S(e = 1, r = 0).$$

We chose the cost and reward parameters (in parentheses) such that these assumptions hold. The intuition is as follows: Clearly, the most preferred outcome would be the absence of misbehavior and reports $(e = 0, r = 0)$. In this case neither damage from misconduct nor investigatory costs arise, which leaves all players with just their endowment and the social welfare unaffected $(\Delta\pi_S = 0)$. If we assume efficient law enforcement, the least favorable outcome is an undetected embezzlement $(e = 1, r = 0)$.

Here the manager reaps a benefit (50), which is outweighed by the cost for the third party (90). This would result in a social net loss ($\Delta\pi_S = -40$). Compared to this, a preferable outcome would be a detected embezzlement ($e = 1, r = 1$). The manager would have to pay a fine (60) which exceeds her benefit from embezzlement (otherwise embezzlement would be a payoff maximizing strategy in the whistleblowing game) and the costs of the investigation (10). On the other side, the third party partially recovers her loss R (80), such that social welfare loss is reduced ($\Delta\pi_S = -30$). We follow a similar argumentation to Hart and Zingales (2017) that a firm does not necessarily maximize its shareholders welfare by maximizing its market value. If the third party could be fully compensated for the damage, embezzlement reduces to be only a distributional issue and the original state could always be reconstituted. The fourth possibility is a false claim ($e = 0, r = 1$). Reporting, although there has not been misconduct, means that there is neither a damage for the third party nor a benefit for the manager, but it creates investigation cost (10) for the manager ($\Delta\pi_S = -10$). This has to be positive to reflect the reputational costs of being investigated for compliance issues. If this would have a cost of zero, it would indicate an indifference towards being investigated which is clearly not the case in reality (otherwise investigation should not rely on reports). Also, if the cost of investigation was too large and would outweigh the recovered loss, a social planner would prohibit an investigation. The four possible outcomes of the whistleblowing game are summarized below in Table 3.1.

		Whistleblowing	
		No	Yes
Embezzlement	No	0	−10
	Yes	−40	−30

Table 3.1: Change in Social Welfare After the Whistleblowing Game

For the trust game, we impose a range from −30 to 60 (with discrete steps of length ten) on the amount c that the manager can send to can her employee. Instead of a binary choice for the manager to punish or not to punish, we allow her to gradate the amount she wants to take from the employee. In this way, we are able to disentangle

the motives to not cooperate or even to retaliate. Assume a manager does not want to cooperate, because she does not expect this to be beneficial, but also she does not want to take endowment of her employee. In this case she could just choose $c = 0$. If punishment was a binary choice, a manager who wants to punish may abstain from retaliation if she perceives the size of the punishment as to high. The gradations of c give the manager the opportunity to differentiate whether she wants to either recover precisely an experienced loss from an investigation (10), or from detected misbehavior (20), or simply guarantee herself a profit in any case if she chooses $c = -30$. For positive values of c the upper bound is set to 60. This guarantees that the employee cannot punish the manager stronger by keeping the entire trusted investment than by filing a report (also 60). The endowment is set sufficiently high (100) that neither party could make a loss nor is restricted in her choice options. Figure 3.2 reports the payoffs for the three roles in a given period conditional on the decisions of the subjects.

$$\pi_{Manager} = 100 + e \times (50 - (60 \times r)) - r \times 10 - c + t$$
$$\pi_{Employee} = 100 + \begin{cases} c \times 3 - t & if\ c > 0 \\ c & if\ c \leq 0 \end{cases}$$
$$\pi_{3rdParty} = 100 - e \times (90 - (80 \times r))$$

Figure 3.2: Payoffs

3.3.2 Treatments

In this section, we present the design of the treatments used to separate the effects of interest. As mentioned above, we vary the legal environment in two dimensions: i) the provision of incentives which means an insurance against a monetary loss and ii) anonymity which means that the employee has not to reveal her reporting decision to the manager. Altering the dimensions one by one results in a total of four treatments as depicted in Table 3.2. The four treatments differ with respect to the choice set for the manager in the trust game conditional on the reporting decision of the employee (incentives) and the date when the manager is informed about the reporting decision (anonymity).

		Anonymity: No	Anonymity: Yes
Incentives	No	Baseline (B)	Only anonymity (A)
	Yes	Only incentives (I)	Anonymity and incentives (AI)

Table 3.2: Treatments

Baseline Treatment (B) In the baseline treatment, equivalent to the illustration in Figure 3.1, the manager knows after stage two about the employee's reporting decision, i.e. *before* she chooses c, and is free to choose a negative c independent of the reporting decision.

Incentives Treatment (I) In treatment I, in which only incentives are introduced, the manager knows whether the employee blew the whistle after stage two as in the baseline treatment. In this treatment the feature of employment protection is modeled such that by filing a report the employee can guarantee her status quo payoff. That means, if there has been a report - truthful or false - c has to be at least zero.

Anonymity Treatment (A) In treatment A, in which only anonymous reporting is granted, the information about whistleblowing is only disclosed after stage four through the investigation, i.e. *after* she chooses c. The choice of c is again unrestricted for any reporting decision. This change in the timing guarantees that the manager cannot condition her cooperative behavior on the actual behavior of the employee. Nevertheless, the manager has the opportunity to retaliate against her employee, if she suspects an unkind action, although she cannot observe it.

Anonymity and Incentives Treatment (AI) In treatment AI, when both incentives and the anonymity provision are in place, the manager knows only after her choice on c whether the employee blew the whistle. In case the manager chose a negative c, it is set ex-post to zero. The timing of treatments which incorporate anonymity is depicted in Figure 3.3.

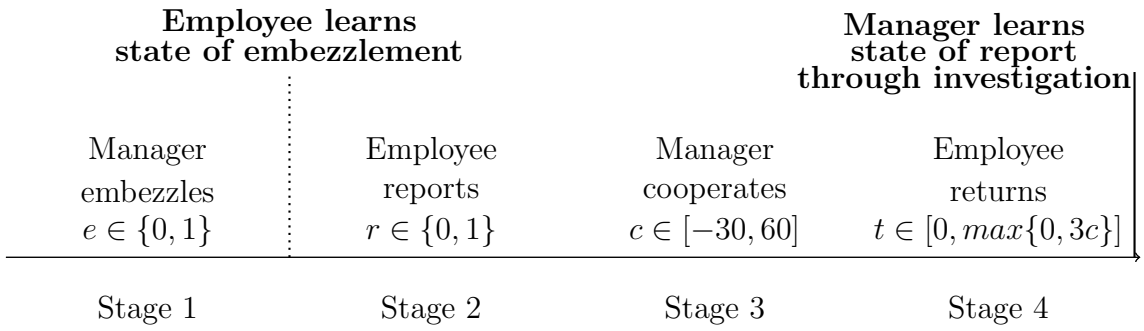


Figure 3.3: Timing in the Treatments With Anonymity

These four treatments allow to identify two possible ways in which whistleblower protection might affect trust of managers towards their employees. The non-anonymous treatments capture the channel of a direct influence through changed observed behavior of the employee. If incentives leads to an increase in truthful or false whistleblowing, this might be perceived as unkind behavior which could drive down trust and thereby the willingness to cooperate. The anonymous settings capture a possible indirect effect. Here the manager cannot observe the actual behavior of the employee, but may form a belief about the whistleblowing probability. If she expects the employee to report truthfully or falsely, the willingness to cooperate might go down – independent of the actual behavior of the employee.

3.3.3 Implementation

The decision whether to implement these treatments with a between-subject or a within-subject design contains several trade-offs. Between designs are more conservative, but may have limitations in relation to testing several variations. On the other hand, within designs are more powerful, but can suffer from confounds (for discussion, see e.g., Charness, Gneezy, and Kuhn, 2012; Moffatt, 2015). A deciding factor for the design choice is the research question at hand and its practical implications. This study is motivated by the debate on supporting whistleblowers by introducing whistleblower protection, for example in the form of incentives and anonymous reporting. Therefore, the most natural design appears to be a within variation to observe a change in behavior after the whistleblower protection is introduced. Of course, confronting the subjects with four different treatments would pronounce the disadvantages of a within

design, for example, the issue of order effects.[12] Also, presenting several changes of the environment to one subject could provoke an experimenter demand effect and thereby bias the behavior. In consequence, we chose to have just one of the dimensions varied for the same subject. Introducing the anonymity environment means a larger modification, since it changes the information structure within a period, while incentives only changes the choice set for the trust game. Therefore, we model the introduction of incentives as a within-subject variation, and to capture the anonymity provision in a between-subject design. Still, the within design has to be implemented carefully. We chose to have multiple periods (8) per treatment for two reasons. First, we want to observe subjects in different scenarios. While reporting is elicited via the strategy method, employees know about the embezzlement decision of the manager. Also, in treatments without anonymity, managers know about the whistleblowing decision of the employee. Second, optimal behavior in this experiment depends on the beliefs about the reciprocal behavior of the other participants. Therefore, we allow the subjects to gain experience over the distribution of types and take the average decision in a role as observational unit. In this way, we reduce the influence of the treatments before the introduction of incentives on those after the intervention.

Framing In line with Mechtenberg, Muehlheusser, and Roider (2017), we framed the experiment in a workplace context and spoke of employers and employees to support the subjects in understanding the hierarchical relation between the players. Furthermore, we chose to phrase the choice about embezzlement in a neutral way and spoke of alternatives (CIRCLE or TRIANGLE) to not induce an experimenter demand effect. However, using payoff tables (see Appendix 3.A) and control questions (see Appendix 3.B), we made clear that the precise consequences for the manager as well as for the third party are understood. We gave a legal reminder that the alternative corresponding to embezzlement means a violation of law. Herewith we model an important feature of unethical decision-making in the real world, since organizations are clearly aware of illegality of such decisions.[13] The employee's decision about a report was phrased as 'filing a complaint' to make them aware of the social undesirability of embezzlement. Drawing attention to unethical behavior may influence the subjects' decisions, which

[12] Complementary to the study design at hand would be a "repeal of whistleblower protection." The reversed order would therefore also address a slightly different research question.

[13] Pruckner and Sausgruber (2013) provide evidence that legal reminders influence unethical behavior.

would be appropriate for our specific fraud-related research question, though.[14]

Procedural details The experiment was programmed with the software z-Tree (Fischbacher, 2007) and conducted in the laboratory of the University of Hamburg, June 2016, using hroot for recruitment (Bock, Baetge, and Nicklisch, 2014). While we asked the control questions at the start of a session, subjects completed a (non-incentivized) questionnaire in which we elicited socio-demographic information (e.g., age, gender, and field of study), risk preferences (via the "100,000 euro" question of Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner, 2011), and their attitudes towards revealing misbehavior (measured on a five-level Likert scale) at the end. To keep the incentives identical for every period over the entire experiment, after the questionnaire has been completed, one period was randomly drawn for payout. We ran five sessions with a total number of 147 student subjects (65% female, ∅ age of 25 years). The majority of the subject were enrolled in economics or business programs. The subjects received payments between 5.50 and 18.50 euro (including a show-up fee of 5 euro) with an average of 10.07 euro.

3.4 Behavioral Predictions

In this section, we establish a set of behavioral predictions concerning the willingness of the employees to blow the whistle, truthfully and falsely, as well as decision of the managers to misbehave and to cooperate conditional on the legal environment. Since the variations in the environment are intended to influence the employee's behavior in the first place, we start with the predictions on truthful and false whistleblowing.

Predictions on truthful whistleblowing (r^t) We first turn our attention to the baseline treatment. According to standard economic theory - in the absence of other-regarding preferences, the employee would be expected not to blow the whistle on embezzlement, since she is not directly affected by the misbehavior and has no further incentives to report. However, a large body of empirical field and experimental studies have accumulated evidence that norms of equity and fairness play an important role (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002).

[14] Alekseev et al. (2017) survey a wide range of experimental literature with respect to the instructions and find that meaningful language could be useful for understanding the environment. For the context of unethical behavior see Abbink and Hennig-Schmidt (2006); Barr and Serra (2009).

Also there is evidence for punishment from unaffected third parties, both if punishment is incentivized or not (Fehr and Gächter, 2002; Fehr, Fischbacher, and Gächter, 2002; Fehr and Fischbacher, 2004). In the context of whistleblowing, "conscience cleansing" might be a motive to come forward as a whistleblower already in the absence of protection (Heyes and Kapur, 2009). On the other side, the manager may consider a report as disloyal behavior driving down her willingness to cooperate in the trust game. Therefore, we can assume that each employee has to weigh the benefits from a report, i.e. reduced social damage from an unfair behavior, against the expected losses from possible retaliation in the trust game.

Based on the experimental evidence for fairness preferences, we expect that already in the baseline treatment a positive fraction of employees blow the whistle truthfully.

In treatment A, this trade-off changes slightly. While the possibility to recover the damage remains the same, the manager now cannot observe reporting and has to make her choice to cooperate independent of the employee's actual decision. Put differently, the benefit side for the employee stays constant, while the costs become independent of her decision. That means she would blow the whistle if she weakly prefers to reduce the damage from embezzlement. Since this a lower threshold than in the baseline treatment, we conclude that the share of truthful reporting should rise with the introduction of anonymity.

In treatment I, in which only incentives are added in contrast to the baseline treatment, the trade-off is affected in a different way. Still, the benefit with respect to recovering social damage from reporting the misconduct remains the same. On the other hand, the employee still might suffer from lower trust, but she is now insured against losses from both, managers that simply do not cooperate and those who retaliate in the trust game if she reports. Therefore, also in treatment I truthful reporting should rise compared to the baseline treatment. The comparison to treatment A depends on the change in expected profit from cooperation in treatment I when a report is sent in contrast to when it is not sent. When the expected profit from cooperation in case of silence is positive, but drops to zero in case of a report, the probability for a report may be lower in treatment I compared treatment A. On the other hand, if the expected profit from not reporting is negative, the probability for reporting may be higher.

When reporting is anonymous and incentivized in treatment AI, the employee again can not be punished in response to her behavior. In addition, reporting offers an

insurance against managers who do not trust in any case. As a result, we expect the reporting frequency to increase compared to both treatments A and I. Taken all together, the predictions on truthful reporting are summarized as follows:

> **RT1**: *In the baseline treatment, truthful reporting will occur with a positive frequency*
>
> **RT2**: *Truthful reporting will be more frequent with anonymous reporting than with non-anonymous reporting for a given status of incentives*
>
> **RT3**: *Truthful reporting will be more frequent with incentives than without for a given status of anonymity*

Predictions on false whistleblowing (r^f) In a similar vein, we can predict the frequency of false claims. In contrast to truthful whistleblowing it is not the preference to recover damage from unjust actions which drives the decision on reporting, but the trade-off between one's own monetary gain and the moral cost of imposing a monetary cost on the manager. Evidence from experiments on unethical behavior (Gneezy, 2005; Mazar, Amir, and Ariely, 2008; Fischbacher and Föllmi-Heusi, 2013; Abeler, Raymond, and Nosenzo, 2018) suggests that individuals attach importance to moral concerns, even if unethical actions are unobserved.

Starting with the baseline treatment, incentives are not provided for the employee to report alleged misconduct and nobody would benefit from this report. Only the manager would suffer a cost. Consequently, false whistleblowing should not occur.

In treatment A, the picture remains the same. While the reporting behavior cannot be observed, there are still no incentives to report falsely, such that there should be no whistleblowing either. There are also no reasons to punish a manager by a false claim for experienced behavior, since the subjects interact in a repeated one-shot games with stranger matching.

In contrast, treatment I offers the employee an opportunity to insure herself against a not trusting manager by guaranteeing herself $c \geq 0$. This means, in the decision upon false whistleblowing she faces a trade-off between securing her endowment on the one side and bearing a moral cost from imposing the investigation cost on the manager plus potentially receiving a lower level of trust in response to this unkind behavior. Based on experimental evidence for risk and loss-aversion (Harrison and Rutstrom, 2008), we

expect that a share of subjects gives more weight to the opportunity to avoid a loss and therefore false whistleblowing to occur in this treatment.

This trade-off changes in treatment *AI*, when reporting is also anonymous. The employee can still insure herself against the potential loss, but the manager cannot retaliate against her actual behavior. Thus, the share of false whistleblowing should increase compared to treatment *I*.

RF1: *False reporting will not occur in the baseline treatment*

RF2: *False reporting will be more frequent with anonymous reporting than with non-anonymous reporting when also incentives are given, but not without incentives*

RF3: *False reporting will be more frequent with incentives than without for a given status of anonymity*

Predictions on embezzlement (e) Predictions about the compliance behavior of the manager can be derived directly from the expected truthful whistleblowing behavior. Embezzling only pays off, if there is no whistleblowing. That means the higher the probability that the whistle will be blown, the lower the expected payoff from embezzlement. Thus, we can formulate the predictions on embezzlement inversely to those on truthful reporting. In addition, again with respect to evidence for costs of unethical behavior and fairness preferences, the expected profit must outweigh the moral costs from harming the third party. Therefore, we expect the managers to embezzle with less than maximal frequency, even if it was maximizing expected profit.

E1: *Embezzlement will occur with less than maximal frequency in the baseline treatment*

E2: *Embezzlement will be less frequent with anonymous reporting than with non-anonymous reporting for a given status of incentives*

E3: *Embezzlement will be less frequent with incentives than without for a given status of anonymity*

Predictions on cooperative behavior (c) Concerning the willingness to cooperate, the behavioral predictions are ambiguous. To start with, standard economic theory would predict the manager not to cooperate at all. Since she would not expect the employee to return anything of the trusted amount, she maximizes her payoff by choosing the smallest possible c. However, evidence from experimental studies strongly suggests that subjects show reciprocal behavior and trust their counterparts (Fehr and Schmidt, 2006).[15] If the trusting behavior is unaffected by the whistleblowing environment, differences in cooperation should not arise across the treatments. Alternatively, if managers are reciprocal players (see e.g., Fehr and Fischbacher, 2002) do perceive reporting as an unkind behavior, according to the previous predictions, cooperation may vary with the institutional framework as well as with the number of whistleblowing cases. That means the higher the reporting frequency of the employee the lower the willingness to trust of the manager. If this behavioral response can be found for truthful claims, the effect could be even more pronounced for false whistleblowing, since a false report is a less reasonable cause for breaching loyalty than a truthful allegation.

Furthermore, cooperation may depend on observed as well as on expected whistleblowing, such that the impact of increased reporting can also play out in the treatments that feature anonymity. While the manager can respond directly to truthful and fraudulent claims in the treatments B and I, she can also form expectations about reporting behavior when whistleblowing is anonymous. As for the compliance decision discussed above, the manager expects a certain probability for a truthful or false report, and takes her cooperation decision in the anonymity treatments also with respect to anticipated behavior. The resulting hypotheses for cooperative behavior are summarized below.

C1: *Cooperation will occur with a positive frequency in the baseline treatment*

C2: *Cooperation will decrease with expected and observed investigations*

C3: *The decreasing effect of investigations on cooperation will be stronger for false claims*

[15] Across studies, subjects usually invest half of their endowment in trust games and receive approximately the invested amount in return.

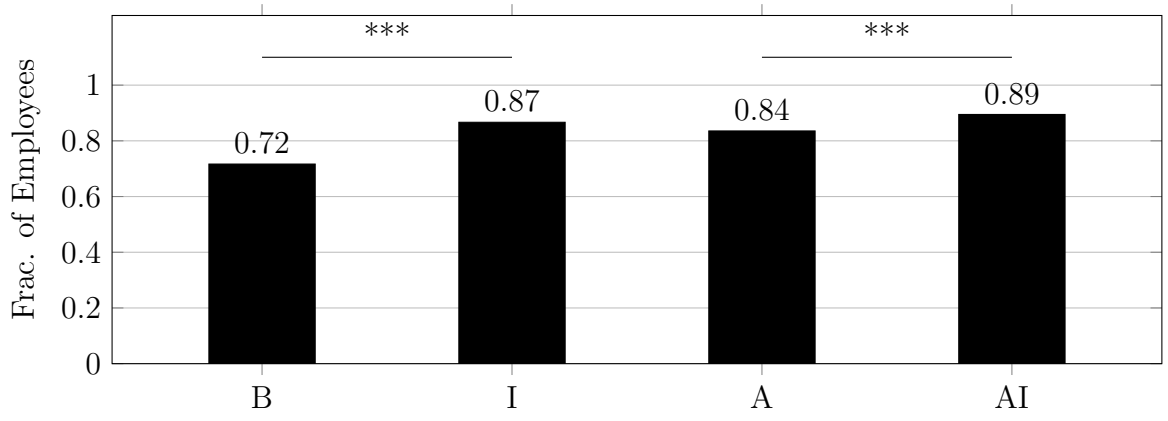
3.5 Results

This section analyzes how the treatments representing different legal frameworks affect the subjects' behavior. We compare the treatment differences to identify the effects of incentives and anonymity provision on aggregated compliance, reporting and cooperative behavior. To test for statistical significance, we follow Moffatt (2015) and use non-parametric tests with subject-role-level averages as observational units. For between-subject differences we apply a Mann-Whitney U test, while we account for within-subject differences with a Wilcoxon signed-rank test. The figures below report the fraction of subjects which opted for a respective decision. There is one bar for each of the four treatments. As for the behavioral predictions, we start with the reporting behavior of the employees and investigate whether the whistleblower protection induces the desired change in the willingness to blow the whistle.

Truthful whistleblowing (r^t) Since we use the strategy method for the employee's decision on reporting, we do not only observe actual whistleblowing, but are able to track separately how the willingness to report truthfully as well as falsely evolves across the treatments, independent of the compliance decision of the manager. Figure 3.4 displays the fractions of employees choosing truthful whistleblowing in the respective treatments. In the baseline treatment, 71.7% decide to report a misbehavior of the manager conditionally on this misconduct actually happening. This supports our prediction **RT1** and is in line with experimental findings for fairness preferences. To evaluate the effect of the protection schemes on the willingness to report, we compare the outcome of the treatments I and A relative to treatment B. For treatment I, we find a significant rise of 16 percentage points to 86.7% ($p < 0.001$). On the other hand, when employees can report anonymously, the fraction rises to 83.6% although this increase is not statistically significant ($p < 0.200$) in treatment A. The highest fraction of whistleblowing is found in treatment AI with 89.5%. Introducing incentives in addition to anonymity leads to a significant increase of 5.9 percentage points compared to treatment A ($p < 0.001$, only 2 out of 38 subjects decrease the reporting frequency). These results provide evidence that both protection schemes affect the employee's trade-off in the desired direction to drive up truthful whistleblowing, thereby lending support to the predictions **RT2** and **RT3**.

Result 1 *Truthful reporting occurs in the absence of protection. (Support for **RT1**)*

Result 2 *Truthful reporting increases with both incentives and anonymity. (Support for **RT2** and **RT3**)*



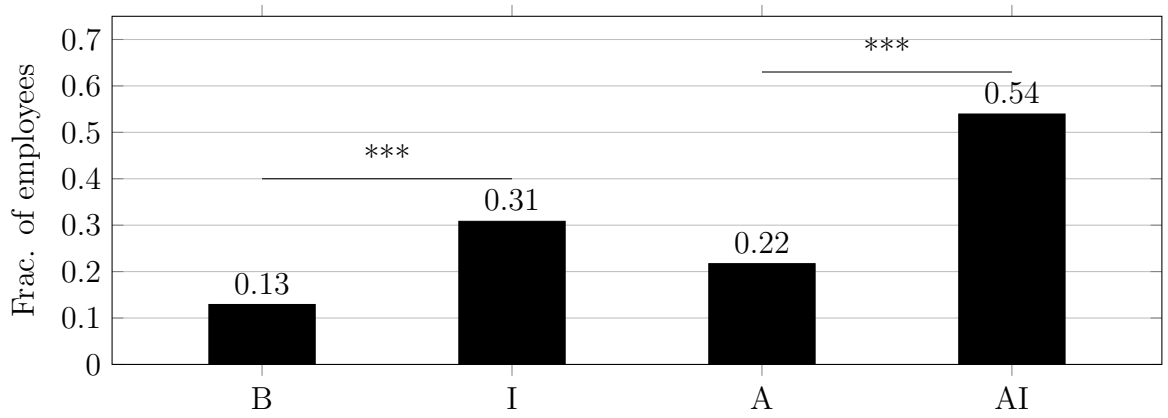
Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Figure 3.4: Truthful Claims Across Treatments

False whistleblowing (r^f) Analogously to the paragraph above, Figure 3.5 displays the willingness to conduct a false report. Surprisingly, we find overall 12.9% of the employees would blow the whistle although there was no misbehavior already in the baseline treatment. This result does not support the prediction **RF1**, since in the absence of incentives false whistleblowing was not expected to occur. A possible explanation could be negative reciprocal behavior induced by undesired decisions by managers in previous periods. Although the employees cannot target the managers whom behavior they disliked, some employees might still want to punish managers in general. In line with prediction **RF2**, we find that the share of false reporting does not increase significantly in treatment A (21.7%, $p < 0.532$). This suggests that anonymity alone does not provoke false claims. However, as expected, introducing incentives in treatment I leads to a significant jump in false reports to 30.8% ($p < 0.001$). When these incentives are introduced additional to anonymous reporting in treatment AI, the share of employees willing to file a false claims peaks with 54.0% ($p < 0.001$). Both these findings support the predictions **RF2** and **RF3** and suggest that subjects react also to the adverse incentives of whistleblower protection.

Result 3 *Employees are willing to report falsely already in the absence of incentives. (Rejects **RF1**)*

Result 4 *Employees' willingness to report falsely increase with the introduction of incentives and with anonymity only when incentives are in place. (Support for **RF2** and **RF3**)*



Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Figure 3.5: False Claims Across Treatments

Considering the results for truthful and false whistleblowing, the experiment already provides evidence for costs as well as for benefits of whistleblower laws. Protection does favor desired behavior and increase righteous reports, but produces adverse effects for false claims at the same time.

Embezzlement (e) Previous results indicate that under whistleblower protection embezzlement would be reported more often. Further, it is of interest whether this changed behavior induced by the legal environment is anticipated by the managers and already deters illegal behavior. Therefore, the focus turns to the compliance decisions of the managers, depicted in Figure 3.6.

In the baseline treatment, a fraction of 41.3% opting for embezzlement, although there are no incentives for the employee to report non-compliance. This supports the prediction **E1** that a significant share managers either anticipate the altruism of the employees or also their own fairness preferences drive the decision not to embezzle

money from the third party. Comparing this to incentivized reporting in treatment I, we find a significant drop of 17.1 percentage points to 24.2% which corresponds to a decrease of 41% in illegal behavior ($p < 0.001$) and supports prediction **E2**. When instead treatment A is contrasted, in which anonymity is granted to the employee, also a lower share of 31.6% decides to embezzle money from the third party (support for **E3**). However, this decline is not statistically significant ($p < 0.578$). In treatment AI, when both protection schemes are in place, only 7.9% of the managers decide to behave illegally. This means significant declines in contrast to treatment A ($p < 0.001$) as well as to treatment I ($p < 0.213$).

Result 5 *Managers choose to embezzle in the absence of whistleblower protection, but not to the maximal extent. (Support for **E1**)*

Result 6 *The frequency of embezzlement decreases with incentives and anonymity. (Support for **E2** and **E3**)*

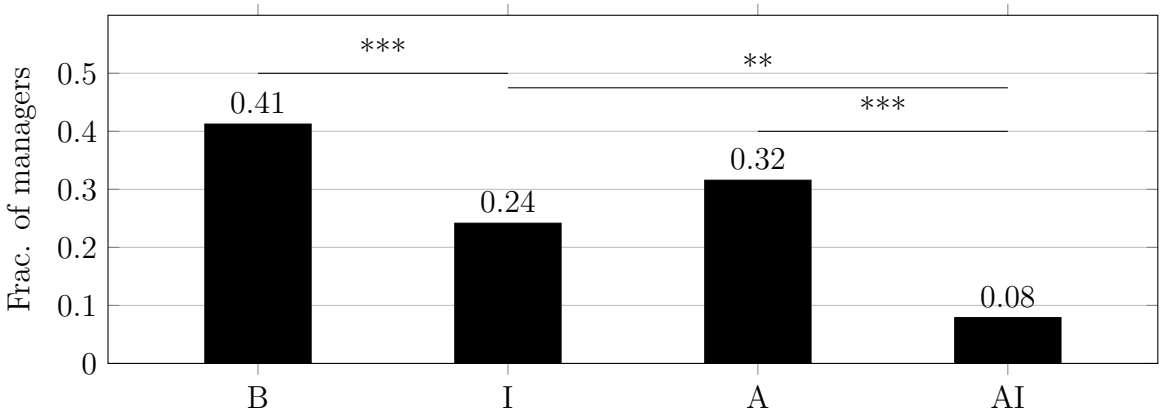


Figure 3.6: Embezzlement Across Treatments

These results suggest that the managers anticipate the reporting behavior of the employees correctly and adjust their cheating frequency downwards with increased probability for whistleblowing. Therefore, the results are in line with the predictions and provide evidence for a beneficial deterrence effect of both whistleblower protection schemes on managerial misbehavior.

Cooperation (c) Having analyzed the whistleblowing and compliance behavior, we now evaluate the willingness to cooperate and the level of cooperation over the different treatments. Figure 3.7a depicts the share of managers that chose to send a positive amount c to their employees in the respective treatment. Since the predictions about cooperative behavior depend on the observed and expected whistleblowing frequency, Figure 3.7b reports the combined truthful and false whistleblowing cases across treatments.

Looking at the baseline treatment first, we find a fraction of 30.4% of the managers choosing to cooperate (whistleblowing cases: 40%. 32% false reports, 8 % truthful reports), which supports our prediction **C1**. The willingness to cooperate moderately increases in treatment A (33.5%, $p < 0.851$), while the number of whistleblowing cases, if at all, goes down to 39% ($p < 0.892$). Although we cannot provide statistical significance, the change in cooperation shows the predicted upward adjustment to the slight decrease frequency of investigations. In treatment I, a similar picture emerges. We find an increase for the overall number of investigations compared to the baseline treatment (46%, $p < 0.345$), and especially for investigations from false reports (24.6%, $p < 0.001$). Correspondingly, the share of the managers choose to cooperate decreases to 25.8% ($p < 0.258$) in response to higher number of whistleblowing cases. [16] These results are in line with the hypotheses **C2** and **C3**, which predict an inverse relation between investigations and cooperation. Nevertheless, they cannot confirm the predictions, since the treatment effects between treatment B and the treatments A and I for whistleblowing and cooperation are not found to be statistically significant. Considering treatment AI on the other hand, only 17.1% of the managers decide to cooperate, which is a significant drop compared to treatment A ($p < 0.019$). This coincides with the highest number of whistleblowing cases (58%)[17], and even more strikingly, with the highest number of investigations caused by false claims (50.0%).[18]

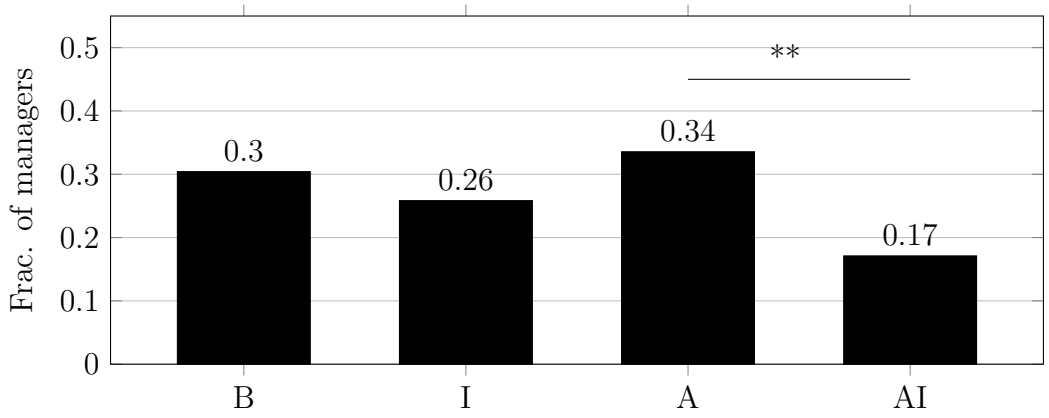
[16] Considering managers conditionally on their embezzlement decision (see Figure 3.D.1 in the appendix), we find a large differences when the employee did not file a report. Those who embezzled money and were not reported invested much more frequently in the trust game compared to those who were reported and those who complied. This shows a pattern similar to a gift exchange, suggesting that loyalty may be a driving factor for productive cooperation. However, note that the number of cases is very small such that we cannot make a claim about statistical significance.

[17] The number of whistleblowing cases in treatment AI is significantly higher than in treatment A ($p < 0.003$) and in treatment I ($p < 0.030$).

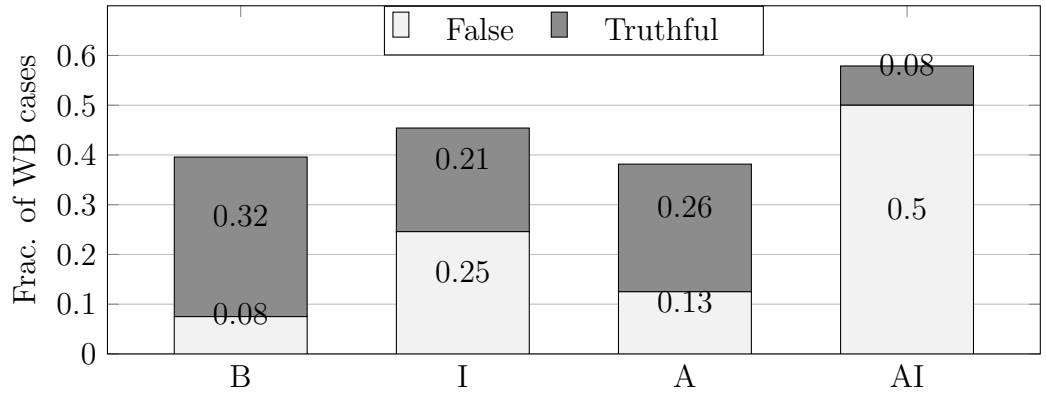
[18] The number of false whistleblowing cases in treatment AI is significantly higher than in treatment A ($p < 0.001$) and in treatment I ($p < 0.002$).

Result 7 *Managers choose to cooperate in the absence of whistleblower protection. (Support for **C1**)*

Result 8 *The willingness to cooperate declines with an increased expected number of investigations. (Support for **C2** and **C3**)*



(a) Cooperation Across Treatments



(b) Whistleblowing Cases Across Treatments

Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Figure 3.7: Trusting and Reporting Behavior Across Treatments

Since treatment *AI* produces significant increases in both overall reporting cases and false claims, we cannot assign the decrease in cooperation based on these results to either of these factors in particular. However, the data suggests that false whistleblowing plays the dominant role, since it accounts for over 85% of the whistleblowing

cases, which is 52 percentage points larger than in treatment A. In contrast, the total whistleblowing cases increase by only 20 percentage points. By disentangling cooperation in treatment AI with respect to the compliance decision (see Figure 3.D.1 in the appendix), we find only 15.7% of the managers investing if they did not embezzle, while those few who opted for misbehavior choose to cooperate in every third case.

Additionally, contrasting the two anonymity treatments provides an intuition that expected false claims cause the trust to decrease. Note that the only difference between the two treatments is given by the incentivized whistleblowing in treatment AI. This means the manager cannot observe any change in actual behavior, but only form an expectation about the employee's choice when she decides about cooperation. Thus, the employee can already report truthfully in treatment A without being retaliated, while there are no incentives for false claims. The managers anticipate an even higher willingness to report truthfully in treatment AI by cutting down illegal behavior substantially. This means, truthful reports cannot arise more often, since embezzlement is almost completely deterred. Only false whistleblowing could increase and cause a damage for the manager. From this we conclude that the treatment effect is strong evidence for arising distrust from whistleblower protection - especially from the increased probability of false claims.

In addition, it is not only of interest whether cooperation takes place, but also to which degree, we turn our focus from the overall willingness to cooperate to the level of cooperation. Since the design allows to vary the level of investment, the arising distrust may lead some managers to adjust the amount that is trusted to the employee instead of shutting down cooperation in general. To account for this, we consider only those managers who chose to cooperate and show the actual investment relative to the maximal amount possible (Figure 3.8).

The results provide a clear picture that there are no treatment differences present for the size of cooperation. Independent of the protection scheme, the trusted share lies within a range of 40 to 44 % of the endowment, which roughly corresponds to the average investment level across experimental studies (Fehr and Schmidt, 2006).

Result 9 *The trust level of those managers who cooperate remains constant independent of whistleblower protection.*

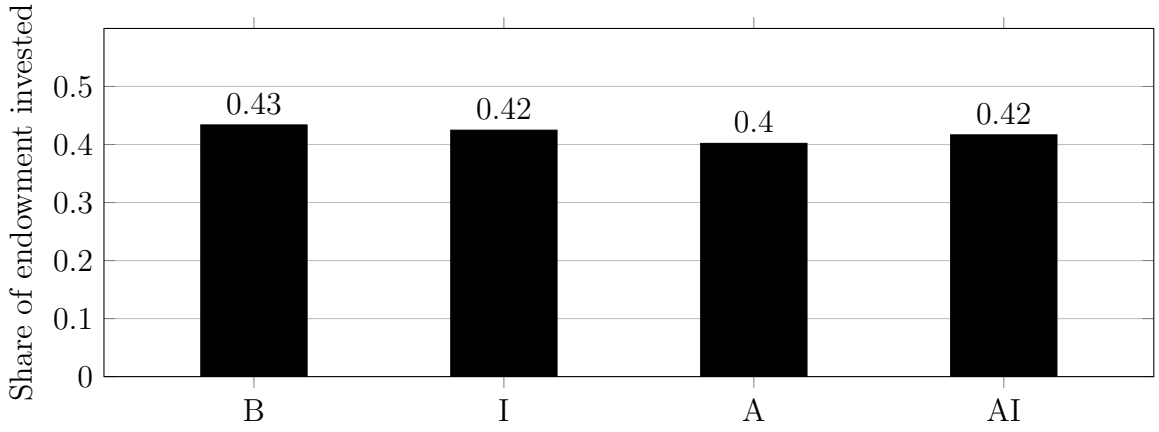


Figure 3.8: Cooperative Level Across Treatments

The results on cooperative behavior provide evidence that the willingness for cooperation depends inversely on the frequency of observed and expected whistleblowing and in particular of false claims. This supports the hypothesis that whistleblower protection can have a detrimental effect on welfare beyond the idiosyncratic costs of false claims. While we find treatment differences for the *share* of managers that are willing to cooperate, neither incentives nor anonymous reporting seems to affect the *amount* subjects are willing to invest.

Change in group payoff To evaluate the costs from forgone cooperation against the benefits from whistleblower protection, we contrast the aggregated payoffs of the groups under the different treatments in Table 3.3. The second column reports the average damage caused by embezzlement per group. Cases of undetected embezzlement result in a damage of 40 for the group, while detected embezzlement causes a loss of 30 (see Table 3.1). Comparing the treatments in this column illustrates the enhancing effect of whistleblower protection on detection and deterrence. The damage from embezzlement decreases in the treatments I (-7.6) and A (-10.0) compared to the baseline treatment and is the lowest in treatment AI (-2.4). The third column represents the average direct costs associated with false whistleblowing (each case causes a loss of 10 for the group). The numbers indicate that especially the incentives for whistleblowing also have a negative impact of welfare. The gain from reduced damage through embezzlement in treatment AI relative to the treatments I and A is in large part used up by the costs from false reports (I: -2.5, A: -1.3, AI: -5.0). The picture

is similar when the profit from cooperation is considered (column 4). While treatment I produces only moderately lower profits from cooperation per group (19.5) in contrast to the baseline treatment (23.5) and treatment I (24.1), the profit in treatment AI turns out be at only roughly 50% (12.9). In consequence, the fifth column shows that groups have on average the lowest payoff in treatment AI (increase of 5.6 compared to the endowment of 300), although it produces the lowest damage from embezzlement. While the cost and benefits from the incentives in treatment I cancel each other out compared to the baseline treatment (payoff increases by 9.4 in both treatments), the groups receive the largest payoffs in treatment A (increase of 12.8).

Treatment	**Damage from**		**Profit from**	**Total change**
	embezzlement	**false reports**	**cooperation**	
B	-13.3	-0.8	23.5	9.4
I	-7.6	-2.5	19.5	9.4
A	-10.0	-1.3	24.1	12.8
AI	-2.4	-5.0	12.9	5.6

Notes: The damage from embezzlement is calculated by the fractions of managers who commit detected embezzlement times (-30) and undetected embezzlement times (-40). The damage from false reporting results from the share of false whistleblowing cases times (-10). The profit from cooperation is calculated by the share of managers who cooperate times the average transfer times three. All numbers report experimental currency units.

Table 3.3: Average Change in Group Payoff After Embezzlement, Whistleblowing and Cooperation

3.6 Discussion

With this paper, we shed light on the potential hidden costs of whistleblower protection. In a workplace setup, a manager could embezzle money at the expense of a third party while being observed and potentially reported by her employee before they enter a trust game. We varied the framework in two dimensions to capture two prominent features of whistleblower protection laws: First, not revealing the information about the reporting decision for the manager prior to the trust game allows to provide anonymous reporting for the employee. Second, restricting the manager's choice set in the cooperation game conditionally on a report, enables the employee to insure herself against retaliation from the manager by blowing the whistle.

Our results confirm that both instruments have the desired effects. We observe an increased willingness to report truthfully illegal behavior by the employees which is anticipated by the managers inducing them to reduce illegitimate practices. This suggests that whistleblower laws offer a rich potential for fighting the damage of corporate fraud through both increased deterrence and detection. On the other hand, the findings demonstrate that also adverse effects of whistleblower protection arise. Since the incentives for reporting are not provided conditionally on a successful investigation, these do not only increase truthful reporting, but also trigger false whistleblowing by the employees.

A novel finding of this paper relates to costs associated with false reporting. Beyond the negative direct impact in the form of costs for authorities or damaged reputation, we point out the importance of observed and expected whistleblowing as unkind behavior for the cooperative climate in a organization – especially false whistleblowing. An increased frequency of unkind behavior may cause an atmosphere of distrust and hampers productive cooperation. In consequence, social welfare may be negatively affected by whistleblower protection although it deters misbehavior.

We chose a simple design for the whistleblowing game, where the employee has precise knowledge about the state of illegal behavior of her superior. Also the employee does not face the risk of leaks under anonymity and investigation as well as incentives are guaranteed consequences of a report. This captures the intended increase in legal certainty for the whistleblower. In reality, when for some laws not all of these assumptions are met, uncertainty may also influence the behavior under the different protection regimes and cause a lower responsiveness of employees (see e.g., Chassang and Miquel, 2018; Mechtenberg, Muehlheusser, and Roider, 2017). However, our approach has the advantage that the results are not driven by ambiguity or risk aversion and serve as a benchmark for future studies that relax these assumptions.

The results of our study provide some implications for the design of a whistleblower protection law. The benefits of reduced fraud may not only be evaluated against inspection and reputational costs arising from false claims, but hidden costs from forgone cooperation have to be taken into account as well. Therefore, a legislator could pass different tailor-made laws for different sectors, since the importance of cooperation may well vary between the industries of an economy. For example, laws that apply for organizations where efficiency is rather driven by compliance than by cooperation, could use strong whistleblower protection to drive down misbehavior of

the management. In contrast, if a company's success heavily depends on productive cooperation, the policy could acknowledge this by avoiding an excessive amount of false claims at the cost of non-maximal deterrence. The results on the cooperative behavior suggest that a law could avoid an "atmosphere of distrust" if the incentives for false claims are not too strong. This could be achieved by providing either anonymity or incentives for reporting, or alternatively, conditioning further incentives on a successful investigation (compare to Mechtenberg, Muehlheusser, and Roider, 2017).

3.A Translated Instructions

Welcome to today's experiment! If you read the following instructions carefully, you can earn a significant payment - depending on your decisions.

Please note, that from now on and during the whole experiment no communication is allowed. If you have any questions, please direct these at one of the experimenters. Neglecting these rules result in exclusion from this experiment and all payments.

All your decisions during this experiment will remain anonymous and cannot be related to you by either the experimenters nor the fellow subjects. Your earnings will be accounted in points. The points you acquire during this experiment will be exchanged for euro at the end. The exchange rate is: **10 points = 50 eurocent**.

General procedure:

There are **three roles** in this experiment: *Manager*, *employee* and *a third party*. These roles are assigned randomly. If you are drawn into the role *manager*, you'll maintain this role throughout the entire experiment. If you start with one of the other two roles, your role will be drawn randomly before each period. In each period you are part of a group consisting of exactly one manager, one employee and one third party. Also the group composition will result from a random draw in every period.

The experiment is divided into two parts consisting of multiple periods. Beneath you find the procedure of a period in part 1. For the second part, you'll receive instructions on your screen immediately before it starts.

Procedure of period in part 1:

Every subject is endowed with 100 points. After the roles are assigned, the manager chooses between two alternatives (CIRCLE or TRIANGLE). CIRCLE has no payoff consequences for any member of the group. TRIANGLE represents violating the law, resulting in a gain (50 points) for the *manager*, and a loss (90 points) for the *third*

party. Again, there are no consequences for the *employee*.

After the manager has made her choice about CIRLCE and TRIANGLE, the employee has to decide whether she wants to file a complaint. This decision is taken separately for both alternatives (complaint if CIRCLE was chosen; complaint if TRIANGLE was chosen). Filing a complaint causes costs for the manager in any case (10 points). If CIRCLE has been chosen and complaint has been filed, the manager has to pay an additional fine (60 points). The third party receives partial compensation for her damage (80 points).

The table below displays all possible combinations of the decisions made by the manager and the employee as well as its respective payoffs for all group members.

Manager chooses alternative	*Employee* files a complaint	Payoffs		
		Manager	*Employee*	*Third Party*
Circle	No	0	0	0
Circle	Yes	-10	0	0
Triangle	No	50	0	-90
Triangle	Yes	-20	0	-10

Subsequently, all group members are informed about the chosen alternative[and whether there has been a complaint].

To conclude a period the manager and the employee play an investment game. First, the manager chooses an amount x between -30 and 60 points. Negative figures mean that points are taken from the employee. Positive mean that points are sent to the employee. If the manager deducts points from the employee these points are transferred and the investment game ends. If the manager sends a positive amount to the employee, it will be multiplied by three. In this case, the employee chooses an amount y between 0 and $3 \cdot$ x which she would like to return to the manager. There are no consequences for the third party in the investment game.

Payoffs in the investment game:

Manager = - x + y points,

Employee = max(x, 3· x) - y points,

Third party = 0.

At the end of a period[all of the group members are informed whether there was a complaint and] your surplus adds up from your **endowment** (100 points), **your revenue from the decisions made** (see table) and **your revenue from the investment game**.

Summary of a period in part 1

1. Manager chooses alternative CIRLCE or TRIANGLE (violation of law)
2. Employee decides upon reporting
3. Every member of a group learns about the chosen alternative [and the reporting decision]
4. Manager and employee engage in an investment game

(5. Every member of a group learns about the reporting decision)

5./6. The surplus is computed

After you have completed the second part and a questionnaire, **one period** is drawn for payout. You'll receive the points you earned in that period converted according to the exchange rate plus 5 euro as show up fee.

Thank you for participating and good luck!

3.B Control Questions

1. Do you keep your role through the entire experiment?

   ☐ Yes, always.

   ☐ No, my role is randomly drawn in each period.

   ☐ Yes, in case I am an manager. If I am an employee or the third party, it may change from period to period.

2. Do you have the same members in your group over several periods?

   ☐ No.

   ☐ Yes, in the second part of the experiment.

   ☐ Yes, always.

3. If the manager chooses TRIANGLE, ...

   ☐ she receives a profit and harms the employee as well as the third party.

   ☐ she does not receive a profit, but harms the employee as well as the third party.

   ☐ she receives a profit and harms the third party, but not the employee.

4. If the manager chooses CIRCLE and the employee files a report, ...

   ☐ all payoffs are unaffected.

   ☐ it causes a cost for the manager. Both the employee and the third party are not affected.

   ☐ it causes a cost for the manager. Both the employee and the third party receive a profit.

5. If the manager sends 30 points in the investment game, how many points does the employee receive?

   _____

3.C Questionnaire

Demographics

1. How old are you? _____
2. What is your sex? ☐ Male ☐ Female
3. What are you studying? _____
4. How much work experience do you have?
   (a) Internships (in month): _____
   (b) Full-time (in month): _____
   (c) Student jobs (in month): _____

Risk preferences

1. Imagine you had won 100,000 euros in a lottery. Almost immediately after you collect, you receive the following financial offer from a reputable bank, the conditions of which are as follows: There is the chance to double the money within two years. It is equally possible that you could lose half of the amount invested. What fraction would you choose to invest?

   ☐ 0 ☐ 20,000 ☐ 40,000 ☐ 60,000 ☐ 80,000 ☐ 100,000

Attitudes towards whistleblowing

1. What is your opinion with respect to the following claims?

   (a) A person should be supported in disclosing serious misbehavior, even if this requires disclosure of insider information.

   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree

   (b) A person should be supported in disclosing already mild misbehavior, even if this requires disclosure of insider information.

   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree

   (c) I would disclose serious misbehavior, even it would cause disadvantages for me.

   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree

   (d) I would disclose already mild misbehavior, even it would cause disadvantages for me.

   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree

   (e) If the chance is larger that misbehavior is detected it could be deterred.

   ☐ Strongly agree ☐ Agree ☐ No opinion ☐ Disagree ☐ Strongly disagree

2. In your opinion, how acceptable are the following actions?

   (a) Disclosing insider information about serious misbehavior by person in authority of an organization.

   ☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable

(b) Disclosing insider information about serious misbehavior by regular employees of an organization.

☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable

(c) Disclosing insider information about serious misbehavior by a friend or family member of an organization's member.

☐ Very acceptable ☐ Acceptable ☐ Neither, nor ☐ Unacceptable ☐ Very unacceptable

3. Imagine you had insider information about serious misbehavior in an organization you are a member of. How important was each of the following items for the decision to tell someone about it?

(a) Persons in authority would support me.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

(b) I would be legally obliged to report.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

(c) Somebody would act to end the misbehavior.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

(d) Only people I choose would know my identity.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

(e) Apart from the people I contact, the information would remain confidential.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

(f) I would remain completely anonymous.

☐ Very important ☐ Important ☐ Neither, nor ☐ Unimportant ☐ Very unimportant

3.D Cooperation Given Embezzlement and Reporting Across Treatments

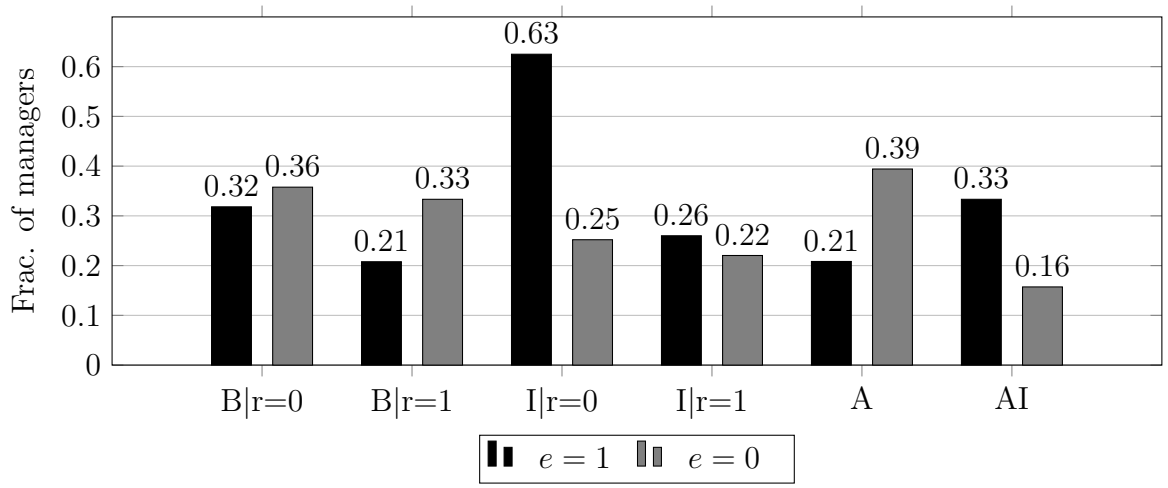


Figure 3.D.1: Cooperation Given Embezzlement and Reporting Across Treatments

Chapter 4.

Gender Differences in Honesty: Groups Versus Individuals[1]

Abstract

Extending the die rolling experiment of Fischbacher and Föllmi-Heusi (2013), we compare gender effects with respect to unethical behavior by individuals and by two-person groups. In contrast to individual decisions, gender matters strongly under group decisions. We find more lying in male groups and mixed groups than in female groups.

Keywords: unethical behavior, lying, group decisions, gender effects, experiment.

JEL Codes: C91, C92, J16.

4.1 Introduction

Unethical behavior is a ubiquitous feature in many economic contexts, and a number of recent experimental studies have analyzed lying as one prominent type of unethical behavior. For example, in Fischbacher and Föllmi-Heusi (2013) individuals are asked to report the (privately observed) realization of a die roll that determines their payoff. Evidence for lying (on the aggregate level) is then obtained by comparing the actual payoff distribution with the uniform distribution, which would result under truth-telling. Other studies have analyzed lying using the sender-receiver setup of Gneezy (2005). All in all, there is strong evidence for lying, but often not to the maximal extent possible; suggesting that there are private costs associated with such unethical behavior (Gneezy, 2005; Charness and Dufwenberg, 2006; Erat and Gneezy, 2012; Gibson, Tanner, and Wagner, 2013).

With respect to gender differences, it seems that males are somewhat more prone to lying than females, but often the effect is small or not statistically significant (Dreber and Johannesson, 2008; Childs, 2012; Erat and Gneezy, 2012; Houser, Vetter, and

[1]This chapter is co-authored by Gerd Mühlheußer and Andreas Roider and has been published as Muehlheusser, Roider, and Wallmeier (2015) in *Economics Letters*.

Winter, 2012; Conrads, Irlenbusch, Rilke, and Walkowitz, 2013; Conrads, Irlenbusch, Rilke, Schielke, and Walkowitz, 2014; Abeler, Becker, and Falk, 2014).[2]

So far, the literature on lying behavior has mainly analyzed decisions by *individuals*; possibly in strategic interaction with other individuals as in tournaments (see e.g., Conrads, Irlenbusch, Rilke, Schielke, and Walkowitz, 2014). However, in many settings, a *group* of individuals must reach a decision *jointly*, e.g., decision-making by committees in economic, social, or political organizations. In fact, there is growing evidence from contexts other than lying that groups often decide markedly different than individuals (for surveys, see Charness and Sutter, 2012; Kugler, Kausel, and Kocher, 2012). On the one hand, groups are better at solving cognitive tasks and act more selfishly (see e.g., Maciejovsky, Sutter, Budescu, and Bernau, 2013; Bornstein, Kugler, and Ziegelmeyer, 2004; Falk and Szech, 2013). That suggests that groups might be more willing to realize the potential monetary gains from lying. On the other hand, there is evidence that "moral reminders" reduce dishonesty (Pruckner and Sausgruber, 2013). Hence, discussions within groups might lead them to lie less. Taken together, it seems a priori unclear whether lying is more prevalent in groups compared to individuals. Moreover, for the lying behavior of groups their gender composition might matter (see e.g., Dufwenberg and Muren, 2006, where gender composition affects groups' giving in a dictator game). Consequently, this paper aims at providing insights on the unethical behavior of groups and individuals, and the role of gender in this context. Gender composition is found to be particularly important under group decision-making. In our view, this has interesting implications for the design of decision-making (and monitoring) processes in organizations.

4.2 Experimental Design

We extend the simple and widely used die rolling experiment of Fischbacher and Föllmi-Heusi (2013), where subjects decide autonomously and anonymously about their (lying) behavior, to a setting where decisions are made jointly in groups. We consider a treatment G where randomly formed groups of two subjects need to coordinate on both who rolls the die and on which realization to declare. As a control treatment I, we replicate the setup of decision-making by individuals as in Fischbacher and Föllmi-

[2]For surveys on gender differences in a variety of economic contexts, see e.g., Eckel and Grossman (2008) and Croson and Gneezy (2009).

Heusi (2013). Subjects were randomly assigned to treatments (and in treatment G, to groups).

The experiment was conducted at the University of Regensburg in June 2014. Participants were recruited through an introductory undergraduate course in economics (economics majors and minors and business majors).[3] Subjects were first asked to complete an unrelated questionnaire inside the lecture hall. They were instructed (i) that their payoff for filling out the questionnaire would be either 0, 1, 2, 3, 4, or 5 euros, and (ii) that the exact amount would be determined in a second phase of the experiment outside the lecture hall, where they would receive further instructions. We made it clear that payoffs would be completely independent from their answers in the questionnaire, and that their behavior in the experiment would remain anonymous.

The die rolling experiment was then played in paper-pencil style in fifteen booths outside the lecture hall that ensured complete privacy of decision-making. Subjects waited inside the lecture hall at their seats, and were only allowed to proceed outside when booths became vacant. Inside the booth, subjects found a fair, six-sided die, a pen, instructions, an anonymous answer sheet (on which the realization of the die roll was to be declared), and a receipt form for each subject. Translations of the instructions and the answer sheet are included in the Supplementary Material. As each booth contained one die and one answer sheet only, in treatment G, subjects had to make a joint declaration, and they were aware that *each* of them would receive the declared payoff.[4] Afterwards, subjects proceeded to the cashier desk. They handed in the anonymous questionnaire(s) and the anonymous answer sheet, where it was checked that the declared amounts coincided with those on the receipt form(s). Then each subject went to privately collect his/her payment. As in Fischbacher and Föllmi-Heusi (2013), subject i's payment (in euros) π_i, is related to the declared outcome of the die roll $r \in \{1, ..., 6\}$ as follows: $\pi_i = r$ for all $r \leq 5$ and $\pi_i = 0$ for $r = 6$. In total, there were 228 participants (124 female, 104 male) of which 108 (120) participated in treatment I (G). The whole experiment took about 2 hours.

[3] As a show-up fee, students who agreed to participate (which all did) received a small bonus towards their final exam.

[4] As participants still had to read the instructions in the booth, they did not need to worry that the time they spent there might be indicative of lying.

4.3 Results

Table 4.3.1 summarizes the distribution of payoffs in the two treatments. In line with the previous literature, a sizeable amount of lying also occurs in our setting. First, the average payoffs in treatments G and I are 3.47 and 3.48, respectively. Hence, they virtually take the same value (3.51) as in the baseline (individualistic) treatment of Fischbacher and Föllmi-Heusi (2013). Both payoff distributions differ significantly from the uniform distribution that would result under truthful reporting leading to an average payoff of 2.50 ($p < 0.001$, two-sided one-sample Kolmogorov-Smirnov (KS) tests). These results are driven mainly by the high frequency of reported 4's and 5's. Comparing our two treatments reveals that - when considering all observations - their payoff distributions do not differ significantly at conventional levels according to a two-sided Mann-Whitney U (MWU) test.[5] However, as shown next, this result masks substantial gender differences. As displayed in Figure 4.3.1(a), in treatment I, the average payoff is somewhat higher for male subjects (3.58) than for female subjects (3.40), and both gender-specific payoff distributions differ significantly from the uniform distribution ($p < 0.001$, two-sided one-sample KS tests). Hence, females are somewhat less prone to lying than men, but the difference is not statistically significant ($p = 0.477$, two-sided MWU test). Based on own calculations, this is again very similar to the baseline treatment of Fischbacher and Föllmi-Heusi (2013), where the respective gender-specific values are 3.60 and 3.37 with $p = 0.133$.

The (slight) tendency of females to lie less than males is, however, amplified in treatment G, where we observe groups that are either "female" (only females), "male" (only males), or "mixed" (one female, one male). As illustrated in Figure 4.3.1(b), compared to treatment I, the average payoff of female groups decreases (to 2.74), while the average payoff of male and mixed groups increases (to 4.00 and 3.71, respectively). Payoffs of female groups are significantly lower than payoffs of male groups or mixed groups (pair-wise two-sided MWU tests with $p = 0.045$ and $p = 0.059$, respectively). The payoffs of male groups and mixed groups are not significantly different from each other (two-sided MWU test, $p = 0.497$). A Jonckheere-Terpstra test indicates that the extent of lying is lowest for female groups followed by female individuals, male individuals, and male groups ($p = 0.026$, two-sided). In fact, while the payoff distri-

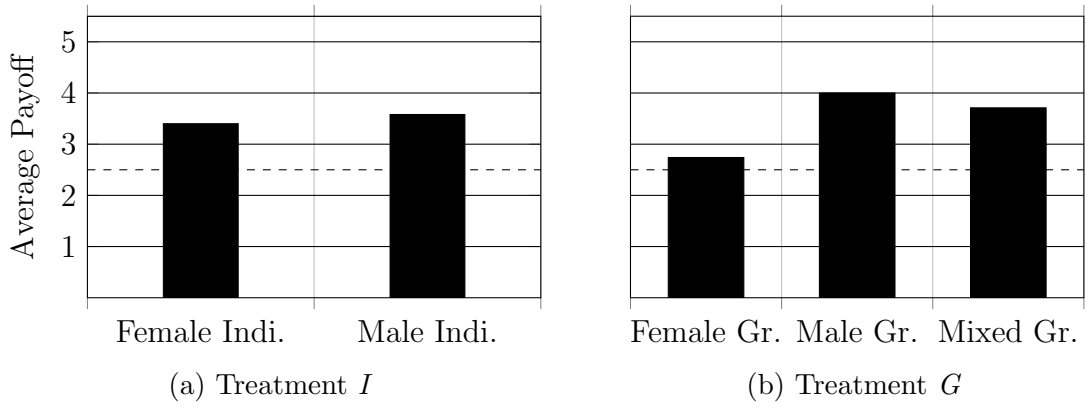
[5] Chytilova and Korbel (2014) conduct an artefactual field experiment on lying with children and adolescents at a high school, where participants were paid in sweets. Their three-person groups obtain a somewhat higher payoff than individuals (3.28 and 2.93, respectively).

Treatment	n	$\overline{\pi}$	$\pi_i = 0$	$\pi_i = 1$	$\pi_i = 2$	$\pi_i = 3$	$\pi_i = 4$	$\pi_i = 5$
I (all individuals)	108	3.48	$.08^{--}$	$.06^{---}$	$.09^{--}$	.19	$.28^{+++}$	$.31^{+++}$
I (females only)	58	3.40	$.05^{--}$	.10	.09	.22	.22	$.31^{+++}$
I (males only)	50	3.58	.10	$.00^{---}$	.10	.14	$.34^{++}$	$.32^{+++}$
G (all groups)	60	3.47	$.05^{--}$	.10	.12	.17	.20	$.37^{+++}$
G (female gr. only)	19	2.74	.16	.11	.21	.11	.21	.21
G (male gr. only)	13	4.00	.00	.08	.08	.15	.15	$.54^{+++}$
G (mixed gr. only)	28	3.71	$.00^{---}$	.11	.07	.21	.21	$.39^{+++}$

Note: n and $\overline{\pi}$ indicate the number of observations and the average payoff, respectively. A minus (plus) sign displays the significance of a two-sided binomial test indicating that the observed relative frequency is smaller (larger) than $\frac{1}{6}$: $^{-}$ ($^{+}$) = 10%-level, $^{--}$ ($^{++}$) = 5%-level, $^{---}$ ($^{+++}$) = 1%-level.

Table 4.3.1: Summary of Payoffs

butions of both male groups and mixed groups differ significantly from the uniform distribution, which would obtain under truthful reporting (two-sided one-sample KS tests, each with $p = 0.001$), this is not the case for female groups ($p = 0.311$). That is, in contrast to individuals (either female or male), male groups, or mixed groups, one cannot reject that there is no lying in female groups.



(a) Treatment I (b) Treatment G

Note: The dotted line indicates a payoff of 2.50, which would obtain on average under truthful reporting.

Figure 4.3.1: Average Payoffs

There are also interesting gender differences with respect to the extent of lying, which

we study by looking at the relative frequencies of 4's and 5's.[6] First, we compare the behavior of male individuals and male groups, where similar fractions report either 4 or 5 (0.66 and 0.69, respectively). However, as illustrated by Figure 2, the fractions of male individuals who report 4 respectively 5 are almost identical. In contrast, male groups more often report 5 (in 54% of cases) than 4 (in 15% of cases), where this difference is significant at the 10%-level of a one-sided binomial test that presumes that 4 and 5 occur with equal probability ($p = 0.0898$). Second, from comparing female individuals and female groups a different picture emerges. From Figure 4.3.2, if anything, female groups are less likely to report 5's than female individuals (and in treatment I (G) one cannot reject that 4's and 5's are reported by equal fractions of female individuals (female groups)). Finally, mixed groups seem to be more similar to male groups than to female groups, as there are more 5's than 4's in mixed groups (where the p-value of a respective one-sided binomial test is, however, only 0.1662).

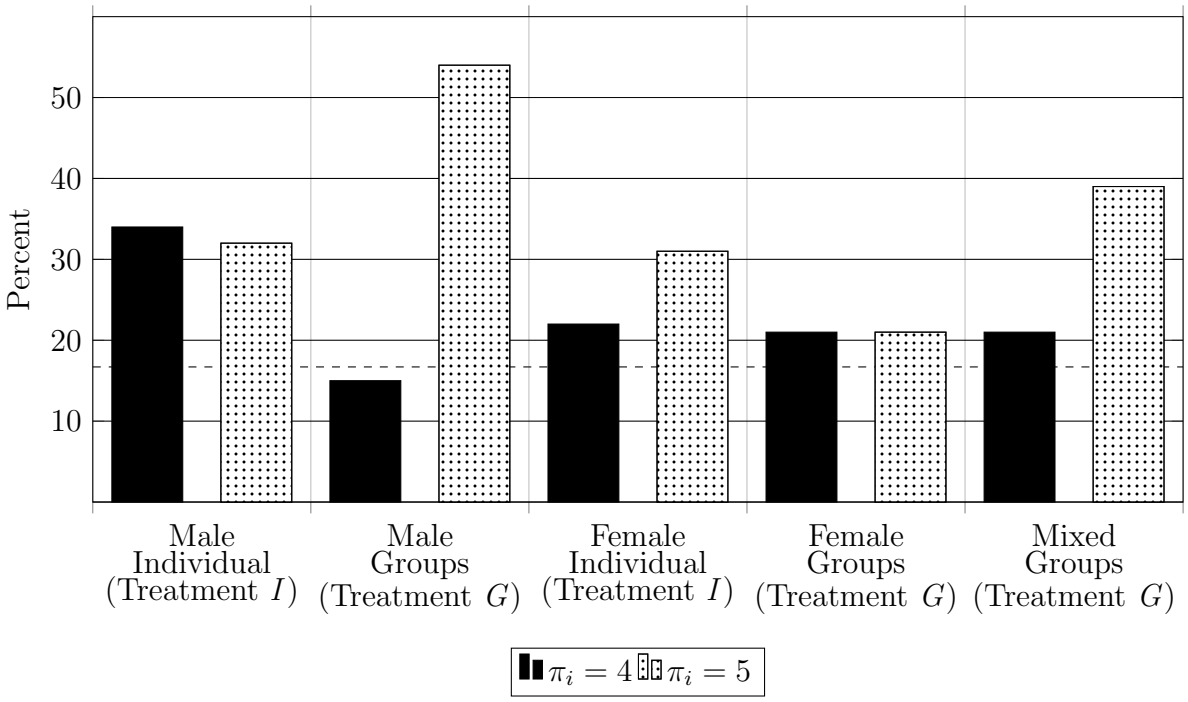


Figure 4.3.2: Frequencies of 4's and 5's by Gender and Treatment

[6]In principle, subjects might also lie to their own disadvantage. However, at an aggregate level, for $\pi_i \leq 3$ none of the frequencies reported in Table 4.3.1 are significantly above the truth-telling benchmark 1/6.

4.4 Discussion

Many important economic, social, or political decisions are taken by groups rather than individuals. We investigate how gender affects unethical behavior in the form of lying. In line with the previous literature, we find no clear evidence for gender differences under individual decision-making on lying. In contrast, in the case of group decision-making, more pronounced gender effects arise; resulting in more (less) aggregated unethical behavior in male (female) groups. Moreover, male groups seem to have a greater tendency towards exploiting the full gains from lying (i.e., securing the maximum payoff of 5) than female groups. Finally, mixed groups with equal shares of males and females behave similarly to male groups. Hence, from the viewpoint of organizational design, our results suggest that in contexts where unethical behavior might be an issue, designers might want to pay particular attention to decisions that are taken by purely male (or male-dominated) groups.

In future research, it would be interesting to study in more detail the forces underlying our results. In this respect, it is well known from other experimental settings that culture and cultural context might be important factors (see e.g., Roth, Prasnikar, Okuno-Fujiwara, and Zamir, 1991; Cox, Lobel, and McLeod, 1991; Herrmann, Thöni, and Gächter, 2008). Cultural context might also be of relevance in our context of unethical behavior, and at least the following two channels might be at work: First, in the light of the discussion in Section 4.1, culture might influence how groups reach joint decisions as compared to individuals. For example, culture might affect how much group members are concerned with their "image" relative to the potential gains from acting dishonestly. Second, also gender differences in behavior might be driven by cultural context. Again, such effects have already been documented in other settings (see e.g., Finucane, Slovic, Mertz, Flynn, and Satterfield, 2000; Gneezy, Leonard, and List, 2009), and they might also be relevant in the context of unethical behavior.

4.A Supplementary Material: Instructions and Answer Sheet

As Supplementary Material, we provide translated versions of the instructions (Section 4.A.1) and the answer sheet (Section 4.A.2) provided inside the booths for treatment *I*. In square brackets, we state the respective adjustments in the text for the two

treatments I and G. The original versions (in German) are available upon request.

To avoid waiting lines at the cashier desks, we had two of them marked with green and red signs, respectively. We used otherwise identical, randomly allocated answer sheets and receipt forms in these colors (in equal proportions) and asked subjects to go to the corresponding color-coded cashier desk after they had completed the experiment.

4.A.1 Instructions

Please read the entire instructions first before you roll the die!

For participating in this experiment [G: each of] you will receive a payoff.

For organizational reasons, the color of your documents determines at which cashier desk you will receive your payoff: If your documents are in red, [G: both of] you may go to the red cashier desk after finishing the experiment. If they are in green, then [G: both of] you may go to the green cashier desk after finishing the experiment.

The payoff that [G: each of] you will receive will be determined by rolling a die:

Outcome of die roll	1	2	3	4	5	6
Payoff in euros	1	2	3	4	5	0

[I: Please roll the die in front of you once.] [G: Please agree upon who of you will roll the die in front of you once.] After that, please circle the outcome of the die roll and the related payoff on the answer sheet. You are free to roll the die more than once, but only the first roll is relevant for your payoff.

In a next step, we ask [G: each of] you to fill out and sign your receipt form (name and payoff) in line with your entry on the answer sheet.

[G: Together] please submit all documents ([I: questionnaire, answer sheet, receipt] [G: both questionnaires, answer sheet, receipts]) at the respective cashier desk. [G: Each of] you will receive [I: your] [G: his/her] payoff there.

If you have any questions, please contact a member of the support team. If not, please roll the die now.

Thank you for your participation!

4.A.2 Answer Sheet

Please circle the combination of the outcome of the die roll and the corresponding payoff:

Outcome of die roll	1	2	3	4	5	6
Payoff in euros	1	2	3	4	5	0

Chapter 5.

An Experiment on Peer Effects Under Different Relative Performance Feedback and Grouping Procedures[1]

Abstract

We conduct a laboratory experiment to test theoretical predictions about subjects' performance in an effort task conditional on their peer group's composition and relative performance feedback. Subjects are grouped either randomly or according to their ability, with the feedback being the best or average performance of their group. While theory-derived hypotheses on aggregate treatment differences cannot be confirmed, we find evidence when gender differences are taken into account. Male subjects perform significantly better when they compare themselves with the best peer instead of the average, while the opposite is true for females. With respect to the grouping treatment, we find that random grouping is beneficial for male subjects, and ability grouping for female subjects. These differences are explained by gender differences in (non-linear) reactions to the reference point and an aversion of females to competitive environments.

JEL Classification: C91, J16, J24, M52

Keywords: Laboratory Experiment, Ability Grouping, Relative Performance Feedback, Peer Effects, Reference Dependent Preferences

5.1 Introduction

In many areas of life, performance feedback is leveraged to induce a change in behavior. For instance, 60% of manufacturing firms reveal performance data to their employees (Bloom and Van Reenen, 2007). Assuming that individuals have reference dependent-preferences (see e.g. Tversky and Kahneman, 1979; Kőszegi and Rabin, 2006), the effect of recognition can depend on the kind of relative performance feedback and on

[1]This chapter is co-authored by Kathrin Thiemann.

the performance distribution in the reference group. This study is the first to test the peer effects for different grouping procedures on performance when the relative performance feedback is exogenous in a theory-guided lab experiment. The results suggest significant gender differences in the reaction to different reference points and grouping procedures as well as peer effects to be non-linear.

We consider different kinds of performance feedback since it may vary with an organization's philosophy. Some firms actively highlight only the top performers (e.g. the "employee of the month", Kosfeld and Neckermann, 2011) in order to motivate employees to perform better. The reference point can also vary with culture as acquired by groups of people that share a religion or ethnic origin. In more competitive cultures individuals may be expected to compete for the top positions. Opposite, in less competitive cultures, social comparison may play a less emphasized role and individuals are expected to compare themselves to the average. Therefore, the question arises whether group composition can be optimized for a given performance feedback in order to maximize group performance.

Thiemann (2017) addresses this issue theoretically, focusing on the question whether ability-segregated classes (also referred to as *ability tracking* or *ability grouping*) or classes with students of heterogeneous ability are preferable. The above-mentioned theory predicts that it depends on the culture of competitiveness of the student body, that is, on the kind of the reference point and the importance of social comparison. The intuition is that a comprehensive school, i.e. a class with heterogeneous students, yields optimal incentives for highly competitive individuals, who want to be the best student in class. Here also subjects with very low ability are motivated by the best student and they exert effort in order to minimize the performance distance. In a system with ability grouping, where high-ability subjects are sorted into a high track and low-ability subjects into a low track, the low-ability students can only compare with the top performer in their class, which is less motivating. When students are less competitive and only compare their performance to that of the average student, the model predicts ability grouping to be optimal. This is driven by stronger motivation in a high-ability group due to the higher reference point compared to a more heterogeneous group. This effect may on average outweigh the negative effect of ability grouping for low-ability students.

In our experiment, we test these predictions and consider environments where subjects perform a real-effort task while they are evaluated either against the *average* or

the *best* performance of their reference group. To affect the ability distribution within the reference group, the members are drawn randomly either from the *entire* pool of participants or only from those of the same ability category (*high* or *low*).

We also test the role of gender concerning the optimal performance feedback and grouping policy, something not considered in Thiemann (2017). Gender might be of importance, since women and men have been found to differ to a huge extent in their preferences for competitiveness (see e.g. Gneezy, Niederle, Rustichini, et al., 2003; Niederle and Vesterlund, 2007; Niederle, 2016). In our experiment, high reference points and pressure for social comparison will create a competitive environment and might cause different effort choices of male and female subjects. The existing research generally finds that men perform better in competitive environments (e.g. tournaments), whereas women's performance does not change in a tournament-based compensation scheme compared to a piece rate.

Our results support that male subjects behaving according to our model's prediction of optimal performance. While we cannot confirm hypotheses on aggregated treatment effects, we find significant gender differences in the reaction to different reference points and grouping procedures. On the subject level, regression analysis suggests that impact of a reference performance differs conditional on whether the best or the average performance is available. Furthermore, these peer effects seem to be non-linear in the distance between an individual's performance and the reference point.

Our study contributes to two fields of economic literature. The first is the empirical literature on peer effects (For an overview see Herbst and Mas, 2015). Recent experimental studies find the mere possibility of being evaluated relative to peers as performance enhancing (Kuhnen and Tymula, 2012) and that this effect is larger for male than for female subjects (Beugnot, Fortin, Lacroix, and Villeval, 2013). Furthermore, peer effects seem to be non-linear in the distance between a subject's performance and the reference point (Gill, Kissová, Lee, and Prowse, 2018). While these studies focus on a single performance feedback, we contrast the effects of different relative performance feedback: the *average* peer achievement and the *best* peer performance.

Second, our study contributes to the literature that addresses the effect of grouping individuals according to their ability. These effects can arise from mutual learning or norm setting within the group. The latter corresponds to the pure peer effect analyzed in lab experiments. A number of field studies have analyzed the influence of ability tracking on student performance in school (see the surveys by Slavin, 1990;

Meier and Schütz, 2008). Effects of ability tracking on mean achievement are usually low and non-significant. Studies usually find that tracking harms low-ability students but benefits high-ability students (e.g. Argys, Rees, and Brewer, 1996; Duflo, Dupas, and Kremer, 2011). Our approach offers two advantages. First, to identify the effect of grouping on performance the assumption of identical resources in both kind of groups has to be met. This is guaranteed in the lab, while it is often violated in practice. For example, the experience or the qualification of teachers may vary between tracked and non-tracked classes (see e.g. Betts and Shkolnik, 2000; Rees, Brewer, and Argys, 2000). Second, the field studies cannot disentangle whether different group compositions affect performance through mutual learning or through different group norms. In the laboratory setting we can exclude mutual learning effects and focus on the latter.

The remainder of the paper is organized as follows. Section 5.2 introduces our theoretical framework and provides the hypotheses. Section 5.3 lays out our experimental design. Section 5.4 reports the results from our experiment on aggregated performance, gender differences optimal performance and the linearity of peer effects. Section 5.5 concludes by discussing limitations and implications.

5.2 Theory

In line with Thiemann (2017) we assume that subjects in our experiment maximize utility by choosing an effort level. Further we assume that effort translates linearly into performance and that subjects have reference-dependent preferences as in Tversky and Kahneman (1979) with relative performance feedback being the reference point. Then subjects face the following optimization problem:

$$\underset{p_i}{\text{Max}}\ u_i(p_i) = (1 - s_i)p_i + s_i \cdot v(p_i - r_i) - c(p_i, a_i) \tag{5.1}$$

$$\text{with}\ \ v(p_i - r_i) = \begin{cases} \lambda \cdot (p_i - r_i) & if\ p_i < r_i \\ (p_i - r_i) & if\ p_i \geq r_i \end{cases} \tag{5.2}$$

$$\text{and}\ \ c(p_i, a_i) = \frac{p_i^2}{2a_i} \tag{5.3}$$

Performance p_i is the number of correctly answered multiplication problems per

period. Before each period, each subject is shown a reference point r_i, that yields information about the performance of the group members. Subjects' utility depends on a direct private component of utility and a comparison oriented component given by the value function $v(\cdot)$. In the experiment the direct private utility from performance is given because of direct remuneration of performance. The utility from the comparison oriented component is assumed to be larger the more competitive a subject is (s_i, with $s_i \in [0, 1]$ is the degree of social comparison). For subjects performing below the reference point, the disutility from the difference to the reference performance is increasing with loss aversion, λ, with $\lambda > 1$. The cost of performance $c(p_i, a_i)$ increases in performance and decreases with ability a_i. A subject's optimal performance is then given by the following best response function:

$$BR_i(r_i) = \begin{cases} (1 - s_i + \lambda s_i)a_i & if \quad p_i < r_i \\ r_i & if \quad p_i = r_i \\ a_i & if \quad p_i > r_i \end{cases} \tag{5.4}$$

Optimal performance depends positively on ability a_i in case the subject's performance is above or below the reference point. If the subject's performance is below the reference point, performance also depends positively on loss aversion (λ) and the degree of social comparison (s_i). In the case where performance equals the reference point, the best response is to stay on this level of performance.

The derived best response function is the basis to compare equilibrium performances across different regimes. First, we compare performances for different reference points: the *average* performance among the other group members and the *best* performance among the other group members. Second, we compare a regime where subjects are randomly grouped with a regime, where subjects are grouped according to ability. In the latter we have groups consisting only of low-ability subjects and groups only with high-ability subjects. We follow the theoretical analysis of Thiemann (2017), where proof is found for four main hypotheses:

H1 *When the best reference point is given, average performance is higher under random grouping than under ability grouping.*

H2 *When the average reference point is given, average performance is higher under ability grouping than under random grouping, if s_i and λ are sufficiently low.*

H3 *Low-ability individuals always perform lower under ability grouping than under random grouping.*

H4 *High-ability individuals benefit from ability grouping when the average reference point is given, and are not affected when the best reference point is given.*

In addition to these formally derived hypotheses we also want to investigate the role of gender. Thiemann (2017) assumes that all individuals in a cultural group have the same degree of social comparison, loss aversion and reference point. Past research, however, has shown that preferences for competition differ to a high extent with gender. In particular women have been found to be more averse to competition than men, i.e. they shy away from entering a competition and perform worse in competitive environments (e.g. Croson and Gneezy, 2009; Niederle and Vesterlund, 2007). Although our setting does not provide monetary incentives for top-ranked performances, an environment where the best performance of a group serves as reference point might still be perceived as a more competitive environment. Also, Jones and Linardi (2014) found that women often show an aversion to standing out, i.e. when their actions are visible they prefer to behave close to the average person. Hence, women would rather choose a reference point at the average to compare with and might not thrive to be the best in the *best* treatment. Also, if women are indeed less competitive, this would imply a lower degree of social comparison in the sense of the model. However, there is psychological evidence that women engage slightly more in social comparison than men (Gibbons and Buunk, 1999; Guimond et al., 2007).

To summarize, there is evidence from the literature indicating that women and men behave differently in competitive environments. Since some of the evidence contradicts the assumptions of the model, we cannot formulate alternative hypotheses for men or women based on this theoretical framework. We will instead resume to exploratory evidence in Section 5.4.2 and analyze the data for men and women separately to see whether differences in behavior are evident.

5.3 Experimental Design

5.3.1 Real Effort Task

In each period, subjects were asked to solve as many multiplication problems of one-digit numbers (3-9) and two-digit numbers (11-99) (see Dohmen and Falk, 2011) as possible within four minutes. Subjects faced linear incentives with a piece rate of 30 eurocent per solved problem. This remuneration scheme does not vary with the feedback and grouping treatments. At the end of the experiment, one period was chosen randomly for remuneration. Every subject was given the same problems in the same order to ensure that the difficulty of the problems was identical. Problems were purposefully designed such that the difficulty would vary to the same extent within each period. In case subjects answered a problem incorrectly, the screen reported "false" and subjects had to repeat it instead of searching for easy problems.[2]

Multiplication problems were chosen as an effort task to ensure that performances during the experiment depend both on ability and effort. On the one hand the given task is a good proxy for cognitive ability and generates heterogeneous outputs that allow for grouping according to ability. On the other hand the task offers sufficient scope to vary effort, since solving the problems needs high concentration and is thus costly. Dohmen and Falk (2011) use this very task in their experiment that compares performance under fixed and variable payment schemes. They find substantial differences in performance. However, since subjects could select into the schemes, the differences may be due to heterogeneous abilities. In favor of a certain elasticity, they report higher levels of effort, stress and exhaustion in the variable payment scheme. Further evidence in favor of elastic responses to incentives is provided by Brüggen and Strobel (2007). This study finds that the number of solved multiplication tasks increases when monetary incentives are higher.

5.3.2 Treatments and Procedural Details

In order to test hypotheses ***H1*** and ***H2***, we implement a two-by-two design to compare mean group performances along the two major treatments: *best* vs. *average* reference point and *ability grouping* vs. *random grouping* (see Table 5.3.1). To test hypotheses ***H3*** and ***H4***, we will compare low and high-ability subjects between these four

[2]For an image of the input screen refer to Appendix 5.B.

main groups. In addition, we have a baseline treatment that is used to group subjects according to ability. In a post-experimental questionnaire we measure individual loss aversion and competitiveness by survey questions in order to test the theoretical optimal performance (see Table 5.3.2 for the general timing of a session).

Reference point		**Grouping procedure**
average	$\times$	**ability grouped**
best		**randomly grouped**
(between-subject)		(within-subject)

Table 5.3.1: Treatments

(a) Baseline Treatment All subjects participated in the baseline treatment, taking place in the first period. Every participant was asked to solve as many multiplication problems as possible in 4 minutes incentivized by the described piece rate. They did not receive any information on other subjects' performance and were neither sorted into groups. They only received information on their own total number of solved problems after the period.

(b) Best vs. Average Treatment The *best* vs. *average* treatments are modeled in a between-subject design, i.e. subjects are either shown the *best* reference point or the *average* reference point after and before each of the four periods that follow the baseline period. By employing a between-subject design we avoid a demand effect that could arise, if subjects are offered two different reference points subsequently. During the experiment subjects are sorted into groups of five. These groups serve the only purpose of providing the reference point. In the *best* treatment we provide subjects with information on the *best* performance of their group. If the subject herself had the best performance we gave information on the second best performance. The subjects from the *average* treatment were given information about the *average* performance of their group, excluding the subject's own performance. This design of the reference points is done in line with the theory, to ensure that subjects receive information about their peers (only), i.e. we make sure that the reference point is exogenous to the subject's own performance. Throughout these four periods subjects are incentivized

by the piece rate, i.e. there is no payment depending on relative performance.

(c) Ability Grouped vs. Randomly Grouped Treatment The grouping treatments are modeled in a within-subject design. All subjects went through two periods of the *randomly grouped* treatment and through two periods of the *ability grouped* treatment. In the *randomly grouped* treatment subjects were randomly grouped with other subjects. This resulted in groups of subjects with more heterogeneous abilities. For the *ability grouped* treatment subjects were ranked according to their performance in the first period (*baseline*). All subjects that performed in the top 50% were sorted into a high track (high-ability type), and those that performed in the bottom 50% were sorted into a low track (low-ability type). Groups under the *ability grouped* treatment were then only randomly composed of subjects within these tracks. This resulted in groups of rather homogenous abilities.

baseline	→ tracking →	grouping procedure I →	grouping procedure II →	questionnaire
(1 period)		(2 periods)	(2 periods)	

4 minutes per period, 30 cents per solved task, one round payoff-relevant

Table 5.3.2: General Timing of a Session

The within-subject design is motivated by a closer fit to the theoretical considerations and the larger statistical power (see Charness, Gneezy, and Kuhn, 2012, for a discussion on between-subject and within-subject design). Since we expect not only the ability grouping to have an motivational effect, but also the tracking decision itself, we required the tracking to be an element of every session specification.[3] We implemented a cross-over design to account for order as well as learning effects.[4] While a between design would be a more conservative approach, the set of variations to be tested may be too large. Based on our theory, we need to compare subsets of the participants with respect to gender, and potentially social comparison and loss aversion. Therefore, we chose to have an within design to double the observations per grouping treatment under a given reference point.

[3]Table 5.4.4 reports evidence for the impact on performance of being tracked.

[4]The within design has one disadvantage that cannot be resolved completely by the cross-over design. Since the reference point is deduced from the past period, in period 4 of the *avrg* the difference between the grouping treatments would be underestimated. We do not find evidence for this distortion in the data. As a potential robustness check, the cross-over design also allows to observe a 2×2 between design if only the periods 2 and 3 are considered.

Procedural details The experiment was programmed with zTree (Fischbacher, 2007) and four sessions with a total of 120 participants were conducted at the experimental laboratory of the University of Hamburg in June and July 2015. We used hroot for recruitment (Bock, Baetge, and Nicklisch, 2014). The subjects were students of the University of Hamburg of which 58 were female. On average, a participant received a payout of 14 euro, including the show up fee of 5 euro. The sessions took about 60 minutes each.

5.4 Results

5.4.1 Summary Statistics and Prima Facie Evidence

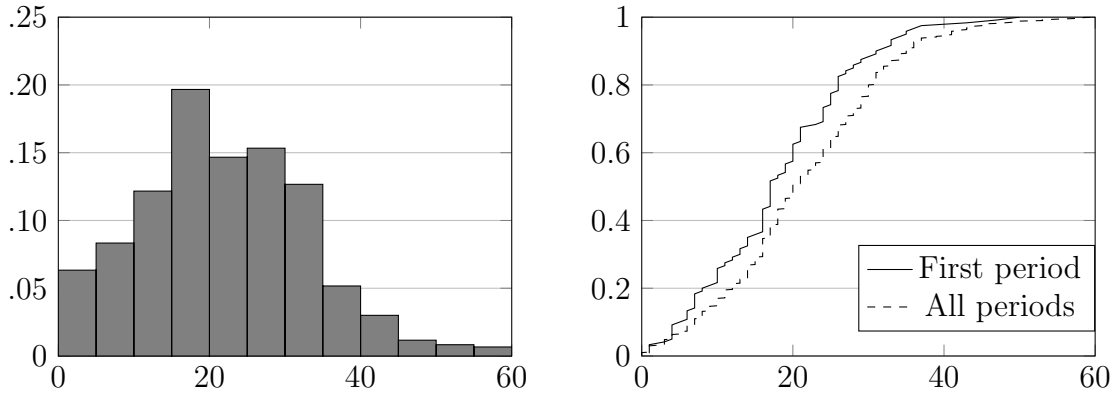
In a first step, we highlight the data on the aggregate level to analyze the heterogeneity of the subjects' ability and the elasticity of the effort task. To test the theoretical predictions on aggregate outcome from Section 5.2, we compare performance under different grouping regimes and reference point settings.[5] In a final step, the analysis focuses on the individual level to investigate the theory-derived optimal performance and the linearity of the peer effects.

Heterogeneity in ability The distribution of correctly solved tasks over the entire experiment shows a large heterogeneity in subjects' performance. We observe a range from zero up to a total of 60 correctly solved multiplications with a mean of 21.4. It can be taken from Figure 5.4.1 (a), that performance is positively skewed around the median of 20.[6] Since the heterogeneous performance may result from the treatments the subjects are exposed to, we consider only the baseline treatment to evaluate the innate ability (Figure 5.4.1 (b), solid line). In the first period, the performance stretches from zero to 50 correct answers with a mean of 18.3 such that the distribution shows a similar picture. Of course, subjects might already be affected by their treatment in the first period. Nevertheless, since the first round performance is significantly correlated with the math grade ($\rho = -0.31,\ p < 0.01$), we interpret this correlation as heterogeneous abilities take effect in this task.

[5] We use a Wilcoxon signed-ranks test for within-subject differences and a Mann-Whitney-U test for between-subject differences, respectively. Since each individual is observed twice in a treatment, we take the average of a subject over the two periods as an observational unit.

[6] Shapiro-Wilk test rejects normality of the data ($z < 0.001$).

With respect to our grouping criterion, Figure 5.4.2 (a) illustrates the mean performance on average and per ability type over the five periods. The results for period one show substantial differences in mean performance of those subjects who perform above the median (26.1) compared to those whose output is below the threshold (10.4). Therefore, we argue that grouping subjects according to ability based on this first period performance is reasonable.



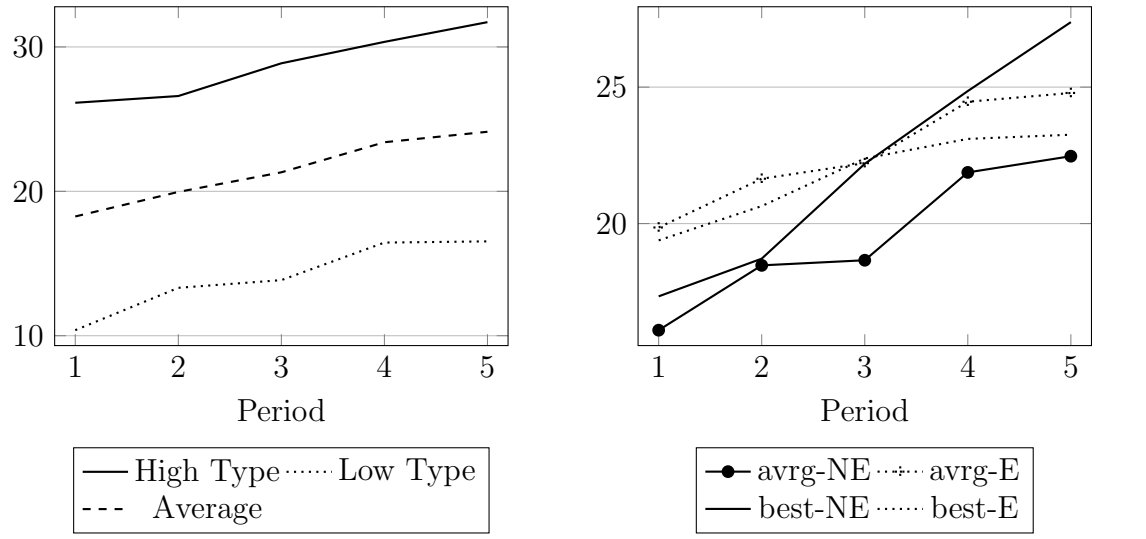
(a) Histogram, *all observations*

(b) CDF, *first and all periods*

Figure 5.4.1: Distribution of Correct Answers

Elasticity of the performance We observe a steady increase in average performance over the periods in which the subjects are exposed to performance feedback (from 18.3 to 24.1, Figure 5.4.2 (a), dashed line). If we evaluate this improvement separately for high-ability and low-ability subjects, we find similar increases for both types. The difference between the two types remains in the range between 13.3 and 15.8 in every period (Figure 5.4.2 (a), solid and dotted lines). This rather constant incline per period suggests that we do not face problems of convergence to an upper limit for neither of the ability types. Of course, this steady increase may not be exclusively explained by an elastic response to non-monetary incentives. Another reasonable explanation would be continuous learning that parallels the increasing relative performance feedback. If the subjects always exert the maximum effort - regardless of the performance feedback - the increase would be entirely driven by improvements in task-specific capabilities and the incentives from relative performance feedback could not factor in. To disentangle the effects from incentives and learning, we take a similar approach as Dohmen and Falk (2011) and Brüggen and Strobel (2007) to evaluate the effort level. In a post-experimental questionnaire we elicit the exhaustion level of the

participants (see Appendix 5.C, question 10). In Figure 5.4.2 (b) we plot the average performance given the reference point and the reported level of exhaustion. While those who experience exhaustion may not be able to react to relative feedback, we should see differences for those who do not report any level of exhaustion. As a first evidence for elasticity of effort in multiplication tasks, we find a larger share of the subjects report that they experienced exhaustion in treatments with maximum performance as the reference point (*best*) for which theory predicts a higher performance ceteris paribus (65% vs. 47 % in the *average* treatments (*avrg*)). Second, if we compare those subjects that do not experience exhaustion - i.e. those who have enough leeway to scale up the effort - we find those in the *best* treatments with a significantly larger average performance, indicating that these can respond to incentives in the form of relative feedback (best-NE: 22.1, avrg-NE: 19.5, $p < 0.08$).



(a) Ability Type

(b) Exhaustion Level and Reference Point

Figure 5.4.2: Average Performance over Time

Test of theory-derived hypotheses on aggregate outcome By contrasting the mean performance of the two grouping scenarios under a given reference point, we test hypotheses ***H1*** and ***H2***. Figure 5.4.3 displays the mean outcome and standard deviation for both random (RG) and ability grouping (AG) given average group performance as reference point (avrg) on the left-hand side, and the best group performance as reference point (best) on the right-hand side. Evaluating performance of all subjects (dark gray bars) under the *best* setting suggests that our experiment cannot confirm

hypothesis ***H1*** (RG $\approx$ AG = 22.7). Also with respect to hypothesis ***H2***, we do not find a significant difference in performance under the two grouping procedures when the *average* reference point is given (21.8 vs. 21.7). One reason for these results might be that the ability composition of the groups under random grouping is not as balanced as assumed in the theory. Figure 5.D.1 in Appendix 5.D shows that the mean ability (measured by first period performance) is sometimes lower than that of groups in the low track and sometimes higher than groups in the high track.

To investigate hypotheses ***H3*** and ***H4***, we compare the mean performances separately for high-ability subjects (light gray bars in Figure 5.4.3) and low-ability subjects (white bars). Hypothesis ***H3*** predicts a generally lower mean for low-ability subjects in an ability grouped setting compared to random grouping. This can neither be supported for the *best* setting (RG $\approx$ AG = 15.2), nor the *average* setting (RG: 15.2 vs. AG: 14.7) on the aggregate level. From hypothesis ***H4*** we expect an output-enhancing effect from ability grouping for high-ability subjects given average group performance as reference point. However, Figure 5.4.3 depicts the mean performance of high-ability subjects in the *average* setting as not significantly different across the grouping treatments (RG: 28.8 vs. AG: 28.4).

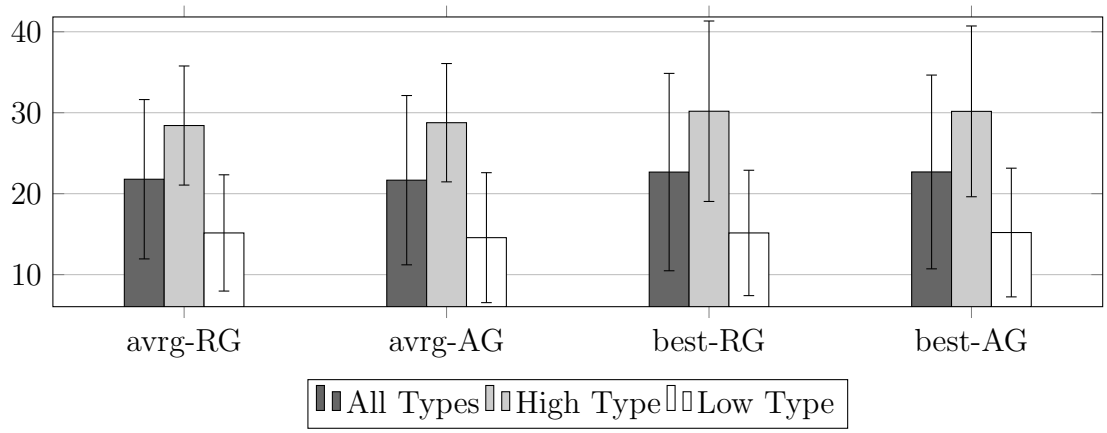


Figure 5.4.3: Mean Performance and Standard Deviation per Reference Point and Grouping Treatment

Apart from the theoretical predictions on treatment differences between the grouping scenarios, we neither observe any difference in mean performance when comparing the *best* and *average* feedback setting (best: 22.7, avrg: 21.7, $p < 0.89$). Still, we find higher peak performance in the *best* setting, with the 95th percentile at a level of 44

compared to 37 under the *average* setting. A variance ratio test confirms that the variance of performance under the *best* treatment is significantly higher ($p < 0.01$). This finding suggests that the best reference point has larger motivational effects at the top of the distribution. In particular, there might be non-linear peer-effects at work that depend on the distance of the subject's performance to the reference point. We will dissect this possibility further with regression analysis in Section 5.4.5.

5.4.2 Gender Differences

The lack of support for the theoretical predictions on the aggregate level might also be driven by systematic differences in performance by gender. As mentioned in Section 5.2 there is robust evidence in the literature that male and female behavior differs substantially in competitive environments. If their responses to the given incentives go into opposite directions, the effects might well cancel out in the aggregate. Since the literature has found that women often exhibit an aversion to competition, which is not incorporated in our theory, we might also expect that our hypotheses are only true for male subjects. We will thus turn the analysis to exploratory evidence on gender differences in behavior.

Overall, mean performance of male subjects is significantly higher than mean performance of females (male: 22.8, female: 19.9, $p < 0.01$). Moreover, male performance has a significantly higher variance ($p < 0.01$). Still, looking at the gender differences in the different treatments, we will see that men are not generally better at performing in the multiplication task.

In Figure 5.4.4 we plot mean performance of male and female subjects (gray and black lines) under the *average* reference point regime (dashed lines) and the *best* reference point regime (solid lines). Comparing the regimes shows that male subjects perform mildly significantly better in the *best* treatment than in the *average* treatment (best: 26.5, avrg: 21.3, $p < 0.09$)[7], while female subjects perform higher in the *average* treatment, without significance (best: 19.7, avrg: 22.4, $p < 0.23$). Also, under the best reference point male subjects perform significantly better than female subjects (male: 26.5, female: 19.7, $p < 0.03$). Under the *average* reference point there is no

[7]Since each individual is observed four times in a treatment, we take the average of a subject over the four periods as an observational unit and use the "bootstrap" technique of the two-sample t-test to calculate p-values (Efron and Tibshirani, 1993).

significant difference by gender (male: 21.3, female: 22.4, $p < 0.64$).

Regression analysis shows that the gender differences in the *best* setting cannot be explained by differences in ability (as measured by the reported math grade and two more controls)[8]. As suggested in Section 5.2 this result might be driven by females exhibiting an aversion to standing out and to competitive environments (e.g. Niederle and Vesterlund, 2007; Jones and Linardi, 2014). Figure 5.4.4 shows that the differences were already evident in the first period where no reference point was shown. Still, subjects had already received the instructions before the first period, thus knowing whether they were in the *best* or *average* treatment. Also, they knew that the first period performance was used to allocate subjects to the high and low track. This might have already created a competitive environment.

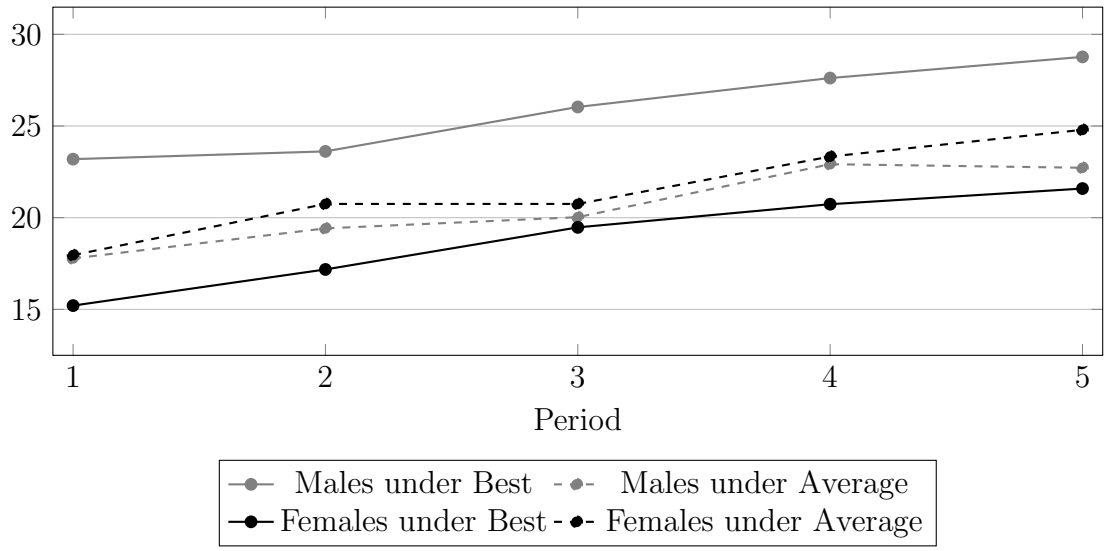


Figure 5.4.4: Mean Performance by Reference Point Treatment and Gender over Time

We further compare mean performance by gender and reference point under the two grouping regimes. Table 5.4.1 suggests that both grouping procedures have an effect on performance, but differently by gender. Overall we find a mildly significant difference in mean performance between the two grouping procedures for women. The opposite is true for male subjects, who perform significantly better under *random grouping*. Splitting this up by reference point regime, we find that there are no significant differences

[8]Female subjects have on average a significantly better (i.e., lower) math grade (females: 2.4 vs. males: 2.7, $p < 0.01$) with the math grade ranging from 1-6 and 1 being the best grade. See Appendix 5.E for the regression results.

Gender	**Ref. point**	**Random gr.**	**Ability gr.**	***p*-value**	***N***	**Share**
female	*both*	20.2	21.5	0.08	58	0.54
	best	19.2	20.3	0.26	34	0.53
	avrg	21.6	23.2	0.16	24	0.52
male	*both*	24.1	22.8	0.04	62	0.63
	best	27.2	25.9	0.25	26	0.58
	avrg	21.9	20.6	0.05	36	0.67

Notes: Wilcoxon-signed-rank test for within-subject comparisons. *Share* is the share of N subjects whose performance changes parallel to the average performance of the grouping treatments.

Table 5.4.1: Mean Performance per Grouping Treatment by Gender and Reference Point

between *random grouping* and *ability grouping* under the *best* reference point. When the *average* reference point is given, we find that female subjects perform indeed higher under *ability grouping* and male subjects under *random grouping*, with only the latter being significant.

If we also split this up for high and low types, we do not find any differences in the reaction to the grouping treatments, i.e. male high and low types both perform better under *random grouping* and female high and low types perform better under *ability grouping*.[9]

Even though we can summarize that ***H1*** is true for male subjects and ***H2*** is true for female subjects, these results are obviously not driven by the mechanisms from the theoretical model. In theory, ***H1*** is true, because only low type subjects gain from *random grouping* and ***H2*** is true because high types gain more than low types lose from *ability grouping*. Both is not the case here.

Overall, this section pointed out significant results when we acknowledge that male and female subjects behave differently. The first main result so far is that performance of men is higher when the *best* reference point is given compared to an *average* reference point, whereas no significant difference is found for women. The second main result is that women perform better under *ability grouping* and men under *random grouping*, especially when the *average* reference point is given. We will use regression analysis to disentangle the channels for the second result.

[9]With on average only 13 observations per group we do not report any statistical tests due to the lack of statistical power.

5.4.3 Testing Optimal Performance

The hypotheses tested in Section 5.4.1 were derived from the theoretical optimal performance as given in Section 5.2. Whether individual subjects behave according to the derived best response function can be tested directly in a system of regressions. If subjects behave optimally, their performance should depend positively on ability and the degree of social comparison. If the subject's performance is below the reference point, performance should also increase with the degree of loss aversion.

The dependent variable is performance of subject i in period t (measured in correctly solved problems). Regression (1) and (3) in Table 5.4.2 only include subjects that performed below the reference point, i.e. the average (or best) performance of their current group members in the last period. Regressions (2) and (4) include those that performed above the reference point.[10] The three covariates of interest are derived from questions that subjects answered in a non-incentivized questionnaire subsequent to the experiment[11]. Estimated coefficients of loss aversion (elicited by a method developed by Abdellaoui, Bleichrodt, and Haridon, 2008) had a mean of 3 and a standard deviation of about 3.5. As a control for ability we use subjects last math grade at school (ranging from 1-6, with 1 being the best grade). The degree of social comparison is proxied by a question on whether subjects would want to perform a task on their own for a piece rate or in a tournament competing for a prize (resulting in a binary variable that is 1 if the tournament was chosen). The regression also includes period and session dummies to control for period and session specific effects, especially for learning effects. Results of Ordinary Least Squares (OLS) regressions with standard errors clustered at the individual level to control for serial correlation in the error term are reported in Table 5.4.2, separately for male and female subjects[12].

[10] There are altogether only three observations, where the performance is equal to the reference point. These observations are not included in the regressions.

[11] See Appendix 5.C for the Questionnaire.

[12] Results for regressions testing optimal performance including interactions with the grouping and reference point regime are reported in Table 5.E.2 in Appendix 5.E. They show that some of the results vary under the two regimes. Under the *best* setting with *random grouping* the indicator for social comparison has a significantly positive effect for males, for women it has a significantly negative effect here. We only find significant effects of ability under *random grouping* and no significant effects for loss aversion below the reference point for any regime.

Variables	Females		Males	
	Below (1)	Above (2)	Below (3)	Above (4)
Loss Aversion	0.368	0.353	0.704**	0.124
	(0.501)	(0.398)	(0.279)	(0.159)
Social Comparison	-3.920	3.485	2.571	5.200
	(3.408)	(2.632)	(2.804)	(3.348)
Math Grade	0.324	-3.596**	-2.134**	-3.987***
	(1.360)	(1.640)	(0.960)	(1.331)
Period FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
R^2	0.10	0.36	0.35	0.48
Adj. R^2	0.00	0.22	0.30	0.41
N	93	51	122	78

Notes: Ordinary least squares regressions. Dependent variable: Number of correct answers. Regressions include periods 2-5. Robust standard errors in parenthesis are clustered at the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.4.2: Testing Theory-Derived Optimal Performance

We see that male subjects roughly behave as predicted by theory. The coefficient of loss aversion has a positive and significant impact on performance only for men whose past performance was below the reference point. Precisely, for male subjects below the reference point an increase of the coefficient of loss aversion by 1 induces on average an increase of solved tasks by 0.7. The indicator for social comparison has a positive impact on performance, but is not precisely estimated. Also ability, measured by the math grade, has a positive impact on performance both above and below the reference point. Female performance, on the other hand, is not in line with the theory. Especially, below the reference point, the given variables cannot explain female performance.

5.4.4 Linear Peer Effects

In the preceding section we have shown that (male) performance increases in loss aversion if the subject's performance is below the reference point. Here we estimate the size of the average effect of the reference point on performance. Typically, these *peer effects* are empirically modeled by the linear-in-means-model, meaning that performance of a single subject is regressed on the average performance of the subjects' reference group (see e.g. Brock and Durlauf, 2001). We proceed in this way for the *average* treatment,

while for the *best* treatment we regress individual performance on the best performance of each group. The following regression with period fixed effects μ_t and covariates $\mathbf{X}_i$ is estimated separately for the *best* and *average* treatment.

$$p_{it} = \alpha + \beta \; refpoint_{it} + \mathbf{X}_i\gamma + \mu_t + \epsilon_i \tag{5.5}$$

The variable $refpoint$ is the average (best) performance of the current group members from the last period that was shown to the subjects before each period. If performance below the reference point increases linearly in loss aversion, the size of the peer effect should be larger in the *best* treatment than in the *average* treatment. Table 5.4.3, hence, reports the results separately for both treatments. The way in which subjects react to a reference point should strongly depend on subject specific characteristics, as suggested by theory e.g. on factors like loss aversion, importance of social comparison and ability. These factors again might vary, for instance, with the cultural background or the gender of the individual subject.

Thus, we estimate a model that only includes $refpoint$ as a first step. The estimated coefficient gives the total impact of the reference point on performance, including any effect that might work through different subject characteristics such as culture, gender or ability. In a second step we include control variables for subject background factors gathered in the questionnaire subsequent to the experiment to see how this changes the impact of the reference point (these are: female, math grade, loss aversion, years since Abitur[13], studies math[14], income[15]). To analyze which factors drive the sensitivity to the reference point, we include some interactions of $refpoint$ with subject characteristics in a third step. We also include an indicator for *ability grouping* and the interaction with $refpoint$ to see whether the impact of the reference point differs by grouping regime. We use an OLS approach with clustered standard errors at the individual level. We expect β to be positive in specifications (1), (2), (4) and (5).

[13] *Abitur* is the name of the diploma awarded to students at the end of secondary schooling in Germany.

[14] The variable *studies math* is a dummy that takes on the value 1 if the subject studies a course that includes mathematics as a major component, such as information systems, economics, business, physics or mathematics.

[15] The variable *income* is an ordered categorical variable taking on the following values of disposable income per months (in euro): 1 = less than 400, 2 = 400-600, 3 = 600-800, 4 = 800-1000, 5 = 1000-1200, 6 = more than 1200.

Variables	Average			Best		
	(1)	(2)	(3)	(4)	(5)	(6)
Reference Point	0.569***	0.470***	0.414	0.298***	0.210**	0.059
	(0.115)	(0.109)	(0.292)	(0.079)	(0.083)	(0.211)
Ability Grouping		-0.100	-13.642***		0.309	-5.209
		(0.717)	(3.959)		(0.726)	(3.695)
Ability Grouping × Reference Point			0.667***			0.172
			(0.190)			(0.120)
Female		-2.327	4.287		-7.154***	-9.881*
		(2.411)	(4.402)		(2.669)	(5.639)
Female × Reference Point			-0.321*			0.084
			(0.181)			(0.141)
Loss Aversion		0.240	0.211		0.234	0.463
		(0.146)	(0.674)		(0.410)	(0.989)
Loss Aversion × Reference Point			-0.003			-0.007
			(0.027)			(0.025)
Math Grade		-2.525***	-0.088		-2.685**	-2.858
		(0.821)	(1.616)		(1.166)	(1.838)
Math Grade × Reference Point			-0.118*			0.007
			(0.068)			(0.047)
Years since Abitur		-0.584	-0.482		0.153	0.154
		(0.415)	(0.397)		(0.244)	(0.248)
Studies Math		1.346	1.260		8.883***	8.623***
		(2.019)	(1.945)		(2.757)	(2.821)
Income		1.063	0.891		0.343	0.294
		(0.739)	(0.674)		(1.018)	(1.024)
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.31	0.37	0.10	0.35	0.36
Adj. R^2	0.14	0.28	0.33	0.08	0.32	0.31
N	240	236	236	240	228	228

Notes: Dependent variable: Number of correct answers per period. Robust standard errors in parenthesis are clustered at the individual level. Regressions include period 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.4.3: Linear Peer Effects

From the results reported in Table 5.4.3, we see that in both reference point treatments individual performance increases in the reference point. Other than expected, the effect is almost twice as large in the *average* treatment. When the reference point is one correct answer higher, individual performance increases on average by more than half a correct answer in the *average* treatment and only by 0.3 correct answers in the

best treatment. In both treatments the impact of the reference point decreases once we control for subject characteristics, but it remains positive and significant. We also see that the indicator for whether the subject studies a subject that includes mathematics as a major component has a highly significant and huge impact on performance only in the *best* treatment. A driver might be that students, who study these more prestigious subjects, are more competitive and performance-oriented (see e.g. Buser, Niederle, and Oosterbeek, 2014). For the *best* treatment, including interactions does not shed any light on what drives the sensitivity to the reference points. In the *average* treatment, we find weak evidence for female subjects (see also Beugnot, Fortin, Lacroix, and Villeval, 2013) and less able subjects (measured by the math grade) being less motivated by the reference point. Also, performance of subjects under *ability grouping* increases more strongly in the reference point than under *random grouping*. A reason for this might be found in non-linear peer effects as discussed in the next section.

5.4.5 Non-Linear Peer Effects

Unlike suggested by theory we have seen in the last section that an *average* reference point, especially under *ability grouping*, has a higher impact on individual performance than the best reference point. A reason for this could be nonlinear effects and diminishing sensitivity with respect to the reference point as suggested by Tversky and Kahneman (1979). To find the effect of the distance to the reference point in our sample we use a differencing method, i.e. the dependent variable is the change in correctly answered problems compared to the period before. With this approach we can avoid multicollinearity of the subjects' performance and the distance to the reference point. We can also eliminate time-invariant factors like subject ability and concentrate on what causes the change in performance between periods. The following regression is estimated:

$$\begin{aligned}\Delta p_{it} =& \alpha + \beta_1 below_{it-1} + \beta_2 absdist_{it-1} + \beta_3 absdist_{it-1} \times below_{it-1} \\ &+ \beta_4 trackdec_{it} + \beta_4 trackdec_{it} \times lowtype_i + \mu_t + \mu_i + \Delta\epsilon_{it} \end{aligned} \tag{5.6}$$

The variable *absdist* is the absolute distance in points of the subjects last period performance to the reference point. The variable *below* indicates whether the subject had performed below the reference point in the last period. The only other thing that changes with t is that subjects are told before the *ability grouped* treatment

whether they were sorted into the low or high track. This is controlled for by a dummy (*trackdec*). We also include an interaction of *trackdec* with *lowtype*, which indicates whether subjects were sorted into the low track. At the cost of explanatory power, we estimate fixed effects models with subject and period fixed effects to eliminate biases due to unobserved subject characteristics and learning effects. We split the sample by gender, to find explanations for the differences in behavior discussed in Section 5.4.2.

To find proof of a peer effect that is larger below the reference point, we would expect $\beta_1 > 0$. In order to find support for diminishing sensitivity as suggested by Tversky and Kahneman (1979), we would expect $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$. Results are reported in Table 5.4.4.

Variables	Female			Male		
	(1)	(2)	(3)	(4)	(5)	(6)
Below the Reference Point	3.611***	0.793	1.791	4.111***	3.416***	1.842
	(0.797)	(1.107)	(1.755)	(0.904)	(1.121)	(1.757)
Absolute Distance to the Reference Point		-0.300**	-0.180		-0.206**	-0.292
		(0.137)	(0.144)		(0.098)	(0.175)
Absolute Distance to the Reference Point						
× Below the Reference Point		0.583***	0.821***		0.180	0.420
		(0.154)	(0.278)		(0.122)	(0.252)
Below the Reference Point						
× Best			-5.138**			2.539
			(2.211)			(2.700)
Absolute Distance to the Reference Point						
× Best			-0.420			0.165
			(0.328)			(0.215)
Absolute Distance to the Reference Point						
× Below the Reference Point						
× Best			0.090			-0.341
			(0.421)			(0.278)
Period of Tracking Decision	-0.037	-1.839	-1.489	0.138	-1.913	-2.159*
	(0.995)	(1.473)	(1.465)	(0.936)	(1.198)	(1.165)
Period of Tracking Decision						
× Low Type		4.609**	5.155***		3.858**	4.230**
		(1.921)	(1.919)		(1.615)	(1.613)
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.06	0.18	0.22	0.10	0.14	0.15
adj. R^2	0.04	0.15	0.18	0.08	0.11	0.11
N	232	232	232	248	248	248

Notes: Dependent variable: Change in performance compared to last period. Robust standard errors in parenthesis are clustered at the individual level. Regressions include periods 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.4.4: Effect of Distance to Reference Point

Specification (1) and (4) show that subjects who were told that they performed below the reference point improve their output in the following period by around four solved problems more than those above the reference point. Including the variable on the distance, specification (2) and (5) show that both female and male subjects exhibit diminishing sensitivity above the reference point, but women's decreasing more than men's. A striking difference between the sexes is found below the reference point, where females show increasing sensitivity while males have a constant sensitivity. This

means females are more motivated the further below the reference point they are. Specification (3) points out that this is only true for the *average* treatment. In the *best* treatment women's output decreases subsequent to being told to have performed below the reference point, no matter the distance. The reference point has thus even a demotivating effect.

Evaluating the output subsequent to the tracking information, we find patterns that also have been found in the previous literature (Kuhnen and Tymula, 2012; Gill, Kissová, Lee, and Prowse, 2018). The mere information that one's performance was in the bottom half, leads to a significant improvement in performance in the following period. In Section 5.4.1 we observed that performance under the *best* setting has a higher variance than performance under *average*. This can be explained by the diminishing sensitivity above the reference point under *average*, while sensitivity to the reference point is constant under *best*. These different patterns in the reaction to the reference point might explain why women on average perform better under *ability grouping*: First, the faster diminishing sensitivity above the reference point makes a reference point that is closer more effective and, second, being told to be in the low track has a larger motivational effect for women than for men.

5.5 Conclusion

We tested theoretical predictions and evaluated the role of gender differences for subjects' performance conditional on their peer group's composition and relative performance feedback in a laboratory experiment. While theory-derived hypotheses on aggregate treatment differences cannot be confirmed, we find evidence when gender differences are taken into account. Support is found only for male subjects behaving according to theory-derived optimal performance, i.e. their performance is driven by ability and by loss aversion if their performance is below the received reference point. Male subjects perform significantly better than women in response to the *best* reference point, and better than under the *average* reference point. Regression analysis shows that women even reduce output in response to being told to have performed worse than the best in their group, underlying that women behave contrary to the theoretical assumptions. This result might be driven by females exhibiting an aversion to standing out and to competitive environments (e.g. Niederle and Vesterlund, 2007; Jones and Linardi, 2014). With respect to the grouping treatments we find that female mean

performance is significantly lower under *random grouping* than under *ability grouping*, while men perform significantly better under *random grouping*. From regression analysis we found that this might be due to gender differences in the (non-linear) reaction to the reference point and to low-ability females reacting stronger to being sorted into the low track. A possible reason we cannot test for might be that women respond rather to a stronger group identity due to similar ability-types under *ability grouping* (see e.g. Croson and Gneezy, 2009). The rather small differences between the grouping treatments might also be due to insufficiently balanced groups under the *random grouping* treatment. In addition, the elasticity of effort in the multiplication task appear to be rather moderate which may hamper the identification of the treatment effects.

Our findings have implications for the design of feedback technologies and grouping procedures. Based on our results, a decision maker may acknowledge the individuals' background. Copying successful designs may not be a promising strategy when the characteristics of the target group are substantially different. The main factors that we identify are loss aversion of the individuals and in particular the gender. In practice, the gender differences that relate to social comparison may be even more pronounced when social recognition is leveraged instead of private feedback (Gerhards and Siemer, 2016).

5.A Translated Instructions

Welcome to today's experiment!
Today you are taking part in an economic experiment. Please note, that from now on and during the whole experiment no communication is allowed. If you have any questions during the experiment, please raise your hand and one of the experimenters will come to your cabin. In this experiment you can earn money by solving multiplication tasks. To solve the tasks you are not allowed to use any helping device, in particular no paper, pencil, calculator or mobile telephone. If you use any such helping device, you will be immediately excluded from the experiment and will get no remuneration. This experiment consists of five multiplication periods of four minutes each (240 seconds). We ask you to solve as many multiplication tasks as possible in one period. The tasks always consist of the multiplication of a one-digit number and a two-digit number. A task will be displayed as long as you need to answer the task correctly. Your remaining time will be displayed at the top of the screen. At the end of the experiment one of the

five periods will be randomly chosen for the remuneration. The number of correctly answered problems in that period will be converted into euro according to the following exchange rate:

1 solved problem = 30 euro cent

In addition everyone receives 5 euro for attendance. At the beginning of the experiment you will have the possibility to test the input-screen in a 30 seconds trial period. After going through the five multiplication periods, we ask you to fill in a short questionnaire. The experiment is divided into three parts. Part 1 consists of one of the above described multiplication periods.

[The order of the following two paragraphs was changed depending on the treatment]

Part 2 [3] consists of periods 2 and 3 [4 and 5]. Here, you will be randomly allocated to a group of five. Your identity will at no point be published to your group members. Before each period you will receive information about the average [best] performance (in correctly answered problems) of your group members in the last period.

Part 3 [2] consists of periods 4 and 5 [2 and 3]. Before period 4 [2] you will be sorted either into track 1 or track 2 based on your performance in part 1. All the participants that performed higher than the median performance in the first period are allocated to track 1. Every subject that performed below median performance is allocated to track 2. Within these tracks again groups of five will be formed randomly before each period. At the beginning of part 3 [2] you will be told into which track you have been sorted. In addition you will again be informed before each period about the average [best] performance of your group members.

If you have questions about these instructions, please raise your hand out of your cabin. One of the experimenters will come to you.

Good luck!

5.B Translated Screens

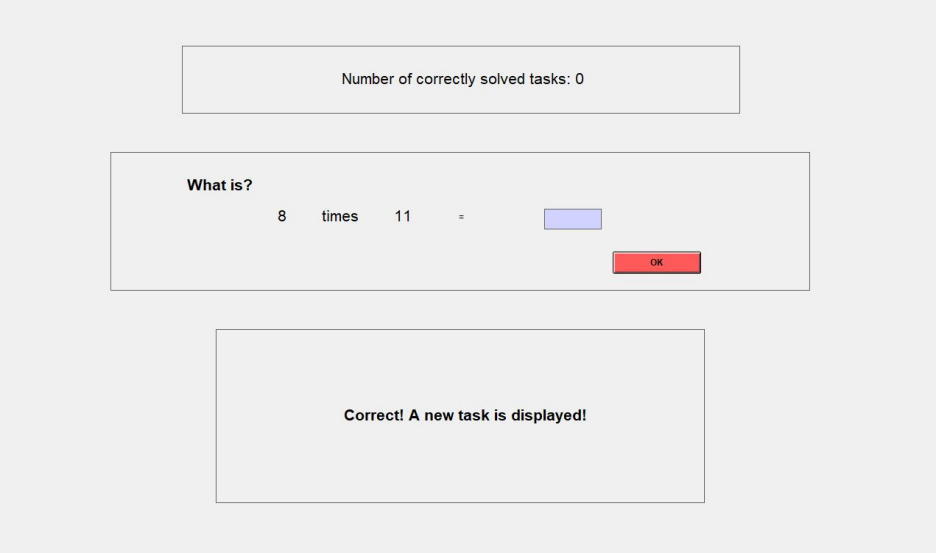


Figure 5.B.1: Input Screen

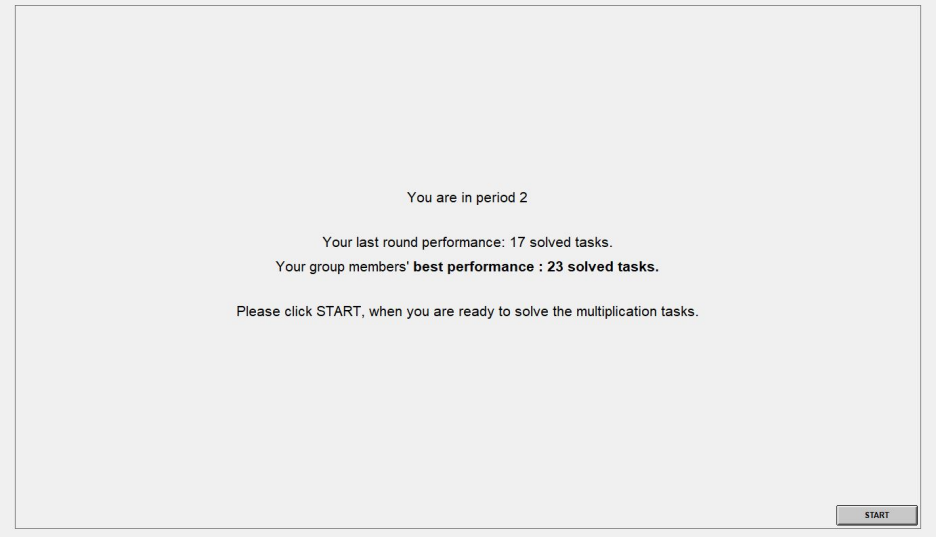


Figure 5.B.2: Feedback Screen

5.C Questionnaire

1. How old are you? _____

2. What is your sex? ☐ Male ☐ Female
3. What are you studying? _____
4. What was your last math grade at your last school? ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6
5. When did you graduate from secondary school? _____
6. How much money do you have at your disposal per month? (including rent) ☐up to 400 euro ☐ 400-600 euro ☐ 600-800 euro ☐ 800-1000 euro ☐ 1000-1200 euro ☐ more than 1200 euro
7. Is German your native language? ☐ Yes ☐ No
8. If no, please indicate your native language? _____
9. Do you have the feeling that you could answer the multiplication problems faster over time due to practice? ☐ Yes, very much ☐ Yes, a little ☐ No
10. Did you get exhausted as time in the experiment went by, so that you could concentrate less? ☐ Yes, very much ☐ Yes, a little ☐ No
11. Imagine you are playing a quiz with 10 questions. Which possibility of earning money would you prefer? A: You get 4 euro for each correct answer. B: You get 60 euro , if you give more correct answers than another unknown person. How do you decide? ☐ A ☐ B

5.C.1 Loss Aversion

Loss aversion of subjects was assessed by a method developed by Abdellaoui, Bleichrodt, and Haridon (2008). Subjects were asked the following three questions subsequent to the experiment:

1. Imagine a fair coin is flipped. You are offered a lottery, in which you can win 100 euro if Head appears and nothing if Tails appears. Instead of playing the lottery you can accept a certain gain. Which of the following gains would you accept?

	reject	accept
10 euro	☐	☐
20 euro	☐	☐
30 euro	☐	☐
40 euro	☐	☐
50 euro	☐	☐
60 euro	☐	☐
70 euro	☐	☐
80 euro	☐	☐
90 euro	☐	☐
100 euro	☐	☐

2. The coin is flipped again. You are offered a game in which you lose 150 Euro if Head appears and lose 50 Euro if Tails appears. Alternatively you can accept a certain loss. Which of the following certain losses would you accept?

	reject	accept
-140 euro	☐	☐
-130 euro	☐	☐
-120 euro	☐	☐
-110 euro	☐	☐
-100 euro	☐	☐
-90 euro	☐	☐
-80 euro	☐	☐
-70 euro	☐	☐
-60 euro	☐	☐
-50 euro	☐	☐

3. The coin is flipped again. You can either reject the game and earn/lose nothing, or you can accept the proposed game. Which of the following games would you accept?

	reject	accept
If Head appears, you earn 30 euro. If Tails appears you lose 50 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 45 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 40 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 35 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 30 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 25 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 20 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 15 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 10 euro.	☐	☐
If Head appears, you earn 30 euro. If Tails appears you lose 5 euro.	☐	☐

The first question is used to elicit the participants' utility in the domain of gains. By presenting a gain prospect x_i its certainty equivalent G_i is elicited. From $u(G_i) = \delta^+ u(x_i)$ the δ^+ can be determined. The second question is used to elicit the certainty equivalent for losses L_i for a prospect of losses (x_i, y_i). With $u(L_i) = \delta^-(u(x_i) - u(y_i)) + u(y_i)$ the δ^- is determined. The third question serves the elicitation of an indifference loss L^* for a given gain G^*. Then the coefficient of loss aversion λ was determined from the following equation: $\delta^+ u(G^*) + \lambda \delta^- u(L^*) = u(0) = 0$. Throughout the elicitation linear utility functions were assumed. For a more detailed description of the procedure see Abdellaoui, Bleichrodt, and Haridon (2008).

5.D Descriptive Statistics

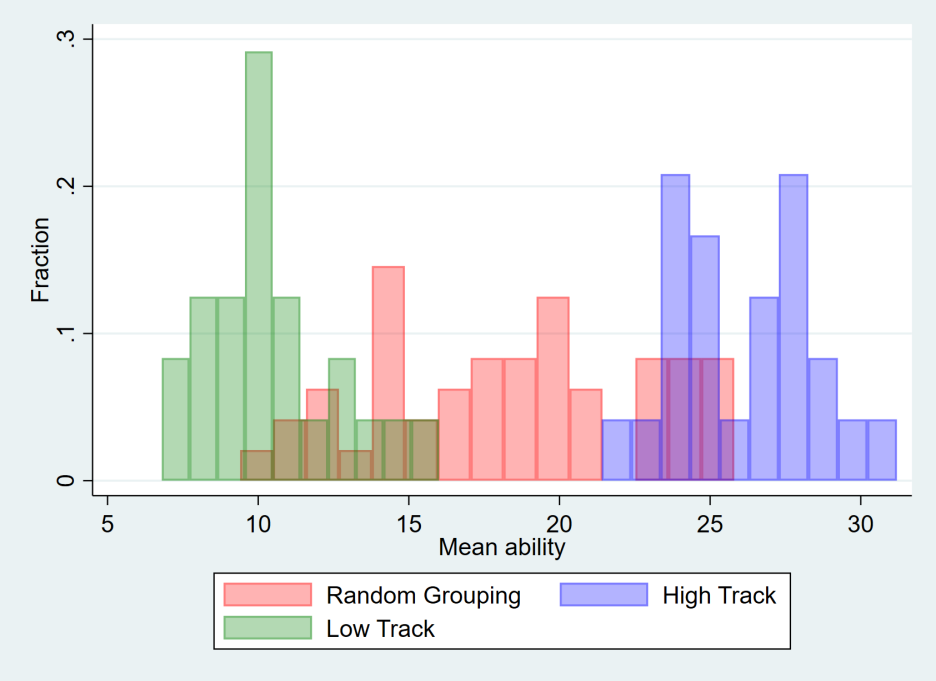


Figure 5.D.1: Histograms of the Mean Ability of the Groups under the Grouping Regimes

	Male			Female		
Variable	Mean	Std. Dev.	N	Mean	Std. Dev.	N
Number of Correct Answers	22.784	12.574	310	19.945	8.972	290
Reference Point	25.924	11.05	248	26.158	10.799	232
Loss Aversion	2.995	3.784	50	3.032	3.023	36
Competitiveness	0.306	0.462	62	0.241	0.429	58
Math Grade	2.726	1.299	62	2.397	1.247	58
Years since Abitur	7.21	5.915	62	5.228	2.573	57
Studies Math	0.583	0.494	60	0.439	0.497	57
Age	26.355	6.529	62	24.086	3.069	58
Income	2.565	1.49	62	2.707	1.315	58
German Native Speaker	0.774	0.419	62	0.724	0.448	58

Notes: The number of observations in the case of "Number of Correct Answers" and "Reference Point" is the number of periods times the number of subjects ($N = t \times n$). Otherwise $N = n$.

Table 5.D.1: Summary Statistics

Variable	NumberAns	Refpoint	LossAv.	Social.	Female	Grade	Abitur
Number Ans.	1.000						
Refpoint	0.296***	1.000					
Loss Aversion	0.095**	0.038	1.000				
Social Comparison	0.079*	0.067	-0.172***	1.000			
Female	-0.128***	0.011	0.005	-0.073*	1.000		
Math Grade	-0.297***	-0.090**	0.116**	0.004	-0.128***	1.000	
Years since Abitur	-0.049	0.039	-0.072	-0.046	-0.210***	0.178***	1.000
Studies Math	0.187***	-0.069	0.081*	0.099**	-0.145***	0.009	-0.048
Age	-0.082**	0.005	-0.068	-0.076*	-0.215***	0.222***	0.932***
Income	0.055	-0.027	-0.060	0.147***	0.051	-0.019	0.174***
German Native	-0.050	-0.104**	-0.130***	-0.162***	-0.058	0.135***	0.131***

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.D.2: Pairwise Correlations

Variable	StudMath	Age	Income	GermanNat.
Studies Math	1.000			
Age	-0.059	1.000		
Income	-0.024	0.151***	1.000	
German Native	-0.181***	0.116***	0.191***	1.000

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.D.3: Pairwise Correlations continued

5.E Regressions

Variables	(1)	(2)
Best	5.239	5.951**
	(3.172)	(2.843)
Female	1.135	-0.443
	(2.334)	(2.162)
Best × Female	-7.902**	-6.196*
	(3.812)	(3.472)
Math Grade		-2.595***
		(0.815)
Years since Abitur		-0.089
		(0.257)
Studies Math		4.309**
		(1.817)
Constant	19.029***	24.232***
	(1.824)	(3.018)
Period FE	Yes	Yes
R^2	0.07	0.20
adj. R^2	0.06	0.18
N	480	464

Notes: Dependent variable: Number of correct answers per period. Robust standard errors in parenthesis are clustered at the individual level. Regressions include period 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.E.1: Gender and Reference Point Regime

Variables	Female		Male	
	Below (1)	Above (2)	Below (3)	Above (4)
Best	-15.898**	14.145**	10.281	25.336***
	(5.871)	(6.338)	(8.406)	(7.360)
Ability Grouping	7.666**	-4.029	3.834	-5.749
	(3.716)	(6.494)	(4.791)	(6.175)
Loss Aversion	0.485	2.206**	0.055	-0.110
	(0.425)	(0.986)	(1.932)	(0.195)
Loss Aversion × Best	0.706	-1.482	0.583	-1.028*
	(0.747)	(1.195)	(1.893)	(0.594)
Loss Aversion × Ability Grouping	0.082	-2.760*	0.520	0.386*
	(0.482)	(1.426)	(1.920)	(0.216)
Loss Aversion × Ability Grouping × Best	-0.430	3.002*	-0.824	1.938
	(0.658)	(1.703)	(1.797)	(1.865)
Social Comparison	5.092	4.301	-4.208	1.719
	(3.520)	(3.238)	(4.347)	(4.076)
Social Comparison × Best	-4.930	-10.606**	14.153***	10.149*
	(4.925)	(5.013)	(4.872)	(5.609)
Social Comparison × Ability Grouping	-5.034	1.135	3.012	3.927
	(3.111)	(4.306)	(3.780)	(4.578)
Social Comparison × Ability Grouping × Best	o.	3.190	-3.172	0.402
		(8.408)	(5.730)	(5.803)
Math Grade	-2.989	-4.688**	-0.539	-0.400
	(1.895)	(1.793)	(1.595)	(2.120)
Math Grade × Best	5.331**	-1.686	-3.176	-8.608***
	(2.119)	(3.977)	(2.314)	(2.630)
Math Grade × Ability Grouping	-2.393	4.588	-1.837	-0.278
	(1.448)	(3.375)	(1.382)	(2.306)
Math Grade × Ability Grouping × Best	0.874	-4.190	1.079	-1.080
	(1.063)	(3.711)	(1.684)	(2.459)
Period FE	Yes	Yes	Yes	Yes
R^2	0.30	0.49	0.36	0.63
Adj. R^2	0.15	0.23	0.26	0.53
N	93	51	122	78

Notes: Ordinary least squares regressions. Dependent variable: Number of correct answers. Regressions include periods 2-5. Robust standard errors in parenthesis are clustered at the individual level. 'o.' means that this variable has been omitted due to perfect collinearity. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.E.2: Testing Theory-Derived Optimal Performance

Variables	Average		Best	
	(1)	(2)	(3)	(4)
Below the Reference Point	4.381***	1.810	2.795***	0.549
	(0.753)	(1.194)	(0.908)	(1.303)
Absolute Distance to the Reference Point		-0.261**		-0.180**
		(0.120)		(0.088)
Absolute Distance to the Reference Point				
× Below the Reference Point		0.505***		0.359***
		(0.178)		(0.110)
Period of Tracking Decision	-0.089	-1.980*	-0.380	-2.950**
	(0.880)	(1.124)	(1.077)	(1.442)
Period of Tracking Decision				
× Low Type		3.138**		6.097***
		(1.444)		(1.991)
Constant	-0.046	1.457	-0.700	-0.920
	(0.796)	(1.166)	(1.145)	(1.294)
Period FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
R^2	0.17	0.22	0.03	0.11
Adj. R^2	0.16	0.19	0.01	0.08
N	240	240	240	240

Notes: Dependent variable: Change in performance compared to last period. Robust standard errors in parenthesis are clustered at the individual level. Regressions include periods 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.E.3: Effect of Distance to Reference Point by Feedback Treatment

Summaries

Chapter 2: *The Contribution of Managers to Organizational Success: Evidence from German Soccer*

We study the impact of managers on the success of professional soccer teams using data from the German Bundesliga, where we are exploiting the high turnover rate of managers between teams to disentangle the managers' contributions. Teams employing a manager from the top of the ability distribution gain on average considerably more points than those employing a manager from the bottom. Moreover, estimated abilities have significant predictive power for future performance. Managers also affect teams' playing style. Finally, teams whose manager has been a former professional player perform worse on average compared to managers without a professional player career.

Wir untersuchen den Einfluss von Trainern auf den Erfolg professioneller Fußballmannschaften mithilfe von Daten aus der Deutschen Fußball-Bundesliga. Dafür nutzen wir die hohe Fluktuation der Trainer zwischen verschiedenen Mannschaften, um den Beitrag der Trainer herauszuarbeiten. Mannschaften, die einen Trainer beschäftigen, dessen geschätzte Fähigkeit in der Spitze der Verteilung liegt, erzielen durchschnittlich erheblich mehr Punkte als die Mannschaften von Trainern, deren Fähigkeiten sich am Ende einordnen. Die geschätzte Fähigkeit der Trainer hat statistisch signifikante Prognosekraft bezüglich zukünftigem Erfolg. Zudem zeigen wir, dass Trainer auch den Spielstil einer Mannschaft beeinflussen. Schließlich betrachten wir den Zusammenhang von beobachtbaren Charakteristika und Erfolg. Dabei finden wir heraus, dass Mannschaften, deren Trainer zuvor professioneller Spieler war, schlechter abschneiden, als solche, deren Trainer kein Fußballprofi war.

Chapter 3: *The Hidden Costs of Whistleblower Protection*

We conduct a laboratory experiment to analyze cooperative behavior between a manager and an employee in the presence of misbehavior and protected whistleblowing.

Before taking part in a trust game with her employee, a manager has the opportunity to embezzle money at the expense of a third party. Her behavior is observed by the unaffected employee who may trigger an investigation by a report. We vary the framework with respect to incentives and anonymity in case of a report and compare misbehavior, reporting and cooperative behavior across treatments. Our results suggest that a whistleblower law could deter wrongdoing, but could also have a detrimental effect on cooperation in organizations when it increases the probability for false whistleblowing.

Mithilfe eines Laborexperiments analysieren wir das kooperative Verhalten zwischen einem Manager und einem Arbeitnehmer, wenn der Arbeitnehmer Fehlverhalten des Managers anzeigen kann und vor Repressionen des Managers geschützt wird. Bevor ein Manager ein Trust Game mit dem Arbeitnehmer spielt, kann sich dieser durch Veruntreuung auf Kosten einer dritten Partei Geld aneignen. Diese Entscheidung wird vom Arbeitnehmer beobachtet und kann von diesem angezeigt werden. Der rechtliche Rahmen wird bezüglich der Anreize für eine Anzeige und der Anonymität des Anzeigenden variiert. Dabei vergleichen wir die Bereitschaft des Arbeitnehmers Fehlverhalten anzuzeigen, die Häufigkeit mit der Manager Geld veruntreuen und die Neigung der Manager eine produktive Kooperation mit dem Arbeitnehmer eingehen. Die Ergebnisse zeigen, dass Anonymität und Anreize zur Anzeige die Bereitschaft zur Anzeige erhöhen und die Häufigkeit des Fehlverhaltens der Manager reduziert. Gleichzeitig sinkt allerdings die Kooperationsbereitschaft der Manager, wenn die Wahrscheinlichkeit für Falschanzeigen steigt.

Chapter 4: *Gender Differences in Honesty: Groups versus Individuals*

Extending the die rolling experiment of Fischbacher and Föllmi-Heusi (2013), we compare gender effects with respect to unethical behavior by individuals and by two-person groups. In contrast to individual decisions, gender matters strongly under group decisions. We find more lying in male groups and mixed groups than in female groups.

Wir erweitern das Würfel-Experiment von Fischbacher und Föllmi-Heusi (2013), um Geschlechterunterschiede bei unethischem Verhalten zwischen Individuen und Gruppen (bestehend aus zwei Personen) zu vergleichen. Im Gegensatz zu individuellen

Entscheidungen spielt das Geschlecht eine große Rolle bei Gruppenentscheidungen. Wir stellen fest, dass häufiger von Gruppen gelogen wird, wenn sie aus ausschließlich männlichen Mitglieder bestehen oder es sich um gemischte Gruppen handelt, als wenn die Gruppen aus ausschließlich weiblichen Mitgliedern bestehen.

Chapter 5: *Peer Effects Under Different Relative Performance Feedback and Grouping Procedures*

We conduct a laboratory experiment to test theoretical predictions about subjects' performance in an effort task conditional on their peer group's composition and relative performance feedback. Subjects are grouped either randomly or according to their ability, with the feedback being the best or average performance of their group. While theory-derived hypotheses on aggregate treatment differences cannot be confirmed, we find evidence when gender differences are taken into account. Male subjects perform significantly better when they compare themselves with the best peer instead of the average, while the opposite is true for females. With respect to the grouping treatment, we find that random grouping is beneficial for male subjects, and ability grouping for female subjects. These differences are explained by gender differences in (non-linear) reactions to the reference point and an aversion of females to competitive environments.

Wir führen ein Laborexperiment durch, um theoretische Hypothesen über die Leistung der Teilnehmer unter unterschiedlichen Gruppenkonstellationen und unterschiedlichem relativen Leistungsfeedback zu überprüfen. Die Teilnehmer werden entweder zufällig oder entsprechend ihren Fähigkeiten Gruppen zugeteilt, in welchen sie entweder Feedback über die durchschnittliche oder über die beste Leistung in der Gruppe erhalten. Während wir die Hypothesen bezüglich der aggregierten Treatment-Effekte nicht bestätigen können, finden wir Evidenz für Geschlechterunterschiede. Männliche Teilnehmer schneiden signifikant besser ab, wenn sie sich mit der besten, anstatt der durchschnittlichen Leistung vergleichen. Für weibliche Teilnehmer gilt das umgekehrte. Bezüglich der Gruppenkonstellation finden wir heraus, dass männliche Teilnehmer von zufälliger Gruppenzuteilung profitieren, wohingegen weibliche Teilnehmer besser abschneiden, wenn sie nach Fähigkeiten gruppiert werden. Diese Unterschiede werden durch Geschlechterunterschiede bei (nichtlinearen) Reaktionen auf das Feedback und eine Aversion weiblicher Teilnehmer gegenüber einer kompetitiven Umgebung erklärt.

Chapter B.

List of Publications

Muehlheusser, G., A. Roider, and N. Wallmeier (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters 128*, 25–29.

Muehlheusser, G., S. Schneemann, D. Sliwka, and N. Wallmeier (2018). The contribution of managers to organizational success: Evidence from German soccer. *Journal of Sports Economics 19*(6), 786–819.

List of Tables

List of Figures

Bibliography

Abbink, K. and H. Hennig-Schmidt (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics 9*(2), 103–121.

Abbink, K., B. Irlenbusch, and E. Renner (2000). The moonlighting game: An experimental study on reciprocity and retribution. *Journal of Economic Behavior & Organization 42*(2), 265–277.

Abbink, K., B. Irlenbusch, and E. Renner (2002). An experimental bribery game. *Journal of Law, Economics, and Organization 18*(2), 428–454.

Abdellaoui, M., H. Bleichrodt, and O. Haridon (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty 36*(3), 245–266.

Abeler, J., A. Becker, and A. Falk (2014). Representative evidence on lying costs. *Journal of Public Economics 113*, 96–104.

Abeler, J., C. Raymond, and D. Nosenzo (2018). Preferences for truth-telling. *Econometrica forthcoming*.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Alekseev, A., G. Charness, and U. Gneezy (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization 134*, 48–59.

Alford, C. (2001). *Whistleblowers: Broken Lives and Organizational Power*. Cornell University Press.

Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Anechiarico, F. and J. B. Jacobs (1996). *The pursuit of absolute integrity: How corruption control makes government ineffective*. University of Chicago Press.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ.

Apesteguia, J., M. Dufwenberg, and R. Selten (2007). Blowing the whistle. *Economic Theory 31*(1), 143–166.

Argys, L. M., D. I. Rees, and D. J. Brewer (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management 15*(4), 623–645.

Arrow, K. J. (1974). *The limits of organization*. WW Norton & Company.

Association of Certified Fraud Examiners (2014). Report to the nations on occupational fraud and abuse: 2014 global fraud study. Technical report, Association of Certified Fraud Examiners, Austin.

Audas, R., S. Dobson, and J. Goddard (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business 54*(6), 633–650.

Bamber, L. S., J. Jiang, and I. Y. Wang (2010). What's my style? The influence of top managers on voluntary corporate financial disclosure. *The Accounting Review 85*(4), 1131–1162.

Barr, A. and D. Serra (2009). The effects of externalities and framing on bribery in a petty corruption experiment. *Experimental Economics 12*(4), 488–503.

Bartuli, J., B. Djawadi, and R. Fahr (2016). Business ethics in organizations: An experimental examination of whistleblowing and personality. *IZA Discussion Paper 10190*.

Benabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies 70*(3), 489–520.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Bertrand, M. and A. Schoar (2003). Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics 118*(4), 1169–1208.

Betts, J. R. and J. L. Shkolnik (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review 19*(1), 1–15.

Beugnot, J., B. Fortin, G. Lacroix, and M. C. Villeval (2013). Social networks and peer effects at work. *IZA Discussion Paper 7521*.

Bigoni, M., S.-O. Fridolfsson, C. Le Coq, and G. Spagnolo (2012). Fines, leniency, and rewards in antitrust. *The RAND Journal of Economics 43*(2), 368–390.

Bigoni, M., S.-O. Fridolfsson, C. Le Coq, and G. Spagnolo (2015). Trust, leniency, and deterrence. *The Journal of Law, Economics, and Organization 31*(4), 663–689.

Bloom, N. and J. Van Reenen (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics 122*(4), 1351–1408.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg registration and organization online tool. *European Economic Review 71*, 117–120.

Bolton, G. E. and A. Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review 90*(1), 166–193.

Bornstein, G., T. Kugler, and A. Ziegelmeyer (2004). Individual and group decisions in the centipede game: Are groups more "rational" players? *Journal of Experimental Social Psychology 40*(5), 599–605.

Brandts, J. and G. Charness (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics 14*(3), 375–398.

Breuer, C. and R. Singer (1996). Trainerwechsel im Laufe der Spielsaison und ihr Einfluss auf den Mannschaftserfolg. *Leistungssport 26*, 41–46.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. *Handbook of Econometrics 5*, 3297–3380.

Brüggen, A. and M. Strobel (2007). Real effort versus chosen effort in experiments. *Economics Letters 96*(2), 232–236.

Buccirossi, P., G. Immordino, and G. Spagnolo (2017). Whistleblower rewards, false reports, and corporate fraud. *CSEF Working Papers 477*.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Butler, J. V., D. Serra, and G. Spagnolo (2018). Motivating whistleblowers. *Management Science forthcoming*.

Callahan, E. S. and T. M. Dworkin (1992). Do good and get rich: Financial incentives for whistleblowing and the False Claims Act. *Villanova Law Review 37*, 273.

Carmichael, F. and D. Thomas (1995). Production and efficiency in team sports: An investigation of rugby league football. *Applied Economics 27*(9), 859–869.

Cassematis, P. G. and R. Wortley (2013). Prediction of whistleblowing or non-reporting observation: The role of personal and situational factors. *Journal of Business Ethics 117*(3), 615–634.

Charness, G. and M. Dufwenberg (2006). Promises and partnership. *Econometrica 74*(6), 1579–1601.

Charness, G., U. Gneezy, and M. A. Kuhn (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization 81*(1), 1–8.

Charness, G. and M. Sutter (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives 26*(3), 157–176.

Chassang, S. and G. P. I. Miquel (2018). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies forthcoming*.

Childs, J. (2012). Gender differences in lying. *Economics Letters 114*(2), 147–149.

Chytilova, J. and V. Korbel (2014). Individual and group cheating behavior: A field experiment with adolescents. *Charles University Prague, IES Working Paper 06/2014*.

Conrads, J., B. Irlenbusch, R. M. Rilke, A. Schielke, and G. Walkowitz (2014). Honesty in tournaments. *Economics Letters 123*(1), 90–93.

Conrads, J., B. Irlenbusch, R. M. Rilke, and G. Walkowitz (2013). Lying and team incentives. *Journal of Economic Psychology 34*, 1–7.

Cox, T. H., S. A. Lobel, and P. L. McLeod (1991). Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of Management Journal 34*(4), 827–847.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47*(2), 448–474.

Dawson, P. and S. Dobson (2002). Managerial efficiency and human capital: An application to English association football. *Managerial and Decision Economics 23*(8), 471–486.

Dawson, P., S. Dobson, and B. Gerrard (2000a). Estimating coaching efficiency in professional team sports: Evidence from English association football. *Scottish Journal of Political Economy 47*(4), 399–421.

Dawson, P., S. Dobson, and B. Gerrard (2000b). Stochastic frontiers and the temporal structure of managerial efficiency in English soccer. *Journal of Sports Economics 1*(4), 341–362.

De Paola, M. and V. Scoppa (2012). The effects of managerial turnover: Evidence from coach dismissals in Italian soccer teams. *Journal of Sports Economics 13*(2), 152–168.

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

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550.

Dreber, A. and M. Johannesson (2008). Gender differences in deception. *Economics Letters 99*(1), 197–199.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–74.

Dufwenberg, M. and A. Muren (2006). Gender composition in teams. *Journal of Economic Behavior & Organization 61*(1), 50–54.

Dworkin, T. and J. Near (1997). A better statutory approach to whistle-blowing. *Business Ethics Quarterly 7*(1), 1–16.

Dyck, A., A. Morse, and L. Zingales (2010). Who blows the whistle on corporate fraud? *The Journal of Finance 65*(6), 2213–2253.

Dyck, A., A. Morse, and L. Zingales (2017). How pervasive is corporate fraud? *Chicago Booth School of Business, mimeo*.

Dyreng, S. D., M. Hanlon, and E. L. Maydew (2010). The effects of executives on corporate tax avoidance. *The Accounting Review 85*(4), 1163–1189.

Eckel, C. C. and P. J. Grossman (2008). Differences in the economic decisions of men and women: Experimental evidence. In C. R. Plott and V. L. Smith (Eds.), *Handbook of Experimental Economics Results*, Volume 1. Elsevier.

Economist Intelligence Unit (2015). Global fraud report, 2015–2016. Technical report, Kroll, London.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability, Vol. 57*. New York and London: Chapman and Hall/CRC.

Erat, S. and U. Gneezy (2012). White lies. *Management Science 58*(4), 723–733.

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

Falk, A. and N. Szech (2013). Morals and markets. *Science 340*(6133), 707–711.

Fehr, E. and U. Fischbacher (2002). Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal 112*(478), C1–C33.

Fehr, E. and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), 63–87.

Fehr, E., U. Fischbacher, and S. Gächter (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature 13*(1), 1–25.

Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature 415*(6868), 137–140.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*, 817–868.

Fehr, E. and K. M. Schmidt (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity 1*, 615–691.

Felli, L. and R. Hortala-Vallve (2016). Collusion, blackmail and whistle-blowing. *Quarterly Journal of Political Science 11*(3), 279–312.

Feltovich, N. and Y. Hamaguchi (2018). The effect of whistle-blowing incentives on collusion: An experimental study of leniency programs. *Southern Economic Journal 84*(4), 1024–1049.

Finucane, M. L., P. Slovic, C. K. Mertz, J. Flynn, and T. A. Satterfield (2000). Gender, race, and perceived risk: The 'white male' effect. *Health, Risk & Society 2*(2), 159–172.

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

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Fizel, J. L. and M. P. D'Itry (1997). Managerial efficiency, managerial succession and organizational performance. *Managerial and Decision Economics 18*, 295–308.

Frick, B. and R. Simmons (2008). The impact of managerial quality on organizational performance: Evidence from German soccer. *Managerial and Decision Economics 29*(7), 593–600.

Friebel, G. and S. Guriev (2012). Whistle-blowing and incentives in firms. *Journal of Economics & Management Strategy 21*(4), 1007–1027.

Gerhards, L. and N. Siemer (2016). The impact of private and public feedback on worker performance–evidence from the lab. *Economic Inquiry 54*(2), 1188–1201.

Gibbons, F. and B. Buunk (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. *Journal of Personality and Social Psychology 76*(1), 129.

Gibbons, R. and J. Roberts (2012). *The Handbook of Organizational Economics*. Princeton University Press.

Gibson, R., C. Tanner, and A. F. Wagner (2013). Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review 103*(1), 532–548.

Gill, D., Z. Kissová, J. Lee, and V. Prowse (2018). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science 65*(2), 494–507.

Givati, Y. (2016). A theory of whistleblower rewards. *The Journal of Legal Studies 45*(1), 43–72.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review 95*(1), 384–394.

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

Gneezy, U., S. Meier, and P. Rey-Biel (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives 25*(4), 191–210.

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

Goodall, A. H., L. M. Kahn, and A. J. Oswald (2011). Why do leaders matter? A study of expert knowledge in a superstar setting. *Journal of Economic Behavior & Organization 77*(3), 265–284.

Graham, J. R., S. Li, and J. Qiu (2012). Managerial attributes and executive compensation. *Review of Financial Studies 25*(1), 144–186.

Guimond, S., N. Branscombe, S. Brunot, A. Buunk, A. Chatard, M. Désert, D. Garcia, S. Haque, D. Martinot, and V. Yzerbyt (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology 92*(6), 1118.

Hall, S., S. Szymanski, and A. Zimbalist (2002). Testing causality between team performance and payroll: the cases of Major League Baseball and English soccer. *Journal of Sports Economics 3*(2), 149–168.

Harrison, G. W. and E. E. Rutstrom (2008). Risk aversion in the laboratory. *Research in Experimental Economics 12*, 41–196.

Hart, O. and L. Zingales (2017). Companies should maximize shareholder welfare not market value. *Journal of Law, Finance, and Accounting 2*(2), 247–275.

Healy, P. M. and K. G. Palepu (2003). The fall of Enron. *Journal of Economic Perspectives 17*(2), 3–26.

Herbst, D. and A. Mas (2015). Peer effects on worker output in the laboratory generalize to the field. *Science 350*(6260), 545–549.

Herrmann, B., C. Thöni, and S. Gächter (2008). Antisocial punishment across societies. *Science 319*(5868), 1362–1367.

Heyes, A. and S. Kapur (2009). An economic model of whistle-blower policy. *Journal of Law, Economics, and Organization 25*(1), 157–182.

Hinloopen, J. and A. R. Soetevent (2008). Laboratory evidence on the effectiveness of corporate leniency programs. *The RAND Journal of Economics 39*(2), 607–616.

Hofler, R. A. and J. E. Payne (2006). Efficiency in the National Basketball Association: A stochastic frontier approach with panel data. *Managerial and Decision Economics 27*(4), 279–285.

Houser, D., S. Vetter, and J. Winter (2012). Fairness and cheating. *European Economic Review 56*(8), 1645–1655.

Howse, R. and R. Daniels (1995). Rewarding whistleblowers: The costs and benefits of an incentive-based compliance strategy. In R. Daniels and R. Morck (Eds.), *Corporate Decisionmaking in Canada*. Calgery: University of Calgary Press.

Jones, D. and S. Linardi (2014). Wallflowers: Experimental evidence of an aversion to standing out. *Management Science 60*(7), 1757–1771.

Kahane, L. H. (2005). Production efficiency and discriminatory hiring practices in the National Hockey League: A stochastic frontier approach. *Review of Industrial Organization 27*(1), 47–71.

Kohn, S. M., M. D. Kohn, and D. K. Colapinto (2004). *Whistleblower law: A Guide to Legal Protections for Corporate Employees*. Greenwood Publishing Group.

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

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

Kugler, T., E. E. Kausel, and M. G. Kocher (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science 3*(4), 471–482.

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

Kuper, S. and S. Szymanski (2009). *Soccernomics*. Nation Books, New York, NY.

Lazear, E. P., K. L. Shaw, and C. T. Stanton (2015). The value of bosses. *Journal of Labor Economics 33*(4), 823–861.

Leibenstein, H. (1966). Allocative efficiency vs. "x-efficiency". *The American Economic Review 56*(3), 392–415.

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics 9*(2), 508–523.

Maciejovsky, B., M. Sutter, D. V. Budescu, and P. Bernau (2013). Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Science 59*(6), 1255–1270.

Marvão, C. M. P. and G. Spagnolo (2014). What do we know about the effectiveness of leniency policies? a survey of the empirical and experimental evidence. *University of Stockholm, SITE Working Paper No. 28.*.

Mazar, N., O. Amir, and D. Ariely (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research 45*(6), 633–644.

Mechtenberg, L., G. Muehlheusser, and A. Roider (2017). Whistle-blower protection: Theory and experimental evidence. *IZA Discussion Paper 10607*.

Meier, V. and G. Schütz (2008). The economics of tracking and non-tracking. *Zeitschrift für Betriebswirtschaft Special Issue 1*, 23–43.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for Experimental Economics*. Macmillan International Higher Education.

Muehlheusser, G., A. Roider, and N. Wallmeier (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters 128*, 25–29.

Muehlheusser, G., S. Schneemann, and D. Sliwka (2016). The impact of managerial change on performance: The role of team heterogeneity. *Economic Inquiry 54*(2), 1128–1149.

Muehlheusser, G., S. Schneemann, D. Sliwka, and N. Wallmeier (2018). The contribution of managers to organizational success: Evidence from German soccer. *Journal of Sports Economics 19*(6), 786–819.

Near, J. P. and M. P. Miceli (1985). Organizational dissidence: The case of whistle-blowing. *Journal of Business Ethics 4*(1), 1–16.

Near, J. P. and M. P. Miceli (1986). Retaliation against whistle blowers: Predictors and effects. *Journal of Applied Psychology 71*(1), 137.

Niederle, M. (2016). Gender. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, Volume 2. Princeton University Press.

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

OECD (2016). *Committing to Effective Whistleblower Protection*. Paris: OECD Publishing.

Pruckner, G. J. and R. Sausgruber (2013). Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association 11*(3), 661–679.

Rees, D. I., D. J. Brewer, and L. M. Argys (2000). How should we measure the effect of ability grouping on student performance? *Economics of Education Review 19*(1), 17–20.

Rehg, M., M. P. Miceli, J. Near, and J. Van Scotter (2008). Antecedents and outcomes of retaliation against whistleblowers: Gender differences and power relationships. *Organization Science 19*(2), 221–240.

Reuben, E. and M. Stephenson (2013). Nobody likes a rat: On the willingness to report lies and the consequences thereof. *Journal of Economic Behavior & Organization 93*, 384–391.

Rosen, S. (1982). Authority, control, and the distribution of earnings. *The Bell Journal of Economics 13*(2), 311–323.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review 81*(5), 1068–1095.

Schmolke, K. U. and V. Utikal (2016). Whistleblowing: Incentives and situational determinants. *FAU - Discussion Papers in Economics No. 09/16*.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In *Beiträge zur experimentellen Wirtschaftsforschung*, pp. 136–168. Tübingen: JCB Mohr (Paul Siebeck).

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research 60*(3), 471–499.

Spagnolo, G. (2008). Leniency and whistleblowers in antitrust. In *Handbook of Antitrust Economics*. MIT Press.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature 49*(2), 326–65.

Szymanski, S. and R. Smith (1997). The English football industry: Profits, performance and industry structure. *International Review of Applied Economics 11*(1), 135–153.

Tena, J. d. D. and D. Forrest (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. *European Journal of Operational Research 181*(1), 362–373.

Thiemann, K. (2017). Ability tracking or comprehensive schooling? A theory on peer effects in competitive and non-competitive cultures. *Journal of Economic Behavior & Organization 137*, 214–231.

Thüsing, G. and G. Forst (2016). *Whistleblowing–A Comparative Study*, Volume 16. Springer.

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

Walker, F. A. (1887). The source of business profits. *The Quarterly Journal of Economics (1886-1906)*, 265.