Economic Behavior in Real Effort Experiments

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Marion Eberlein
aus Troisdorf

Bonn 2008

Dekan:	Prof. Dr. Christian Hillgruber
Erstreferent:	Prof. Dr. Matthias Kräkel
Zweitreferent:	Prof. Dr. Erik Theissen
Tag der mündlichen Prüfung:	23.10.2008

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Acknowledgements

For valuable comments and discussions I would like to thank (in alphabetical order): Jürgen Arns, Almut Balleer, Eva Benz, Benjamin Born, Jörg Breitung, Thomas Dohmen, René Fahr, Armin Falk, Simon Gächter, Sebastian Goerg, Oliver Gürtler, Bernd Irlenbusch, Johannes Kaiser, Matthias Kräkel, Sandra Ludwig, Julia Nafziger, Robert J. Oxoby, Judith Przemeck, Anja Schöttner, Reinhard Selten, Dirk Sliwka, Erik Theissen, Klaus Utikal, and Gari Walkowitz. Moreover, financial support from the Deutsche Forschungsgemeinschaft (DFG), grant KR 2077/2-3 and IR 43/1-1, is gratefully acknowledged.

Table of Content

Chapter 4

Chapter 5

List of Tables

List of Figures

Chapter 1

Overview

The development of appropriate methods for analyzing human behavior is a central issue in the social sciences. In *empirical* economics, two basic kinds of methods can be distinguished. The initial method is observation of natural occurrences in economy. The data most commonly used are gathered in the field by chance ("happenstance" data), such as national income accounts, unemployment rates, wages, inflation rates, etc. They are a by-product of ongoing uncontrolled processes. Still a great deal of research is based on fundamentals derived in that way. Another method is based on experimental data which are created under controlled conditions in the laboratory, explicitly for scientific purposes. Laboratory experimental data, which are gathered in a controlled manner in an artificial environment, have gained more and more importance in economics in the last twenty years.

To investigate a certain research question, there are many cases where happenstance data are adequate and cheap. Then experiments are not worthwhile. But in a lot of cases happenstance data are inadequate and experimental data can be obtained at reasonable cost. According to Friedman and Sunder (1994) such cases present the best opportunities for experimental work. In particular, experimental laboratory data are relative easy to interpret. One reason for this is that ceteris paribus changes are possible and therefore causal inferences easier to identify. The experimenter controls the conditions and documents the instructions. Hence, another experimenter can easily replicate a certain experiment and can therefore support or reject its claimed results.

The superior control possibilities of laboratory experiments are not called in doubt. But contrarily to field experiments, by which the experimenter investigates subjects' behavior in their natural environment, the question whether the conditions implemented in the laboratory are also present in reality will probably always be subject to some uncertainty and debate (see Falk and Fehr 2003). Not infrequently, a lack of realism is complained, that is, that the design of the laboratory environment does not closely enough reflect a real-

world environment of substantive economic interest. One way of adding more realism to a laboratory experiment is to conduct so-called "real effort" experiments. In a typical laboratory experiment, the choice of work effort is represented by an increasing monetary function, i.e., instead of choosing real efforts, subjects choose a costly number. In real effort experiments, subjects have to exert efforts in working on an actual task.

The experiments, which are presented in this dissertation, all include real effort tasks that subjects have to work on and are paid for. For example, subjects have to solve arithmetic products or answer general-knowledge questions. With this, specific characteristics in the population, such as for example overconfidence, which plays a role in chapters 3 and 4, can be realistically investigated.

The reminder of the dissertation proceeds as follows:

In Chapter 2 it is experimentally investigated whether the so-called in-group/out-group bias leads to a favoring of own team members as candidates in promotion (by voting for them) relative to other teams and their members. In contrast to psychological approaches, monetary incentives for voting choices are implemented and objective performance criteria defined and thus the extent of the in-group/out-group bias is exactly measured. The data show that face-to-face interaction with team members leads more subjects to favor own team-mates than in anonymous interaction. Moreover, not only the frequency but also the average extent of positive team identity is higher with face-to-face interaction according to objective performance measures. A further finding suggests that only anonymous team interaction often leads to substantial discrimination of own team members (i.e. negative team identity), which also is an interesting new finding and extends previous findings of psychologists on the in-group/out-group bias.

Chapter 3 deals with an experiment which investigates managers' self-predictions of their subsequent performance and, based upon, their choice of a collaborator. The data show that managers' self-predictions are not biased anymore after they have been informed about the performance of a reference group. In spite of this, most managers do not rationally choose a collaborator given their beliefs. In a second treatment, superiors (who are assumed to be at a higher hierarchy level than the managers) obtain various information, e.g. about managers' self-predictions, and have to predict the managers' performances. The data

show that superiors adapt their predictions into the direction of the managers' self-predictions, although not completely. Particularly, superiors think that their managers' self-predictions are biased if they are lower than the average performance of the reference group. Based upon their predictions, superiors have to select a collaborator for their managers. The data show that superiors' collaborator choices do not significantly differ from the managers' choices. This proves due to excellent information processing by both, managers and superiors, which on the whole leads to very similar predictions of managers' subsequent performance.

It is a well-known phenomenon that people have difficulties to assess their ability correctly: Often they overestimate their (relative) abilities. Chapter 4 deals with an experiment that investigates whether the self-assessments of subjects improve when they receive feedback and there are incentives to make a correct self-assessment. Subjects' reactions to feedback in several subsequent rounds are investigated to see not only if, but also when and how they react. The data show that feedback influences subjects' self-assessments but that the total improvement over time is just moderate. Furthermore, differences in the reactions of subjects (e.g. to what extent they follow feedback), in the robustness of their belief about their relative ability, and how they process feedback are observed. The effects depend on the kind of feedback.

Chapter 5 investigates the influence of real effort and performance differences on donating behavior in an experiment that introduces a real effort task into the solidarity game. In a tournament, performance on this real effort task determines rank, initial income distribution, and losing probability in a subsequent lottery. This design enables me to investigate if donators take into account performance differences between losers of the lottery and between themselves and a loser. To control for the influence of performance on donators and recipients, I run a further treatment in which subjects do not have to work on the real effort task and the assignments to an income and to a losing probability are random. A further additional treatment which combines the real effort and the non-real effort treatment is conducted to clearly separate between countervailing real effort effects. The data reveal different kinds of donating behavior.

Chapter 2 [1]

Positive and Negative Team Identity in a Promotion Game

2.1 Introduction

Most company representatives emphasize the importance of creating team spirit or team identity for a good team performance. Generally, team identity is regarded as a positive value especially in the labor environment. But there is also a dark side of team identity that may appear when a manager has to reach or take part in decisions on a business matter that does not only concern his own team and, in particular, could cause comparative disadvantages for his team. Because of team identity he may in this case use all his influence for a decision in favor of his team, even if this does not lead to an efficient decision as a whole.

An application to a particular case is an enterprise where the responsible heads of department are meeting in order to select one of their staffs as a candidate for a management recruiting program. The complete personal records of each staff are assumed to be known. Their performance is consequently well-known to each decision-maker. Then the question has to be raised whether the heads of department prefer to choose a staff of their own department in spite of obvious criteria in favor of staffs from other departments.

These aspects of team identity - the so called in-group/out-group bias, which leads to a favoring of the own group and their members relative to another group's members - have been investigated in a few experimental economic studies: Charness et al. (2007) investigate whether group membership and the saliency of the own group influences behavior in the Battle of the Sexes and the Prisoner's Dilemma Game. They find that significantly more subjects choose the strategy which maximizes the payoff of the own group when the group is most salient compared to when it is less salient. In the treatment where groups are most salient, the group of one player (who plays against someone from the out-group) sits as audience in the same room and is informed about the player's payoff,

[1] This chapter is based on Eberlein and Walkowitz (2008).

which also influences the group's payoff. Chen and Li (2007) investigate how team identity affects various dimensions of social preferences in the Dictator Game and in games with a response possibility like, e.g., reward or punishment. They show, besides other results, that subjects show more charity and less behindness-aversion to in-group than to out-group members. In addition to that, subjects reward good intentions of in-group members more than those of out-group members and forgive bad intentions of in-group members more than those of out-group members. Furthermore, Eckel and Grossman (2005), Goette et al. (2006**)** and McLeish and Oxoby (2007) investigate whether subjects behave more cooperatively to members of the in-group than to members of the other group (in a Public Goods Game respectively in Prisoner's Dilemma Game respectively in a bargaining game). As a main result, each of these studies shows that subjects behave significantly more cooperatively to their in-group mates than to out-group members.[2] In the experimental study, which is presented in this chapter, I want to abstract from aspects of two-person-games but consider a promotion decision problem that involves more than one representative of each team.

There are much more experimental studies which focus on the favoring of own group members relative to out-group members in psychology. In this field, the in-group/out-group bias has extensively been investigated since the sixties. There is a considerable research on the assessment of group members' attitudes and characteristics in comparison to other groups (e.g. Sherif et al. 1961, Rabbie and Horwitz 1969, Tajfel et al. 1971, Doise and Sinclair 1973, Tajfel and Turner 1979, Rabbie et al. 1989, Mummendey et al. 1992, Schaller 1992, Wit and Wielke 1992, Yamagishi et al. 1999, Otten and Wentura 1999, Crisp and Hewstone 2000). As a common and consistent result it is shown that the own group is assessed more positive than the out-group. The subjects' attitude concerning the evaluation of the own group is substantially more affirmative. There are much fewer investigations in the field of psychology with a view to the evaluation of performance (e.g. Sherif et al. 1961, Ferguson and Kelley 1964, Brewer 1979, Kraiger and Ford 1985,

[2] There are many laboratory economic experiments focusing on group identity in real groups, which also exist outside the laboratory, as for example nationalities. For example, in the investment game, Buchan et al. (2006) find that American students exhibit an in-group bias: They are more willing to trust other American students than students of other nationalities. Ruffle and Sosis (2006) observe that kibbutz members cooperate more with members of their own kibbutz than with city residents. Solow and Kirkwood (2002) compare contributions in a Public Goods Game between members of a real group, members of an experimentally created group, and strangers. They only find significantly higher contributions when comparing contributions of members of a real group with contributions of strangers.

Downing and Monaco 1986). As a general result it can also be stated that the performance of the own group is assessed more favorably than the one of other groups.

To the best of my knowledge, none of the studies so far carried out on team identity focuses on the aspect of competition within teams, next to competition between teams. It is not clear if team identity endures when also the own team-mates compete for attaining a specific goal. Will people then favor their team-mates in relation to other groups' team-mates or does competition neutralize team identity or even lead to a discrimination of own team-mates? Imagine a subject who does not win a tournament for promotion or just thinks that his chances are very small. For this subject it may be even worse when one of his team-mates wins the tournament and not a third person from another team, who is perhaps not even personally known to this person.[3] The closer relationship to the team-mate on the one hand and the distance and more abstract perception of the out-group member on the other hand, exposes this subject to a higher action level of envy, revenge, or even discrimination against the own team-mate (see ,e.g., Suls and Wheeler (2000) for an overview of Social Comparison Theory).

My design enables the investigation of both positive and negative team identity[4] by incorporating competition – also between team-mates – to a promotion decision problem. From an efficiency point of view, the subjects with the best performances shall be promoted to an elite-team in my experiment. Each subject has to assign promotion points to each of the other subjects. These promotion decisions may be distorted for two different reasons. First, the performance evaluation of certain subjects may be influenced by their membership to a certain group, e.g., a subject may favor his own team-mates for promotion although they are objectively not good enough to join the elite-team. Second, a subject who wants to join the elite-team himself may try to discriminate his team-mates so that his chance to get into the elite-team becomes higher than theirs.[5] Within the experimental

[3] A similar aspect is discussed by Grund and Sliwka (2005) in a theoretical tournament model with inequity averse competitors. The authors argue that lateral promotions reduce inequity costs compared to vertical promotions: In vertical promotions the winner of the promotion tournament becomes the superior of his former colleague and both are faced with the result of the tournament permanently. Contrarily, in lateral promotions neither the winner nor the loser face their former colleague anymore. Thus, feelings of envy (and compassion) are higher in vertical than in lateral promotions.

[4] With "positive team identity" I refer to the favoring of own team-mates, with "negative team identity" to the discrimination of own team-mates. A more detailed definition is given in subsection 2.3.2.

[5] This aspect is related to sabotage in tournaments. In the respective literature competitors cannot only try to win the tournament by exerting much productive effort and therefore reaching high output, but can also try to reduce the competitors' output by sabotaging them. Because only the relative order of the amount of outputs

design of my work, I can isolate and analyze such distortion effects on performance evaluation of in-group and out-group members.

Compared to the psychologists' approach, I apply an experimental procedure that has the following advantages: First, in contrast to psychological analyses, objective performance criteria are defined and thus the extent of the in-group/out-group bias exactly measured. In the psychological studies, the authors only quantify to what degree the own group is evaluated higher than others. But there is no objective evaluation criterion that enables the authors to measure the exact degree of team identity. For example Downing and Monaco (1986) investigate whether subjects evaluate the own team members better than the out-group when they have to show ski exercises. Then the authors compare whether the own group or the other groups receive better evaluations, but an objective criterion which group is actually better is missing. However, in my experiment, subjects solve multiplication tasks and I can exactly rank them according to their performance. Hence, objectively, it is clear who should be elected into the elite-team in my experiment. Second, in contrast to psychological studies I provide monetary incentives which counteract in-group favoring. Each subject is paid an extra amount of money which depends on the performance of the elite-team. Hence, an incentive is given to elect subjects with the best performances into the elite-team. If, in spite of this environment, a team favoring is observed, it provides a stronger evidence for team identity than without monetary incentives.

My data show that face-to-face interaction between team members leads more subjects to favor own team-mates than in anonymous interaction. Moreover, not only the frequency of positive team identity is higher but also the average extent. There are not only more subjects who favor own team-mates with face-to face interaction, but own team-mates are favored to a higher degree according to objective performance measures. This aspect is not investigated by psychological studies on team identity. Another striking result suggests that anonymous group interaction often leads to discrimination of own team members, i.e. negative team identity. This is also an interesting new finding that enriches previous findings of psychologists and enhances the understanding of team interaction.

---

between the rivals determines who wins the tournament, sabotage can be optimal for a competitor. For theoretical literature on sabotage in tournaments see, e.g., Lazear (1989), Kräkel (2005), and Gürtler (2008). For experimental studies in this field, see, for example, Harbring et al. (2007), and Harbring and Irlenbusch (2008).

The remainder of this chapter is organized as follows: In the next section, I present my experimental design, treatments, and procedure. Afterwards, I present the results. The final section discusses my findings and concludes.

2.2. Experimental Design, Treatments, and Procedure

2.2.1 Basic Experimental Design

The basic experimental design consists of three subsequent stages: i) First real effort performance stage, ii) Voting-for-promotion stage, iii) Second real effort performance stage. At the beginning of the first stage, each subject draws an individual code which is kept secretly and guarantees full anonymity. In the following, I describe the features of each of these stages in detail.

i) First Real Effort Performance Stage (Stage 1)

At the first stage, each of twelve invited subjects gets a list with 100 arithmetical calculations to be carried out within a time limit of fifteen minutes. The set is made up of 100 simple products of numbers which can be multiplied easily. For each correct solution subjects earn 3 Cents. Auxiliary means except pen and paper are not permitted.

ii) Voting-for-Promotion Stage (Stage 2)

After the multiplication work, subjects are instructed about the remaining two stages of the experiment. In particular, they are informed that, at stage 3, equally structured and comparable arithmetical tasks have to be solved, for which again fifteen minutes of time are provided and 3 Cents for each correct solution are paid. In contrast to the first stage, there are two different roles for the players at stage 3: role A and role B with respectively five and seven players. The five individuals playing in role A receive a bonus of 10 € each. Contrarily, those seven subjects playing in role B do not get such a bonus. Furthermore, each player (independent of role A or B) earns an extra payment, in addition to his individual achievement, that depends on the performance of the players in role A. It amounts to the average number of multiplications correctly calculated by the players in role A at stage 3 multiplied by 3 Cents. By this mechanism an incentive is given to the players to have the best five of them in role A. On the other hand, each subject has an

incentive to be in role A himself. The subjects do not learn which role they played until the end of the experiment.[6]

Before the calculations of the second set of multiplication tasks are actually carried out, the players for role A are elected by all the twelve participants of the experimental session. After the participants have been instructed about this fact and also about stage 3, a piece of paper is handed out to everyone containing a list of the code names of the other eleven subjects together with the number of products they have solved correctly at stage 1. Subjects are not informed about their own performance, which could otherwise influence their voting. In order to decide upon the role of each player, every individual has to propose which subjects shall act in role A. No player can nominate himself. According to Borda's rule, 11 points shall be attached to the code name of that subject who is preferred most for acting in role A, 10 points to the person who is preferred second most, etc. To the code name of that person who is preferred least for role A, 1 point is to be assigned. Subsequently, the subjects are ranked according to the total number of points they received from all their fellows. Those five persons with the most points are appointed to role A.[7] The others have to act in role B.[8]

Because the subjects get an extra payment which depends upon the average achievement of the members of the elite-team (players in role A), there is an incentive to choose the most qualified persons for the elite-team. But also another, a "strategic" behavior is possible: In order to increase their own chances to get into the elite-team, subjects might try to weaken the position of their strongest competitors. In the most extreme case they might assign only 1 point to the player with the best achievement and 11 points to the player with the worst achievement in stage 1 (in the following I will refer to this strategy as the *competitive* one).[9]

---

[6] This is to prevent a possible influence on a player's effort when he knows that he has already earned the 10 € reward as player in role A.

[7] See, e.g., Young (1974) for an axiomatization of Borda's rule. As Borda's rule violates the principle of "independence of irrelevant alternatives" it is possible to manipulate its outcome by introducing extraneous alternatives (see, e.g., Young 1995). In my experiment, no additional alternatives (i.e. other subjects) are available – the number of teams and team members is constant.

[8] In case of equality of points, which does not allow to clearly define the five persons with the most points, a random decision is made. But this case never occurred.

[9] Although it is no equilibrium that every subject uses the competitive strategy, subjects might use this strategy because they think that it is profitable.

As subjects have been informed that the multiplication tasks they have to solve at the next stage are equally structured to those they have already solved, they know that someone who performed very well at stage 1 is likely to perform very well at stage 3, too. Subjects therefore know who should – objectively – join the elite-team. This also means that they can estimate the resulting costs (i.e. the lower extra payment) if they instead favor a team-mate for the elite-team who does not belong to the five best ones.

iii) Second Real Effort Performance Stage (Stage 3)

At this stage, subjects have to carry out multiplication tasks within a fifteen minutes' time limit. The calculation tasks are the same as at stage 1, but in another order. This allows me to exactly compare the performances at stage 1 and 3. For each correct answer 3 Cents are paid.

2.2.2 Experimental Treatments and Procedure

I apply three different treatment conditions to investigate which environment creates team identity and by which means this is affected, especially with having in view the favoring or discrimination of own team-mates as distinctive essential outcomes. Each of my treatments is equally structured in the three stages described above, but different in team identity enhancing factors.

Anonymous Treatment (ANT)

In a session of the first treatment, the Anonymous Treatment (ANT), twelve subjects are randomly assigned to four groups (i.e. teams), each made up of three team members. However, neither team members nor members of any other team are able to see each other after team division or to have the opportunity to communicate at any time during the experiment.

The scores of the multiplication tasks of the first stage are summed up for all the members of a team so that a common team score can be determined. These team scores are compared between the four teams. Each subject of the team with the highest team score receives a team bonus of 5 €. By this a bit of (presumably weak) team identity may be

generated because of common economic interest.[10] Which team finally gets the team bonus is announced not before the end of the experiment. In that way a difference in team identity, which may be caused by a different angle of winners and losers, shall be avoided.

But already at this early phase of the experiment, the subjects are informed about how their team results are assessed. In detail, each subject is informed about how many right calculations his team fellows achieved together, but is not told how many of the tasks he himself has solved correctly.

After every subject has received such an intermediate status of specific team performance, they are instructed about the remaining stages of the experiment and get the voting list thereafter. By this list, subjects learn all other subjects' code names and scores and whether they belong to their own team or not. Subjects not adjoined to the own team cannot be connected with a certain other team. Therefore, none of the other team scores is known.[11] Taken together, by the team score competition and through the handling of team membership – using the denotion "group" in the instructions and adjoining subjects to a certain team per se – subjects are given the possibility to develop some degree of team identity. Special emphasis lies hereby on economic team identity, which can arise from the common will to win the team bonus.

Group Treatment (GRT)

By the GRT, I expand the spectrum of possible team identity in the way that I also introduce social factors in addition to the same factors already applied in ANT. For this, I organize a comprehensive communication phase prior to stage 1, where members of each team are given the opportunity to meet with their team fellows and to talk face-to-face in a separate room. They are admitted and advised by an assigned experiment assistant, who is responsible for this specific room during the whole experiment and can give supplementary support. In order to support and enhance social identity the team members in each room are asked to carry out some tasks together. By performing a simple task together and directing effort to a common objective, I want to ensure that the subjects get somewhat acquainted.

---

[10] Psychological studies show that a "minimal group condition" can already result in favoring members of the own group (see Bewer (1979) for an overview). Psychologists usually employ as a minimum group condition a reasonable division into groups, for example Doise and Sinclair (1973) tell subjects that they are divided into two groups according to their preference to one of two photographs.

[11] This is to prevent a possible influence on voting, e.g., by disadvantaging the team with the best team score.

First, the three members of a team have to determine a name for their team. This name has to consist of three words, each beginning with the initials of one of the team members' forenames. After having agreed on a team name, they have to write it on a cardboard and affix it outside to the door. While the team is thinking about an appropriate name, the experiment assistant is leaving the room in order to avoid disturbance of the social team identity finding process. Thereafter, subjects have to read two short newspaper articles about the lawn of the soccer world cup 2006 in Germany and about the first bio-energy village in Germany. For these tasks they get five minutes' time. Only two copies of the articles are provided for each three-person-team to enhance cooperation and coordination among the team members. Then, the team gets three minutes to deal with a fragmentary text with statements concerning the text read before. For each right answer completing the lacking parts of the article, 50 Cents are paid and directly donated to the Beethoven Memorial House in Bonn.[12] By donation instead of paying off, I want to foster both subjects' effort and high-minded approach.[13]

The described communication phase is used as an instrument for the process of establishing a social identity. It is important to create a relaxed atmosphere prevailing during the whole phase. To support this intention, team rooms allow sitting around a shared table. After completing the communication phase, the economic phase begins, which is identical to ANT. During this stage, the subjects must not talk and have to use their anonymous code names, which the other team members do not know and cannot link to the other two persons of the team.[14] Each team member is now seated at a separate table in the team room.

---

[12] I choose a social objective in form of the donation to the Beethoven Memorial House because it stresses the regional relation and thus may strengthen identity.

[13] I could, of course, pay off the earnings immediately to the subjects but that would correspond more to an economic identity, which I want to attain at a later point of time by paying the team bonus. Moreover, I want to avoid a shifting of marginal incentives for elaborating the multiplication tasks, which may happen if already the communication phase is profitable.

[14] In my experiment I have to distribute papers with certain information to certain subjects when informing them about the scores of their two team-mates and about the contents of their voting list (remember that a subject's score does not appear on his list). At distributing I have to be careful not to hurt anonymity. Therefore, I use the following procedure: The assistant in the room picks up a subject's secret code name, which is written on a piece of paper and placed folded on the subject's desks. The code name cannot be read by the assistant or the other two persons in the room. Afterwards, the assistant takes it to the experimenter waiting outside the room, who then finds out the associated paper, which is then given to the subject by the assistant together with the code picked up before. Again this is done in the same secret way as at the start of the procedure so that the other two fellows and the assistant herself cannot read the information. Also the experimenter is not completely informed because, through this procedure, he cannot see any of the subjects in the room so that he is incapable of associating code names with persons. By this it is ensured that each subject gets the necessary particular information without that anyone gives up anonymity.

Writing the instructions for both ANT and GRT, I use the term "group" only to describe the formal division into different teams. This is done very carefully to avoid any framing bias or demand effects.[15]

Individual Control Treatment (ICT)

In ICT, I exclude all factors which may cause economic or social team identity. So the ICT includes only the basic design and subjects are not separated into different groups. I use this treatment as a reference for the experimental treatments ANT and GRT. By a comparison, I am particularly able to analyze whether the underlying voting mechanism leading either to an objective or a competitive strategy is influenced by my treatment conditions.

Table 2.1 shows an overview of my treatments and its features.

Table 2.1: Treatment specifics

	Economic identity				**Social identity**			
	Team features				**Communication phase**			
Treat-ment	**creating „teams“**	**team bonus**	**information about team performance**	**information about team composition**	**see/ talk**	**joint task**	**joint feedback**	**group name**
ICT	-	-	-	-	-	-	-	-
ANT	+	+	+	+	-	-	-	-
GRT	+	+	+	+	+	+	+	+

The pen and paper experiment was conducted at the Rheinische Friedrich-Wilhelms-University Bonn. All individuals were students from this university. Subjects were recruited via the internet by using ORSEE software (Greiner, 2004) announcing the possibility to earn an amount of money dependent on their behavior. The treatments lasted about 90 – 120 minutes. On average, subjects earned 15.50 €. The Individual Control Treatment (ICT) consisted of two sessions with twelve subjects per session. The other two treatments (ANT and GRT) were made up of three sessions with a total of 36 participants per treatment.

[15] The instructions of GRT can be found in the appendix, section 2.5

2.3 Results

In this section I present the results. First, I analyze the strategies that underlie subjects' voting behavior. Next, I explore whether team identity has a substantial impact on subjects' votes, using the underlying strategies as a reference for comparison. I develop quantitative and qualitative measures to investigate team identity. Finally, I enter into a discussion about the efficiency of the composition of the elite-team.

2.3.1 Reference Strategies

When investigating whether subjects favor or discriminate their team-mates, a reference strategy is needed to which such deviations can be measured. I distinguish two strategies which can serve as a reference: the *objective* and the *competitive* one. The individual degree of the observed in-group/out-group bias can differ depending on what reference strategy is pursued. A strict objective strategy is defined as a ranking where the subject with the best performance gets the most (i.e. 11) points, the one with the second best performance the second most (i.e. 10) points, etc. Hence, the participant with the lowest performance receives the least (i.e. 1) points. Contrarily, a strict competitive strategy is defined conversely: The subject with the highest performance receives only 1 point and the participant with the worst performance gets the most (i.e. 11) points, etc.

When a subject basically wants to send high-performance players to the elite-team and therefore gives most points to them but modifies this strategy a bit in favor of team fellows, team identity should, as a matter of course, be measured in relation to the objective ranking. On the other hand, when a subject decides in favor of a competitive strategy, he can show team identity by changing the team fellows' rank relatively to this strategy. If all the subjects' rankings were e.g. compared to the objective rankings and team identity was determined on that basis, an underlying strategy would be attributed which does not correctly describe the behavior of certain subjects and thus would yield a wrong conclusion.

To use the appropriate reference strategy, I have to define an exact measure by which I can decide whether a subject basically uses the objective or the competitive strategy. Thus, for

each subject, I calculate a measure m_1 to determine the individual reference strategy. In the following I define m_1 formally.[16]

With a subject's *score*, I denote a subject's number of correctly solved products at stage 1. With score s_i at position i ($i=1,...,11$) of a subject's ranking[17] the number K of all elements s_i greater or equal than the preceding element s_{i-1} can be calculated by

$$k_i = \begin{cases} 1 \text{ for } (s_i - s_{i-1}) \geq 0 \\ 0 \text{ otherwise} \end{cases} \quad \text{, with } i = 2,\ldots,11,\ \sum_{i=2}^{11} k_i = K$$

Conversely, for the L elements in decreasing order

$$l_i = \begin{cases} 1 \text{ for } (s_i - s_{i-1}) < 0 \\ 0 \text{ otherwise} \end{cases} \quad \text{, with } i = 2,\ldots,11,\ \sum_{i=2}^{11} l_i = L$$

It follows $k_i + l_i = 1$ for every i. Thus, it holds $K + L = 10$. The measure m_1 is defined as $m_1 = K/10$. There is a mainly increasing order if $K > L$. Consequently this yields to $K > 10 - K$ and $K/10 > 0.5$. Besides, from $K \leq 10$ follows $K/10 \leq 1$. Hence, a measure is yielded for mainly increasing order of scores (which corresponds to predominantly competitive ranking) $0.5 < K/10 \leq 1$ or $0.5 < m_1 \leq 1$. Conversely, it can easily be shown for $K < L$ that $0 \leq K/10 < 0.5$ respectively $0 \leq m_1 < 0.5$, i.e. for mainly decreasing order of scores respectively predominantly objective ranking. For $K = L$, i.e. $K/10 = 0.5$, the order has no predominant direction. In this case the order is considered as objective.[18] Hence, for predominantly objective ranking it is $0 \leq m_1 \leq 0.5$.

To sum up: When a subject's ranking is predominantly arranged in an increasing order of performance, the value of m_1 takes on values between 0.5 and 1 ($0.5 < m_1 \leq 1$) and the competitive strategy is assigned. In the other case, where $0 \leq m_1 \leq 0.5$, the objective strategy is assigned.

---

[16] Please see appendix (section 2.5) for a calculation example.
[17] With a subject's ranking I refer to his distribution of points toward the other subjects in his voting list. The subject who is preferred most in a subject's ranking is at position i=1 (i.e. gets 11 points).
[18] A tie occurred only twice in the whole data set.

It may be assumed that the choice of an underlying strategy depends on the given set of team identity factors. As the ICT is the treatment where no team identity factors are given, this may yield to more competitive voting behaviors. Contrarily, the objective strategy may occur more often in ANT and GRT. An overview of the applied reference strategies per treatment can be found in Table 2.2.

Table 2.2: Relative frequencies of strategies dependent on treatment (in %)

Treatment	ICT	ANT	GRT
Objective	79.17	69.44	83.33
Competitive	20.83	30.56	16.67

As can be seen from Table 2.2, the objective strategy is applied more often than the competitive one in all three treatments. Moreover, the objective strategy is applied most in GRT and least in ANT. But when testing whether the frequencies of the chosen strategies (objective or competitive) differ between my treatments, no significant difference can be found.[19]

This means that the choice of strategy does in most cases not depend on the given team identity factors. Therefore, the application of either strategy does not depend on whether no team identity is provoked, economic team identity is introduced, or a communication phase is established in addition to economic identity to foster team identity.

Because the objective strategy is used most often, most subjects seem to have the election of a high-performance elite-team in mind when voting. But as they also would benefit from being elected to the elite-team, their voting may be based on another kind of reasoning: If subjects want to enhance their own chance to get into the elite-team on an indirect way, their behavior depends on their belief of the other subjects' strategy choices and on subjects' assessment of how many tasks they themselves have solved correctly at stage 1.[20] If a subject believes that he has performed rather bad and believes that many other subjects

[19] Results from two-sided Chi square tests: Pearson chi^2=0.6960, p=0.404 when comparing ICT and ANT, Pearson chi^2=1.9251, p=0.165 when comparing GRT and ANT, Pearson chi^2= 0.1670, p=0.683 when comparing ICT and GRT.

[20] As already mentioned, no subject knows his own performance. But subjects are asked in the questionnaire how many tasks they think to have correctly solved at the first stage.

use the competitive strategy, he may vote objectively to increase his chance to join the elite-team.

Continuing this train of thoughts it can analogously be argued that if a subject believes to have performed really well and believes that many other subjects vote objectively, he may further enhance his own chance of getting into the elite-team by voting competitively. With my dataset, I can test by the calculation of a point bi-serial correlation coefficient for ICT, ANT, and GRT whether there is a correlation between the assessed performance and the choice of strategy, but no sufficient correlation can be found.[21]

2.3.2 Quantitative Measurement of Team Identity

Method

Based on the subjects' individual lists of preferences as stated in the ranking, I can investigate whether their voting is biased by team identity, given a certain reference strategy. As a prerequisite, team identity has to be measured in a suitable way. When evaluating voting behavior, each subject has to be considered individually because each subject gets a different voting list, as subjects do not appear on their own list. I define positive (negative) team identity as follows: When a subject assigns his own team-mate to a better (worse) rank – given his reference strategy – positive (negative) team identity is shown.

At the beginning, I investigate quantitatively whether subjects show team identity or not and neglect the extent of team identity for the moment. For each subject I verify typical characteristics in view of team identity in his ranking. For a short characterization, I use the 3-tuple (positive team identity, negative team identity, no team identity) with a dichotomous variable (1 denotes "yes" and 0 "no") for each characteristic. For example, if a subject does not dislocate any team member, his personal feature is (0,0,1).[22] By these

[21] Subjects' statements about their self-assessment have to be treated with some caution because they are not paid for a correct self-assessment. In spite of that, their self-assessments are relatively accurate. On average, they underestimate themselves a bit: Average deviations from actual scores are: -1.58 (*SD* = 7.27) in ICT, -3.97 (*SD* = 11.56) in ANT, and -4.44 (*SD* = 7.81) in GRT.

[22] If a subject shows positive team identity to both team-mates his feature is (1,0,0) and if a subject shows negative team identity to both team-mates his feature is (0,1,0). There are subjects who do not treat both team-mates equally. If a subject ranks one team-mate higher than in his reference ranking and the other

features, I investigate the quantity of team identity bias occurring in the treatments and whether the detected frequencies differ across ANT and GRT.[23]

Working Hypothesis

Because of the findings from literature presented at the beginning of this chapter and since more team identity enhancing factors are given in GRT than in ANT, I state the following hypothesis concerning a comparison between the frequencies of positive and negative team identity in GRT and ANT:

Hypothesis 2.1: ***(i) In GRT more subjects show positive team identity than in ANT.***
(ii) In GRT less subjects show negative team identity than in ANT.

Frequencies of Positive and Negative Team Identity

Table 2.3 displays the percentages of subjects who show positive, negative, or no team identity in ANT and in GRT.[24]

Table 2.3: Relative frequencies of positive, negative, and no team identity (in %)

	Positive team identity		Negative team identity		No team identity	
Treatment	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
ANT	33.33	66.66	30.56	69.44	66.66	33.33
GRT	52.78	47.22	11.11	88.89	58.33	41.67

Interestingly, in both treatments positive and negative team identity can be detected. Considering ANT, 33.33% of all subjects participating in ANT show positive team identity to at least one team-mate and 30.56% show negative team identity to at least one team-mate. In GRT, 52.78% show positive team identity to at least one team-mate and only 11.11% show negative team identity to at least one team-mate.

To examine the first part (i) of my working hypothesis, I test "positive team identity" and "no positive team identity" across ANT and GRT. As a result, I find that the frequencies of

lower, the feature is (1,1,0). If one team-mate is not dislocated but the other is dislocated, his feature is (1,0,1) when favoring this team-mate, or (0,1,1) when discriminating this team-mate.

[23] This analysis is not undertaken for ICT since no teams are formed.

[24] In Table 2.3, the percentages of subjects who are written under "Yes" and "No" for a certain category for a certain treatment add to 100%. Because each subject can treat his team-mates differently, percentages across categories for a certain treatment do not necessarily add to 100%.

these two characteristics differ significantly between the treatments (Chi square test, Pearson chi^2=2.7758, p=0.048, one-sided), with positive team identity occurring more often in GRT. Thus, exploring whether subjects show positive team identity or not and neglecting the extent of this identity indicates that significantly more subjects in GRT show positive team identity than in ANT. Consequently, my working hypothesis 2.1 (i) cannot be rejected. The data show that giving subjects the possibility to talk to each other face-to-face, so that they get to know each other, and to solve some tasks together, with a high-minded aim in view, result in more positive deviations from the reference ranking concerning the own team-mates than without these additional social group identity factors.

To examine the second part of my working hypothesis, I test the two categories "negative team identity" and "no negative team identity" between ANT and GRT. I find that the frequencies of these characteristics also differ significantly between the treatments (Fisher exact test, p=0.040, one-sided[25]), with negative team identity occurring more often in ANT. Thus, omitting the communication phase leads more subjects to discriminate their own team-mates by shifting them down in their ranking. Getting acquainted with one's team-mates and fulfilling a task together seem to work against this discrimination and neutralize negative team identity or even change it into more positive attitudes toward team members. Hence, my working hypothesis 2.1 (ii) can also not be rejected.

This is a striking result: In GRT, goodwill toward own team-mates is not only expressed by favoring them more often than in ANT (i.e. placing them higher in the individual ranking) but also by not putting them into a worse position compared to the reference ranking. Thus, in GRT team identity appears twice – in positive and in not negative actions undertaken by own team-mates.

In the next step I test if the number of subjects who show "no team identity" to at least one team-mate and the number who shows "any kind of team identity" differ between ANT and GRT. I find that the quantities do not differ significantly across treatments (Chi square test, Pearson chi^2=3.1280, p=0.209, two-sided).

[25] A Fisher exact test is applied here because the required minimum number of observations per category for the conduct of a Chi square test (which is 5) is not satisfied.

The so far accomplished statistical tests are not totally independent of each other.[26] To overcome this methodological issue I additionally investigate combinations of the three investigated characteristics. In view of my research topic, the *complete* characterization tuples (1,0,0) and (0,1,0) - subjects who show *only* positive or *only* negative identity to both team-mates - are of particular interest. So I test whether the frequencies of the complete characterization tuples [(1,0,0) or not (1,0,0)] and [(0,1,0) or not (0,1,0)] substantially differ across ANT and GRT. In addition to that, I test the frequencies of the complete tuples [(0,0,1) or not (0,0,1)] in ANT and GRT. Table 2.4 gives an overview of the distribution of these specific cases.

Table 2.4: Relative frequencies of only positive, only negative, or only no team identity (in %)

	Only positive team identity (1,0,0)		**Only negative team identity (0,1,0)**		**Only no team identity (0,0,1)**	
Treatment	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
ANT	8.33	91.67	13.89	86.11	47.22	52.78
GRT	33.33	66.66	0	100	44.44	55.56

As can be seen from Table 2.4, 8.33% of all subjects participating in ANT show "only positive team identity", i.e. favor both team-mates.[27] A higher percentage of subjects (13.89%) shows "only negative team identity", i.e. discriminate both team mates. This is quite different when considering GRT: 33.33% of all subjects participating in GRT favor both team-mates while there is no subject in this treatment who discriminates both team-mates.

When testing if the frequencies of subjects who show "only positive team identity" or "not only positive team identity" differ between ANT and GRT, I find a significant difference

[26] The specifications 0 or 1 in the categories "positive team identity", "negative team identity", and "no team identity" are not totally independent of each other: If a subject neither shows positive nor negative team identity to any team-mate, it necessarily follows that he shows no team identity to any team-mate. This is why also the tests of the categories are not totally independent of each other: As I have shown, the test concerning "no team identity" yields no significant difference between ANT and GRT and the test concerning "positive team identity" shows that there is more positive team identity in GRT than in ANT. Then it is logical that there are more subjects in ANT showing negative team identity than in GRT. But, in any case, the results of the specific tests are meaningful, because within these tests the data are independent.

[27] Also in Table 2.4, the percentages of subjects who are stated under "Yes" or "No" of a certain category for a certain treatment add to 100%. Again, across these categories, the percentages do not necessarily add up to 100% for a certain treatment.

(Fisher exact test, p=0.009, one-sided). There are significantly more subjects in GRT than in ANT who show positive team identity to both team-mates. When testing if the frequencies of subjects who show "only negative team identity" and of subjects who do not show "only negative team identity" differ between ANT and GRT, there is also a significant difference found (Fisher exact test, p=0.027, one-sided). Here, significantly more subjects discriminate both team members in ANT than do subjects in GRT. Considering "only no team identity", I find no significant difference among the two treatments (Chi square test, Pearson chi^2=0.0559, p=0.813, two-sided).[28]

Finally, I investigate if there is a correlation between a subject's underlying strategy and the occurrence of positive or negative or no team identity. It seems plausible that someone who uses the competitive strategy is mainly focused on joining the elite-team. Therefore, he may regard his team-mates as (closer) competitors and is not willing to give them a higher chance to join the elite-team by favoring them because of envy. Perhaps he even wants to discriminate them in order to reduce their chances to win the elite bonus – independent of their performance at stage 1.

Concerning ANT, I find a negative correlation between using the competitive strategy and showing positive team identity (phi=-0.3, phi-correlation coefficient for two nominal and dichotomous variables). This means that a subject who uses the competitive strategy, less frequently shows positive team identity than when using the objective strategy. There is no correlation between the chosen strategy and negative team identity in ANT. However, I find a positive correlation between using the competitive strategy and showing no team identity (phi=0.4). This is in line with competitive behavior that includes neither positive nor negative team identity. In GRT, there is a positive correlation between the competitive strategy and negative team identity (phi=0.3). When a subject uses the competitive strategy, he more frequently shows negative team identity than when using the objective strategy. Again this finding can be interpreted in such a way that a subject who is competitive preferably does not want his team-mates to join the elite-team and therefore discriminates them. For the other cases no correlation can be found.

[28] These three tests of tuples are totally independent of each other because a subject who is not characterized by tuple (1,0,0) is not necessarily characterized by tuple (0,1,0) or by (0,0,1).

The competitive strategy seems to counteract favoring of own team-mates in both treatments, but through different ways: While it leads to less occurrences of positive team identity in ANT it provokes more occurrences of negative team identity in GRT.

In order to get a better insight into the driving forces behind team identity I conduct an order probit regression analysis. Table 2.5 displays the estimated coefficients and their effect on a subject's probability to show positive, negative, or no team identity to a certain team-mate.

In my regression I model "team identity" toward a certain team-mate as the dependent variable which can either be negative (-1) or positive (1), or zero (0) (for the case that neither type of team identity occurs). According to my research agenda, I investigate the impact of the independent variable TREATMENT (ANT with anonymous interaction of team-mates and GRT with face-to-face interaction). To find out whether a subject's team identity is influenced by the team-mates' absolute and relative performances I include the independent variables ASSOCIATED SUBJECT'S SCORE (a team-mate's score at stage 1) and ASSOCIATED SUBJECT'S RANK (a team-mate's rank in the reference ranking) in my regression. A high absolute and relative performance of a team-mate are expected to positively influence a favoring of him. Moreover I include the variable REFERENCE STRATEGY with which I now examine the general effect of the chosen reference strategy on team identity. Furthermore, my regression contains the variables TEAM BONUS EXPECTED, OWN ACTUAL SCORE, SELF-ASSESSMENT OF OWN ACTUAL SCORE, and SEX.[29] Whether a subject expects his team to win the team bonus or not may also guide his behavior toward team-mates, moderated by his own actual performance at stage 1 (which is unknown to the subject) and, more important, the assessment of his actual performance at stage 1. The expectation of the team bonus – dependent on team members' scores –, together with a positive self-assessment, might lead to a more positive, generous, and favoring attitude toward team-mates. When the team bonus is not expected this may lead to negative feelings (e.g. anger, disappointment, frustration) inducing implicit negative reciprocity toward team-mates. A voter might blame his team members for not winning the team bonus resulting in punishment by discriminatory voting behavior. However, it is also plausible to expect poor players to form a coalition of "losers".

[29] Of course I observe a subject's actual score in the multiplication task. Furthermore, I ask subjects whether they expect to have won the team bonus, how they assess their performance and whether they are male or female in the questionnaire at the end of my experiment.

Table 2.5: Ordered probit estimates for dependent variable TEAM IDENTITY

Independent Variable	Coefficient (Standard error)
TREATMENT	.63558 ***
	(.21022)
ASSOCIATED SUBJECT'S SCORE	.01466 *
	(.00790)
ASSOCIATED SUBJECT'S RANK	.06407 *
	(.03828)
REFERENCE STRATEGY	.27737
	(.24168)
TEAM BONUS EXPECTED	.20902
	(.21259)
OWN ACTUAL SCORE	.00970
	(.01111)
SELF-ASSESSMENT OF OWN ACTUAL SCORE	-.00678
	(.01103)
SEX	.17260
	(.24418)
Cut1	1.13856
	(0.89671)
Cut2	2.83161
	(0.91487)
N	140
(Pseudo) R^2	0.0761
LR-χ^2	20.69
Prob > chi2	0.0080

Asterisks indicate variables as being significant at 1%*, 5%**, and 10%*.**

Table 2.5 shows that TREATMENT highly significantly influences subjects' voting attitudes toward team-mates.[30] Thus, in GRT, positive team identity is significantly more likely to occur, and at the same time, negative or no team identity are less likely to be exhibited. As expected, a higher ASSOCIATED SUBJECT'S SCORE and a better ASSOCIATED SUBJECT'S RANK do also have a positive influence on voting behavior that favors associated team-mates. The underlying REFERENCE STRATEGY – in general – does not seem to sufficiently guide subjects in their voting behavior. Also TEAM BONUS EXPECTED, OWN ACTUAL SCORE, SELF-ASSESSMENT OF OWN ACTUAL SCORE, and SEX do not significantly influence the probability of favoring or discriminating a team-mate or to show no team identity toward him.

[30] In the regression 1 [-1] represents positive [negative] team identity and 0 stands for no team identity. I apply 0 [1] for ANT [GRT], objective [competitive] reference strategy, no expectation of group bonus [expectation of group bonus], female [male] as dummy variables.

2.3.3 The Extent of Team Identity

To test the extent of team identity, I develop two measures, m_2 and m_3. As already explained, I determine measure m_1 for each subject first in order to find out whether I have to use the objective or the competitive strategy as reference for team identity measurement.

Measure m_2

After the determination of the reference strategy, measure m_2 is applied. It is used for investigating *where* members of the own team are placed in the ranking. The more positions they are shifted, the higher is the degree of team identity. Measure m_2 is constructed by the actual number of shifted positions - in relation to the reference sequence - and the maximum number of ranks which are possible to leap over (in the same direction).[31] When a team fellow is shifted from position i to position J, i.e. from rank r_i to rank r_J, *relative to the reference ranking*, the degree of shifting is measured by the difference in ranks normalized by division through the maximum shift possible in the same direction. Formally this yields for either team fellow:

$$\hat{m}_2 = \text{sign} \cdot (r_i - r_J)/(r_i - R), \text{ with } r_i = i,\ i = 1,\ldots,11 \text{ [32] and}$$

$$\text{sign} = \begin{cases} +1 \text{ for upward} - \text{shifting} \\ -1 \text{ for downward} - \text{shifting} \end{cases}$$

$$R = \begin{cases} r_1 \text{ for upward} - \text{shifting} \\ r_{11} \text{ for downward} - \text{shifting} \end{cases}$$

I distinguish whether an individual is shifted to a higher or to a lower position. In the first case, i.e. positive team identity, the measure is taken as positive. In the latter case, i.e. negative team identity, it is set negative. It is also calculated by the number of positions the considered person is shifted – but downwards – divided by the possible maximum number the person can be shifted – but of course also counted downwards – and affixed with a negative sign. If shifting is impossible because $r_i = R$ already holds, then $\hat{m}_2 = 0/0$ is formally yielded, which is set to 0 by definition. Hence $-1 \leq \hat{m}_2 \leq 1$.

[31] Please see appendix (section 2.5) for a calculation example.
[32] Again, $i=1$ refers to that subject who gets assigned 11 points in a subject's ranking.

Finally, team identity of the voting subject is measured by the average degree of shifting he carried out for his two team fellows: $m_2 = 0.5 \cdot (\hat{m}_2(\text{fellow1}) + \hat{m}_2(\text{fellow2}))$.[33] Because it is impossible that $\hat{m}_2(\text{fellow1}) = \hat{m}_2(\text{fellow2}) = -1$ or $+1$[34], it is $-1 < m_2 < 1$. In the case that one fellow cannot be shifted in one direction, because $r_i = R$ already holds and it can therefore not be decided whether the subject is team-minded or not, an average is not computed and only the one team-mate that can be shifted is taken into consideration.

Measure m_3

In order to measure not only the subjects' team identity by the number of shifted ranks, a further measure m_3 is developed, which additionally takes the amount of overleaped scores into account.[35] Hereby the measurement of team identity may be enhanced because there is an essential difference whether an in-group member who is ranked up, for instance to the preceding position, replaces a subject who solved 20 multiplications more or only 1. Indeed, one of the underlying questions for a subject confronted with the voting task is: How much is it worth for me to get my team fellow favored?

Formally, m_3 is constructed quite similar to m_2. I take the difference between the score of the shifted subject and the score of the subject whose rank is replaced. Then this quantity of scores is normalized by division through the maximum difference of scores that is feasible by a shifting into the same direction.

I denote for a specific subject the sequence of the other 11 subjects' achieved scores *in the reference ranking* as $\{t_i\}$ with $i=1,\ldots, 11$, $0 \leq t_i \leq 100$, $t_i \geq t_{i+1}$ for the objective ranking and $t_i \leq t_{i+1}$ for the competitive ranking.[36] When a team fellow is shifted from rank i to rank J, I define for either team fellow

$$\hat{m}_3 = \text{sign} \cdot (t_i - t_J)/(t_i - T), \text{ with}$$

[33] If a subject favors one team-mate and discriminates the other, calculating the average results in a neutralization of the two effects. At the most extreme case, m_2 can add up to zero. Thus, a m_2 of zero does not necessarily mean that a subject does not show team identity. However, by taking the average I want to account for the fact that a subject might intentionally discriminate one team-mate and might offset this by favoring the other.

[34] Since it would mean both team-mates are shifted to the first respectively last rank.

[35] Please see appendix (section 2.5) for a calculation example.

[36] Again, $i=1$ refers to the subject who is most preferred in a subject's ranking (i.e. gets 11 points).

$$\text{sign}=\begin{cases}+1\ for\ upward-shifting\\ -1\ for\ downward-shifting\end{cases}$$

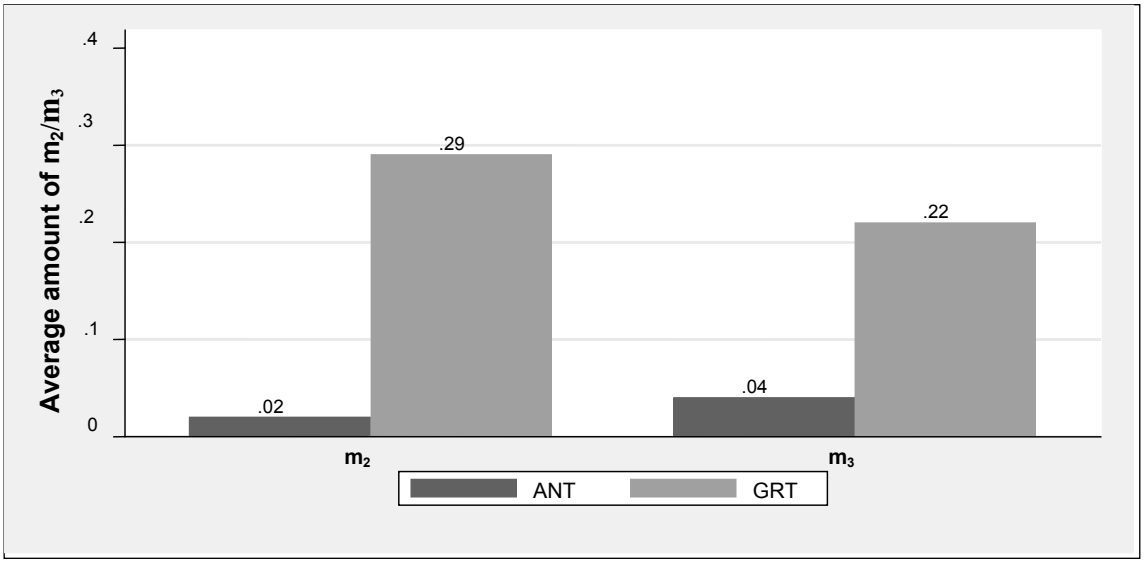
$$T=\begin{cases}t_1\ for\ upward-shifting\\ t_{11}\ for\ downward-shifting\end{cases}$$

and $m_3 = 0.5 \cdot (\hat{m}_3(\text{fellow1}) + \hat{m}_3(\text{fellow2}))$.[37] It is $-1 < m_3 < 1$. Again, if formally $\hat{m}_3 = 0/0$, it is set to 0 by definition. The average is only computed when team identity is feasible in both directions for both team-mates. Otherwise only $\hat{m}_3$ is used. If the considered subject is shifted to a lower rank, the last rank is taken as the point of reference, i.e. in case of objective ranking the one with the lowest score and in case of competitive ranking the one with the highest score. When shifting a subject to a lower position, a negative number is assigned, so that measure $\hat{m}_3$ for an individual can vary in the range $-1 \leq \hat{m}_3 \leq 1$.

Values of Measures m_2 and m_3

In the first part of this subsection I consider the *averages* of m_2 and m_3 which can be found in Figure 2.1. As can be seen, team identity is, on average, shown in both treatments. The

Figure 2.1: Average amounts for measures m_2 and m_3 dependent on treatment



[37] Again, I am aware that conducting the average of team identity towards the two team-mates can cancel each other out if one team-mate is favored and the other is discriminated (see footnote 33 for a brief discussion of this issue).

averages of m_2 and m_3 are both higher in GRT (0.29 and 0.22) compared to ANT (0.02 and 0.04). Thus, in GRT, the extent of positive team identity represents on average 29% of the maximal feasible upward rank shift and 22% of the maximal feasible upward score shift. Contrarily, in ANT, only 2% (4%) of the disposable favoring scale is applied. Moreover, m_2 is higher (smaller) than m_3 in GRT (ANT). Thus, accounting for the actual scores and not only for the ranks leads on average to a comparatively smaller extent of team identity in GRT but to a higher extent of team identity in ANT.[38]

I can now test whether the average extent of team identity differs between ANT and GRT. Starting with m_2, a significant difference is found (Mann-Whitney U test, p=0.004, two-sided), which shows that the values of m_2 are substantially higher in GRT. A similar result can be found for m_3: When testing the values of m_3 in both treatments against each other, a significant difference is found (Mann-Whitney U test, p=0.009, two-sided). Again, m_3 is higher in GRT on average. These results show that the extent of team identity is higher in GRT than in ANT on average.[39]

As can be seen from Figure 2.2 and Table 2.6, m_2 and m_3 are zero for most subjects in both treatments. Thus, according to these two measures, many subjects do neither favor nor discriminate their team-mates.[40] In addition to that, in both treatments there are subjects who show positive values of m_2 and m_3 (positive team identity) and subjects who show negative values of m_2 and m_3 (negative team identity). As can be seen, the relative frequencies of positive m_2 and m_3 are higher in GRT than in ANT. Moreover, negative amounts of both measures are more often found in ANT than in GRT. This confirms the results of the last subsection where it is shown that positive (negative) team identity is significantly more often found in GRT (ANT) than in ANT (GRT).

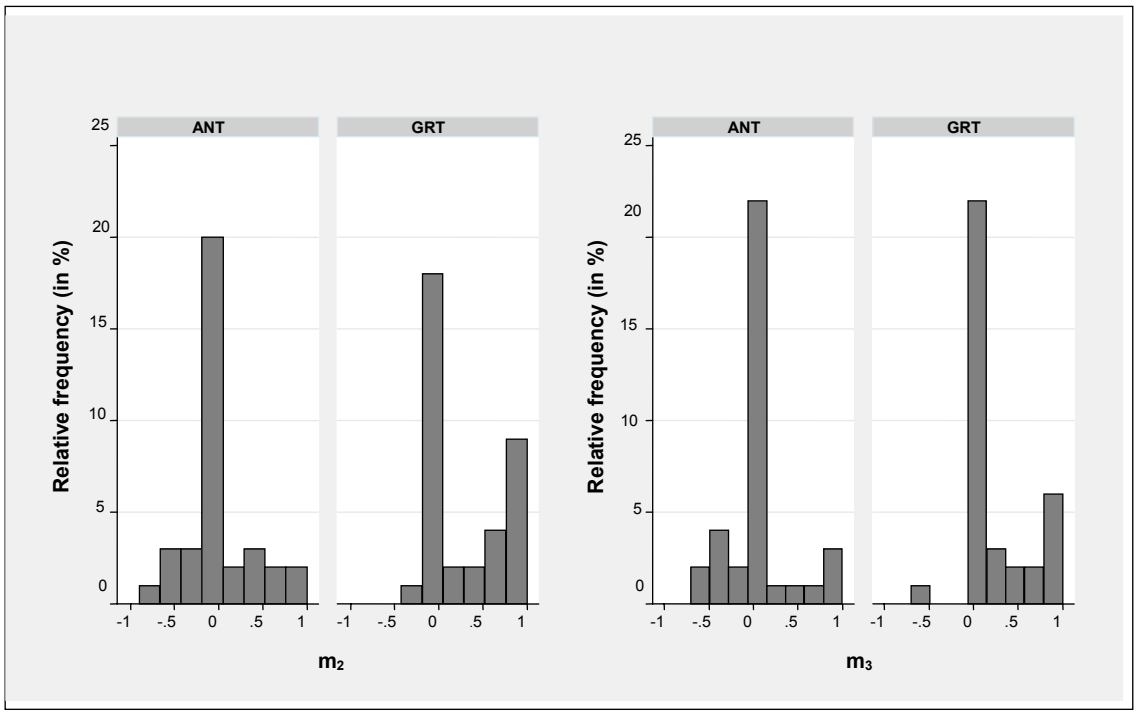
---

[38] This is due to the fact that in GRT team-mates whose score equals a non team member's score are often favored relative to this non team-mate.

[39] From a statistical point of view, it might be a little critical to test the values of m_2 and m_3 because each subject gets different information from the voting list (because a subject himself is not included). Therefore, the team identity of a subject depends on slightly different information. Consequently, I have to be careful with interpreting the results when testing the values of m_2 and of m_3. However, both measures are normalized and thereby comparable across subjects and deliver useful additional information.

[40] Basically, a m_2 or m_3 which is equal to zero can also result if $\hat{m}_2$ (fellow1) =- $\hat{m}_2$ (fellow2) and accordingly for $\hat{m}_3$. But this case occurs only once in the whole data set.

Figure 2.2: Distributions of m_2 and m_3 in both treatments



Moreover, as can also be seen in Table 2.6, the average extent of positive team identity is higher in GRT than in ANT. Thus, not only the frequency but also the extent of positive

Table 2.6: Relative frequencies (RF, in %) and averages (AV) of m_2 and m_3 when positive, negative, or zero values of m_2 and m_3 are considered

	Positive values of				Negative values of				Zero values of			
	m_2		m_3		m_2		m_3		m_2		m_3	
Treatment	**RF**	**AV**	**RF**	**AV**	**RF**	**AV**	**RF**	**AV**	**RF**	**AV**	**RF**	**AV**
ANT	25.0	0.49	22.2	0.52	27.78	-0.35	27.78	-0.29	52.78	0	50.0	0
GRT	47.2	0.67	38.9	0.61	5.56	-0.46	2.78	-0.49	52.78	0	41.67	0

team identity is higher with face-to-face interaction than without. Concerning the extent of negative team identity, the average is also higher in GRT than in ANT.[41] However, the average extent of negative team identity is calculated for only two (one) persons in GRT concerning m_2 (m_3).

[41] When testing the differences of the averages across treatments I do not find consistent significances.

2.3.4 Composition of the Elite-Team

An interesting question is whether the composition of the elite-team differs dependent on the treatment condition. All subjects profit from a high average performance of the elite-team. Consequently, the composition of the elite-team can be called efficient from an ex post point of view, when it is formed of the five subjects with the best performance at stage 3.

Taking a look at the distribution of performances at the last stage, the data show that subjects elect very good elite-teams, but not the efficient ones. To investigate the extent to which subjects in a certain session fail to compose the efficient elite-team, I consider the average of the actual elite-team's result in a session and compare it with the average result of those five persons with the highest scores at stage 3 in this session. To investigate inefficiencies on the treatment level, I calculate the averages per treatment which can be found in Table 2.7. The average difference per treatment is lowest in ICT (1.3), higher in GRT (2.3), and highest in ANT (4.9). Thus, subjects in ANT vote most inefficiently. But also in GRT and even in ICT subjects fail to elect the (complete) efficient elite-team. On average, the election of those subjects who should join the elite-team, because of their performance at stage 3, is quite similar across treatments.

Table 2.7: Comparison of actual and efficient elite-team per treatment (ex post point of view)

	ICT	ANT	GRT
Average result of elite-team	94.7	89.9	94.9
Average result of efficient elite-team	96	94.8	97.2
Difference	1.3	4.9	2.3
Average number of efficiently elected subjects	3.5	3.3	3.7

The previous analysis is carried out from an ex post point of view because the performances in the last stage are considered. But subjects in my experiment only know the performances of the first stage at the time of voting. In the light of this information subjects act efficiently when the five subjects who performed best in the first stage are elected to the elite-team.[42]

[42] Although I use the same multiplication tasks in stage 1 and stage 3 (in just another order), subjects' performances across stages 1 and 3 change a bit.

As can be seen in Table 2.8, a different picture occurs when regarding the composition of the elite-team from an ex ante point of view. GRT is the only treatment where the best five subjects of stage 1 are elected to the elite-team. This means, in this treatment subjects elect an efficient elite-team, given their information at this stage.

Table 2.8: Comparison of actual and efficient elite-team per treatment (ex ante point of view)

	ICT	ANT	GRT
Average result of subjects at stage 1 who are elected to the elite-team	90.6	85.1	91.1
Average result of the best five subjects at stage 1	91.9	87.9	91.1
Difference	1.3	2.8	0
Average number of efficiently elected subjects	4	4	5

To investigate how efficiently subjects vote in my treatments from an ex ante point of view, I first calculate for each session the difference between the average performances at stage 1 of those subjects who are elected to the elite-team and of those who are the best five at stage 1 and, second, the average per treatment. I find that the composition of the elite-team is efficient from an ex ante point of view in GRT. The average difference is 1.3 in ICT and 2.8 in ANT.

It seems that the election of the elite-team is "distorted" because – besides team identity – several subjects use the competitive ranking. In GRT, in contrast to ANT, the influence of competitive ranking, of positive, and of negative team identity seem to cancel each other out resulting in an efficient outcome from an ex ante point of view. Contrarily, the elite-teams in ANT are worst – both from an ex ante and an ex post point of view.

2.4 Summary and Discussion

In this chapter I systematically investigate whether there exists an in-group/out-group bias that leads to a favoring of own team members as candidates in promotion relative to other teams and their members.

In contrast to psychological approaches, monetary incentives for voting choices are implemented. Using an experimental approach, I am able to isolate and analyze distortion effects on performance evaluation of in-group and out-group members.

Moreover, objective performance criteria are defined and the extent of the in-group bias exactly measured.

My results show that face-to-face interaction and communication and a joint team task can lead more subjects to favor own team-mates than in an anonymous team environment. In-group members are more willing to deliver a voluntary voting support to increase team-mates' probability of promotion. This finding clearly confirms the first part of my hypothesis. Moreover, not only the frequency of positive team identity is higher but also the average extent. There are not only more subjects who favor own team-mates, but own team-mates are favored to a higher degree according to objectively measurable performance.

My second main finding, which also supports the second part of my working hypothesis, is that mere anonymous team interaction can lead to a substantial disadvantage for own team members. This very interesting result extends and even contradicts previous findings of psychologists. When giving subjects the possibility to discriminate their team-mates in voting for promotion, they make use of negative shifting in their promotion ranking. Consequently the introduction of teams (groups) does not only lead to a favoring of own team-mates, as extensively investigated by psychologists, but also negative team identity can be detected.

How can the negative bias in performance evaluation and voting for promotion be explained? One reasonable explanation is that own team members represent a less abstract reference group for the decision making team member than the members of the other teams. In this case the initial (positive) fact, that a team is created, results in a negative outcome for the other team members. This can be due to the fact that the other team members are closer to the decision maker and therefore more exposed to competition. A second assertion that points into the same direction and that also gets support by participants' written comments is that under anonymous conditions a subject attempts to be the "winner" of the team, who is elected and promoted to the elite-team. To reduce their team-mates' chances they are discriminated by assigning less votes to them. In the case when a subject thinks that his chances are very small for getting into the elite-team, he wants to prevent other team members to join it because of envy. Instead, the advantage is left to a third person, who is more distant.

A third contribution of this chapter is that I show that a totally anonymous promotion process does not lead to an efficient composition of the elite-team because several subjects behave competitively and disadvantage own team members in the voting stage. Contrarily, in a situation where subjects can easily communicate and have joint group tasks, the negative influence of competitive behavior seems to be cancelled out by positive team identity resulting in an efficient elite-team from an ex ante point of view. From this perspective, GRT is the treatment which would be favored by the management of an organization. Hence, company leaders should care for intensive face-to-face communication and identification creation among team members. The sole creation of virtual or rapidly changing teams, where members have no constant contact, can lead to inferior outcome when subjects shall be determined for promotion. Creating a team context with individuals pursuing their own interest may even negatively affect the efficiency of selected teams. To counteract this, employers should foster social identity between team-mates and not exclusively focus on economic goals.

Another finding is that the mere introduction of teams, a joint team task, an implemented team bonus, and face-to-face communication do not substantially influence subjects' choice of the reference strategy for voting. This may be due to the fact that a certain distribution of "competition-types" and associated beliefs inherently exists, which remains unaffected by experimental treatment conditions. However, on average there is a slightly decreasing tendency that these types are detected in GRT.

Future work can focus on more diverse effort tasks, which are more sensitive to environmental and organizational changes, to investigate how performance and promotion are connected. The real effort task in my experiment is one-dimensional and just requires simple mathematical skills. Other characteristics or abilities (like e.g. social competence) were not needed to solve the given tasks. By implementing different types of tasks that require complementary skills it can be investigated how the change of multi-level performances interacts with occurrences and extends of positive and negative team identity.

Furthermore, the composition of teams can be varied. For instance, by the introduction of culturally mixed teams it could be studied how these factors influence team identity and promotion processes. Also interesting is the influence of crossed categorization on the in-

group/out-group bias: Consider for example an extension of my experiment, where each team consists of two Germans and one Chinese. Would the Chinese then favor his own team members or the Chinese students of the other teams?

Finally, another interesting aspect is the formation of social networks across borders by intentionally fostering or disadvantaging colleagues. This fact seems to be important nowadays with working places being frequently changed and abandoned and former team-mates being spread all over the world. Accounting for these aspects will lead to a better understanding of team interaction – within and between heterogeneously composed teams – in globalized business environments.

2.5 Appendix

A calculation example

In the first session of GRT, subject *c* with team-mates' scores 82 resp. 71 states the following individual ranking which is compared to the objective ranking:

individual ranking	objective ranking
82	99
99	98
98	82
78	78
71	75
75	75
75	71
69	69
66	66
52	52
47	47

The objective ranking is taken as reference point for subject *c* because the performances in his ranking are predominantly decreasing: $m_1 = 3/10 = 0.3$

Measure m_2:

$\hat{m}_2(82) = 2/2 = 1$

$\hat{m}_2(71) = 2/6 = 0.33$

$m_2 = (1+0.33)/2 = 0.67$

Measure m_3:

$\hat{m}_3(82) = 17/17 = 1$

$\hat{m}_3(71) = 4/28 = 0.14$

$m_3 = (1+0.14)/2 = 0.57$

Instructions of GRT in German:

Informationen über das Experiment

Sie nehmen nun an einem wirtschaftswissenschaftlichen Entscheidungsexperiment teil. Bitte lesen Sie sich die Beschreibungen der einzelnen Experimentsstufen jeweils sorgfältig durch. Wenn Sie etwas nicht verstehen sollten, schauen Sie bitte noch einmal in die entsprechende Anleitung. Falls Sie dann noch Fragen haben, geben Sie uns bitte ein Handzeichen.

Am heutigen Entscheidungsexperiment nehmen 4 Gruppen mit jeweils 3 Entscheidern teil. Jede Gruppe ist in einem separaten Entscheidungsraum untergebracht. Die Mitglieder unterschiedlicher Gruppen werden im Gesamtverlauf des Experiments nicht miteinander in Kontakt treten. Die Informationen zwischen den einzelnen Gruppen werden durch die Experimentatoren ausgetauscht, welche jederzeit für Sie verfügbar und ansprechbar sind. Das Entscheidungsexperiment ist in mehrere Stufen unterteilt, welche Sie alle hier, in Ihrem Raum, bearbeiten. Nach Beendigung des Experiments erhalten Sie Ihre Auszahlung, ebenfalls in diesem Raum. Sie werden daher Ihren Raum für die Dauer des gesamten Experimentes nicht verlassen.

Bevor es losgeht…

Zuallererst möchten wir Sie bitten, sich gemeinsam einen Namen für Ihre Gruppe auszudenken. Dieser Name muss aus genau 3 Wörtern bestehen, welche jeweils mit den Anfangsbuchstaben der Vornamen der einzelnen Gruppenmitglieder beginnen. Der Gruppenname wird dann auf das vorhandene Kartonpapier geschrieben und draußen an der Tür des Gruppenraumes befestigt. Dies erleichtert uns die Orientierung und die Zuordnung während des Experimentes. Für die Namensfindung haben Sie von jetzt an 10 Minuten Zeit.

Experimentsbeginn

Zu Beginn des Experiments haben Sie nun die Möglichkeit, durch Ihre **Gruppenleistung** einen **Spendenbeitrag** zum **Erhalt des Bonner Beethovenhauses** zu erspielen. Der von Ihnen erspielte Betrag wird direkt im Anschluss an das heutige Experiment von der Experimentsleitung an den „**Verein Beethoven-Haus Bonn**“ überwiesen.

Im Folgenden erhalten Sie dazu nun **2 verschiedene Texte**. Ihre Aufgabe ist es, die vorliegenden Texte innerhalb eines Zeitraums von **5 Minuten** zu lesen. Nach Ablauf der Zeit werden die Textblätter durch den Experimentator eingesammelt. Im Anschluss daran haben Sie **3 Minuten** Zeit, einen **Lückentext** mit Aussagen zu den gelesenen Texten zu bearbeiten. Für jede richtig beantwortete Frage bekommen Sie einen Betrag von **0,50 €** auf Ihrem **Gruppen-Spendenkonto** gutgeschrieben. Nach Beendigung dieser Aufgabe wird Ihnen mitgeteilt werden, wie hoch der von Ihnen erspielte Spendenbetrag - inklusive einer Aufstockung durch die Universität Bonn, um den Betrag abzurunden - ist.

Einleitung

Von nun an werden Sie Ihren persönlichen Code verwenden. Sie bekommen ein Startgeld in Höhe von **3 EURO** auf Ihrem persönlichen Experimentskonto gutgeschrieben. Im Verlauf des Experiments können Sie weitere Geldbeträge hinzuverdienen. Die Höhe Ihres Zugewinns hängt von Ihren Aktionen und Entscheidungen und von den Aktionen und Entscheidungen der anderen Teilnehmer ab.

Des Weiteren können Sie im Verlauf des Experiments verschiedene Bonuszahlungen erzielen. Ob Sie diese erhalten, hängt von Ihren Aktionen und Entscheidungen und von denen der anderen Teilnehmer ab.

Während der folgenden Experimentsschritte interagieren Sie mit Entscheidern innerhalb Ihrer Gruppe und mit Personen in anderen Gruppen. Sie erfahren zu keinem Zeitpunkt die wirklichen Namen der Entscheider in den anderen Gruppen und können deren Codes nicht einer bestimmten Person zuordnen. Genauso erfahren diese zu keinem Zeitpunkt Ihre Identität. Ihren Code können weder die Personen der anderen Gruppen noch die Personen Ihrer eigenen Gruppe mit Ihnen in Verbindung bringen. Auch Sie können die Codes Ihrer Gruppenmitglieder nicht einzelnen Personen zuordnen. Für die gesamte Dauer des Experiments ist es von nun an sehr wichtig, dass Sie nicht mehr mit den anderen Mitgliedern Ihrer Gruppe, hier in Ihrem Raum, sprechen.

Aufgabenbeschreibung für Stufe 1

Im Folgenden erhält jede Person in jeder Gruppe eine identische Liste mit einfachen Rechenaufgaben. Ihre Aufgabe ist es, davon so viele wie möglich richtig zu lösen. Zur Bearbeitung dieser Rechenaufgaben haben Sie **15 Minuten** Zeit. Für jede richtig gelöste Aufgabe werden Ihnen **3 Cent** auf Ihrem persönlichen Experimentskonto gutgeschrieben. Die Bearbeitungsreihenfolge der einzelnen Rechenaufgaben können Sie beliebig wählen. Zur Lösung der Aufgaben sind keine weiteren Hilfsmittel (außer Papier und Stift) zugelassen.

Nach Ablauf der 15 Minuten wird Ihre **persönliche Leistung** ermittelt. Darüber hinaus werden die **Gesamtergebnisse** aller 4 Gruppen miteinander **verglichen**. Dies geschieht, indem vorher die Einzelergebnisse der Mitglieder einer Gruppe zu einem **Gruppenergebnis** aufaddiert werden. Allen Mitgliedern derjenigen Gruppe, die so das **beste Gruppenergebnis** erzielt, wird zusätzlich zu ihrem Einkommen aus der persönlichen Leistung ein fester **Gewinnerbonus** in Höhe von **5 EURO** gutgeschrieben. Ihr Gesamteinkommen dieser Stufe berechnet sich demnach aus:

Anzahl richtig gelöster Aufgaben • 3 Cent
[+ Bonus (nur im Gewinnfall!)]
= Ihre persönliche Gesamtauszahlung für Stufe 1

Ob Ihre Gruppe gewonnen hat und Sie damit zusätzlich den Gruppenbonus erhalten, erfahren Sie am Ende des Experiments. Sie bekommen jedoch vorab schon Informationen darüber, wie die Leistung Ihrer Gruppe einzuordnen ist.

Stufe 2 und Rollen-Bestimmung

In der sich nun anschließenden letzten Stufe des Experiments bekommen Sie erneut eine Liste mit einfachen Rechenaufgaben. Diese Rechenaufgaben sind **strukturgleich** mit denen der ersten Stufe. Zur Bearbeitung dieser Aufgaben haben Sie wieder **15 Minuten** Zeit. Für jede richtig gelöste Aufgabe bekommen Sie **3 Cent** auf Ihrem persönlichen Experimentskonto gutgeschrieben. Die Bearbeitungsreihenfolge der einzelnen Rechenaufgaben können Sie wieder beliebig wählen.

Im Gegensatz zur vergangenen Stufe gibt es hier jedoch nun **2** verschiedene **Spielerrollen**: Spielerrolle **A** und Spielerrolle **B**, wobei insgesamt **5 Spieler** in der Rolle **A** und **7 Spieler** in der Rolle **B** agieren werden. Spieler in der Rolle **A** erhalten zu Beginn dieser Stufe eine einmalige feste Zuzahlung in Höhe von **10 EURO**. Spieler in der Rolle **B** erhalten diese Zuzahlung nicht. **Alle** Spieler (unabhängig, ob A oder B) erhalten zusätzlich zu ihrem Einkommen aus ihrer persönlichen Leistung eine zusätzliche Zahlung **in Abhängigkeit der Leistung** der Spieler mit der Spielerrolle **A**. Diese zusätzliche Zahlung berechnet sich aus der **durchschnittlichen Anzahl richtig gelöster Aufgaben** aller Spieler mit der Rolle **A**. Unabhängig von ihrer Rolle auf dieser Stufe müssen alle Personen **dieselben** Aufgaben rechnen. Sie erfahren erst am Ende des Experiments, in welcher Rolle Sie auf dieser Stufe agiert haben. Wie die Rollen für die einzelnen Teilnehmer bestimmt werden, wird Ihnen in Kürze erläutert. Bitte beachten Sie: Auf dieser Stufe gibt es keinen Gruppenbonus!

Nach Ablauf der 15 Minuten wird Ihre **persönliche Leistung** ermittelt. Darüber hinaus wird die **durchschnittliche Leistung** der Spieler mit der Spielerrolle **A** berechnet. Ihr Gesamteinkommen für diese Stufe errechnet sich demnach aus:

Spielerrolle A	**Spielerrolle B**
+ 10 EURO	
+ Anzahl richtig gelöster Aufgaben • 3 Cent	Anzahl richtig gelöster Aufgaben • 3 Cent
+ Durchschnittliche Anzahl richtig gelöster Aufgaben der Spieler in Rolle **A** • 3 Cent	+ Durchschnittliche Anzahl richtig gelöster Aufgaben der Spieler in Rolle **A** • 3 Cent
= Ihre persönliche Gesamtauszahlung für Stufe 2	= Ihre persönliche Gesamtauszahlung für Stufe 2

Wie werden die Spielerrollen bestimmt?

Bevor Sie die Rechenaufgaben bearbeiten, werden die Spielerrollen **A** (=**5** Spieler) und **B** (=**7** Spieler) durch die Mitglieder aller Gruppen gewählt. Dies geschieht durch folgendes Verfahren:

Jedes Gruppenmitglied in jeder Gruppe erhält einen Zettel, auf dem alle restlichen Teilnehmer des Experiments mit der **Anzahl richtig gelöster Aufgaben** von **Stufe 1** aufgelistet sind. Um Anonymität zu wahren, erfahren Sie jeweils nur den persönlichen Code der anderen Person, und, ob sie zu Ihrer eigenen Gruppe gehört oder nicht.

Um die Spielerrollen aller Teilnehmer zu bestimmen, müssen Sie nun angeben, wen Sie **am liebsten** in der Rolle **A** agieren sehen möchten. Dem Code der Person, die Sie **am liebsten** in der Rolle **A** agieren sehen möchten, sollten Sie **11 Punkte** zuordnen, dem Code der Person, die Sie am **zweitliebsten** in der Rolle **A** agieren sehen möchten, sollten Sie **10 Punkte** zuordnen, usw. Dem Code der Person, die Sie **am wenigsten** in der Rolle **A** agieren sehen wollen, sollten Sie **1 Punkt** zuordnen. Jede Punktzahl von 11 bis 1 dürfen Sie dabei nur einmal vergeben!

Anhand der Punktlisten aller 12 Teilnehmer wird anschließend ermittelt, welche 5 Personen in der Rolle A und welche 7 Personen in der Rolle B agieren. Die **individuellen Punkte** werden dafür für alle Teilnehmer **aufaddiert**. Die **5** Personen mit den **meisten Stimmen** agieren dann in Rolle **A**. Alle restlichen Teilnehmer agieren in Rolle **B**. Können aufgrund eines Stimmengleichstandes keine 5 Personen eindeutig ermittelt werden, so wird zufällig entschieden, welchen der stimmengleichen Spieler die Spielerrolle A zugeordnet wird.

Chapter 3 [43]

Whom will you choose? Collaborator Selection and Selector's Self-Prediction

3.1 Introduction

Hiring of new employees is an important task for managers of corporate enterprises. The principal wants his managers to engage the most competent collaborators because they produce the highest output. Sometimes it can, however, be observed that managers not necessarily choose the most qualified collaborators.[44]

In the theoretical economic literature there are some studies that explain why it may not be rational for managers to choose the most qualified candidate. One possible explanation is presented by Friebel and Raith (2004). They argue that a manager makes every effort to avoid becoming replaced by a more qualified collaborator if his ability is directly compared to that of the collaborator. Therefore, the manager uses his authority to recruit strategically. Nevertheless, he takes into account the underlying disadvantage: Recruiting a less qualified collaborator increases the risk of a low team output, which negatively affects his further career.

Glazer and Segendorff (2005) assume that a manager is interested in credit claiming for successful work. They show that choosing the worst candidate can be an equilibrium strategy. Segendorff (2000) demonstrates that, under certain circumstances, it can be rational for a competent risk-averse manager to choose an incompetent employee in order to have a scapegoat in case of failure. According to Beniers (2005), a further reason for choosing a less competent collaborator can be loyalty: Very competent collaborators may have good side opportunities and so are usually less loyal.

[43] This chapter is based on Eberlein and Przemeck (2008b).
[44] This is for example the case when managers think that they do not deserve their position (see Rodriguez-Bailon et al. 2006).

Furthermore, the so-called anti-herding-literature deals with situations in which a person wants to signal its own (high) competence. In the context of this chapter, mainly Levy's article of 2004 has to be mentioned. In her model a decision-maker can be supported by a consultant. She shows that the most able decision-makers do not confer with a consultant in order to signal their excellent ability.

Besides the already mentioned aspects, there are numerous other reasons for a manager to hire a less competent collaborator. It can often be observed that a manager has to choose a collaborator for his department to undertake a project but has only little experience in the field of that project. Thus he does not know if he will perform well. If he believes that his achievement is not noticeable, he may tend to choose a less qualified collaborator.[45] In particular, the professional reputation he enjoys in his department may be lessened if he chooses a candidate who is more qualified than he himself. He even may be replaced. Therefore, the manager's self-prediction and the conditions influencing his self-prediction play a significant role in such a hiring process. It is plausible that a collaborator who is better than the manager will have better prospects for his career if the superiority in ability can be identified in the long run by a principal or the firm owners. Thus, choosing a less qualified collaborator may prevent the manager from decreasing authority and from loosing his position in the worst case. In such a case, the principal probably anticipates this inefficient behavior of the manager. One possibility to reduce this inefficiency is to offer the manager incentives to choose the best collaborator. This can be done by organizing a tournament between the departments of a firm in order to stimulate outstanding working results. Then an award is given to the manager's department for obtaining a high output or a good achievement compared to all other departments of the firm. In that case, a manager has to take into account all aspects that influence the chance of winning the tournament when choosing a collaborator: First, the abilities of the other departments' managers, second, which collaborators these will engage, and, third, his own expected ability and his chosen collaborator's ability.

To sum up, a trade-off can be observed in the process of employing a new collaborator: On the one hand, an exceptionally good collaborator increases the output of a manager's department and therefore increases the probability of winning a tournament prize. On the

[45] Choosing a less qualified collaborator means choosing one whose expected outcome is lower than the expected outcome of the manager.

other hand, a manager's superiority and even his position may then be at stake. If this is of higher importance for the manager, the manager should decide in favor of a less qualified candidate. However, a manager chooses a collaborator before the project is undertaken. Thus he has to predict his own ability. If a manager, who *overestimates* his ability, wants to employ a less able staff, he may unintendedly hire one who is more able than he himself. If a manager, who *underestimates* his ability, wants to employ a less able staff, it may happen that he engages a much too bad one. If he knew his real ability, he might choose a better collaborator (although still worse than he himself). Besides, it is also important how certain the manager is of his self-prediction. If he is not sure of the accuracy of his self-prediction and wants to hire a less qualified collaborator, he may choose a much less qualified one to ensure in any case that his prospective performance is better than the performance of his collaborator.

My experiment investigates managers' self-predictions of their subsequent performance and, based thereupon, their choice of a collaborator. To my knowledge, the influence of self-prediction on collaborator choice has not yet been investigated. Next to this aspect, my experiment examines if the selection behavior differs if a superior (who is assumed to be at a higher level of the hierarchy than the manager) chooses a collaborator for a manager, based upon his prediction of the manager's performance. By informing the superior of the manager's self-prediction, I can investigate how a superior evaluates this information. It is not clear whether a superior takes a manager's self-prediction for sure. According to psychological and economic literature on overconfidence, managers often state too high self-predictions.[46] If a superior anticipates this, he will predict the manager's performance to be lower than the manager's self-prediction. As a consequence, a superior may select another collaborator for the manager than the manager would do.

Regarding the results of my experiment, the data show that managers' self-predictions are not biased anymore after they are informed about the performance of a reference group. About 35% of the managers choose a collaborator who is worse than the self-prediction of their subsequent performance. Moreover, most managers do not rationally choose a collaborator given their beliefs. Concerning my second treatment, the data reveal that

[46] The phenomenon of overconfidence has extensively been investigated in psychology (e.g. Alpert and Raiffa 1969, Svenson 1981, and Weinstein 1980). In the field of economics there is also a growing literature which deals with overconfidence empirically (e.g. Camerer and Lovallo 1999, Fellner et al. 2004, Russo and Schoemaker 1992) and theoretically (e.g. De Long et al. 1991, Daniel et al. 1998 or Gervais and Odean 2001).

superiors adapt their predictions into the direction of the managers' self-predictions, although not completely. They adjust their predictions but keep their former prediction in mind and do not deviate too much from it. Interestingly, superiors think that their managers' self-predictions are too low if they are lower than the average performance of the reference group. I find that superiors' collaborator choices do not significantly differ from the managers' choices. This proves due to excellent information processing by both, managers and superiors, which on the whole leads to very similar predictions of managers' subsequent performance.

The question whether subjects know that others are overconfident has already been investigated by Cesarini et al. (2006) and Ludwig and Nafziger (2007). Cesarini et al. (2006) observe that subjects anticipate others' overconfidence. Ludwig and Nafziger (2007) also show this but only for an environment in which subjects are very familiar with the task. These studies considerably differ from my experiment because I investigate performance *predictions (before* the task has been carried out) and their evaluation by a third person. In Cesarini et al. (2006) subjects have to answer numerical questions. They have to state a lower and an upper limit for each question so that they think the stated interval actually contains the true answer to the question with a probability of 90% (confidence interval estimation). Moreover, subjects have to assess the other subjects' average accurateness, i.e., they have to assess in how many of the stated intervals the right value is on average contained. Because subjects assess that the number of these intervals are, on average, less than 90% of all stated intervals, Cesarini et al. (2006) conclude that subjects anticipate that others are overconfident. This approach is completely different to mine. In my experiment managers have to state how many multiplication tasks they will solve correctly. If, after being informed about the managers' self-predictions, superiors systematically predict a lower performance for the managers than the managers themselves, I say that the superiors think that managers are overconfident. In Ludwig and Nafziger (2007), subjects are informed about the average assessment of some other subjects and have to evaluate if these subjects are on average over- or underconfident. Then, they have to assess the average performance of these subjects. Contrarily, in my experiment superiors have to predict the managers' performance *before* they get to know their self-predictions as well as *afterwards*. Thus I can exactly investigate superiors' information processing and whether they expect that managers are overconfident or not. This is also a difference to Cesarini et al. (2006), where subjects do not get any information

about the subjects they have to assess. Furthermore, in contrast to the other experiments superiors get to know the self-prediction of a certain associated manager in my experiment. Ludwig and Nafziger (2007) also run a treatment in which subjects assess the bias of a certain subject, but they use a completely different approach. In their treatment, subjects have to evaluate for each possible individual assessment (strategy method) whether a subject, who hypothetically states one of these assessments, is over-, underconfident or unbiased. In my experiment, superiors know that their managers really made a certain self-prediction. Therefore they more intensively think about the accurateness of this particular value when they assess the manager's performance. Next to these aspects, I go one step further than Cesarini et al. (2006) and Ludwig and Nafziger (2007) in examining a decision (the collaborator choice), which is based upon the performance predictions.

The remainder of this chapter proceeds as follows. Section 3.2 describes the design of my baseline treatment and the design of the two main treatments. Section 3.3 presents the results of the baseline treatment, section 3.4 the results of the first main treatment, and section 3.5 the results of the second main treatment. In section 3.6 the results of both main treatments are compared, while the final section concludes.

3.2 Procedure, Experimental Design, and Hypotheses

The pen and paper experiment was conducted at the Bonn Experimental Economic Laboratory from September to December 2007. Subjects were recruited via the internet by using ORSEE software (Greiner, 2004) announcing the possibility to earn an amount of money dependent on their behavior. 20 subjects took part in the baseline treatment and 17 in each of the two main treatments. One session lasted 70 minutes on average and subjects earned approximately 12 €.

3.2.1 Design of the Baseline Treatment (BT)

The BT is conducted to find a typical reference group that will be used for comparison in the two main treatments. The BT is divided into two distinct stages: At stage 1, subjects carry out simple multiplications. At stage 2, they have to choose a hypothetical collaborator and to assess the collaborator choices of the other subjects in their group.

Stage 1:

Subjects get 10 minutes of time to solve simple multiplication tasks. They are only allowed to use pen and paper. For each correctly solved task, a subject obtains 4 Cents. Thereafter, the subjects are arranged into groups of four and are told how many multiplications they themselves solved correctly.[47] Moreover, they are informed about the number of correctly solved tasks of the other three persons in their group.

Stage 2:

At stage 2, every subject chooses a hypothetical collaborator with an integer result from the interval [0,110]. Thus subjects can choose between numerous collaborators who differ in the number of hypothetically correctly solved multiplications ranging between 0 and 110.[48] As it seems impossible to solve 110 tasks correctly in 10 minutes, a chosen hypothetical collaborator with the result of 110 is better than all the subjects in my experiment. This is important for my experiment, because I am mainly interested in the question whether subjects choose a better or a worse collaborator compared to their own performance.

Each subject and his chosen collaborator form a team. The selection of a collaborator affects the team result and the subject's payoff. The team result is defined as the sum of the subject's result in the task and that of the hypothetical collaborator. Before subjects select a collaborator, they are asked which (hypothetical) collaborator they expect to be chosen by the other three members of their group.[49] Afterwards they choose a collaborator themselves. Thus, each four-person-group forms four teams, each team consisting of a subject himself and his chosen collaborator.

A subject obtains 5 € + 1 Cent · CR if he chooses a collaborator who has a *lower* result than he himself (with CR denoting half of the selected collaborator's result). In contrast, a subject is paid 2.50 € + 1 Cent · CR if he chooses a collaborator with a result *higher* than or equal to his own result. In my experiment the difference (5 € – 2.50 €) represents the

[47] Before dividing subjects into groups, I first determine their results (i.e. the number of correctly solved tasks) and the distribution of the results. Then I assign the subjects to quartiles according to their results. Afterwards, I divide the subjects into five groups so that each four-person-group has exactly one member of each quartile. This procedure is used to make sure that each group is "typical".

[48] Note that each subject always gets his desired collaborator so that different subjects may choose collaborators with the same result.

[49] Subjects get 50 Cents for each correctly stated belief.

additional department's esteem for its manager, if the manager has higher abilities than his collaborator. The payoff structure ensures that rational subjects choose the best candidate (the one with the result of 110) in case of the selection of a more qualified one. If a subject prefers a less qualified collaborator, he should choose the candidate with just one result less than he himself. A tournament between the teams of a four-person-group is organized. If a subject's team result belongs to the two highest of his four-person-group, 7.50 € will additionally be paid to him. Thus the two rewards of 7.50 € represent the winner prizes of the tournament.[50]

3.2.2 Design of the Managers Treatment (MAT)

One of the five four-person-groups in BT is determined to constitute the reference group for the two main treatments. Each of the 17 "managers" attending MAT is compared to this reference group. Hence, each person in this treatment represents the fifth person in a group consisting of the four persons of the reference group and the considered subject himself. Using this method I obtain a comparatively great number of independent observations because the decisions of the, so to say, "fifth persons" are independent of each other.

MAT is divided into different stages: At stage 1, managers have to state predictions concerning their subsequent performance, at stage 2 they choose a collaborator, and at stage 3 they carry out multiplication tasks. Thus, contrarily to BT, managers perform the tasks at a later point of time.

Stage 1:

At stage 1, each manager is shown a typical multiplication task of stage 3 (namely "6 · 79"). Afterwards, they have to state a self-prediction (SP1) of how many tasks they will be able to solve correctly at stage 3. They know that they shall later work on the task for 10 minutes and that auxiliary means (except for pen and paper) are not permitted. They know that they will earn 4 Cents for each correctly solved multiplication task and that they will get a payoff for their self-prediction that is the higher the more precise their self-prediction is: A manager receives 1.80 € if his self-prediction exactly corresponds to his subsequent result, i.e. to the actual number of correctly solved tasks at stage 3. If the self-prediction

[50] The term 1 Cent · CR is chosen so that 5 € + 1 Cent · CR is always higher than 2.50 € + 1 Cent · CR. Therefore, choosing a better instead of a worse collaborator is only reasonable if expecting to win a winner prize.

deviates from the subsequent result, a manager is paid 1.80 € – (2 Cents · |actual number of correctly solved tasks – predicted number of correctly solved tasks|). Thus, the payoff is the lower, the more the actual performance at stage 3 differs from the predicted value.[51]

After stating their first self-prediction (SP1), managers are informed about the number of tasks the members of the reference group have solved correctly. If the managers thereupon want to change their former self-prediction (SP1), they can do that. In this case, their second self-prediction (SP2) is taken as the new basis for the determination of their payment as described above. If they do not change their self-prediction, payoffs are calculated on the basis of their first self-prediction. By a change of self-predictions, it can be investigated if and how managers renew their prediction of their subsequent performance as reaction to new information.

Stage 2:

At stage 2, managers are informed about the design of BT and the incentives that were given to the reference group's persons. The managers have to estimate which collaborator has been chosen by a certain person of the reference group. 50 Cents are paid for a correct belief. Moreover, they are asked how sure they are of their beliefs. Thereafter, managers select a collaborator from the interval [0, 110]. Each manager and his chosen collaborator compose a team. A manager obtains 5 € + 1 Cent · CR if he chooses a collaborator with a *lower* result than his own at the subsequent stage 3. On the contrary, a manager obtains 2.50 € + 1 Cent · CR if he chooses a collaborator with a result *higher* than or equal to his own result at stage 3. A manager's team result is compared to the four team results of the reference group. If his team belongs to the two best ones, he is additionally paid a bonus of 7.50 €.

Stage 3:

At stage 3, managers have to work on the multiplication tasks for 10 minutes. They get 4 Cents for each correct result. Thus I can determine whether a considered manager chooses

[51] Of course, I want to avoid that managers try to solve as many tasks as predicted before and finish working on the task when reaching the predicted score. I pay 4 Cents for each correct answer but payoff is only reduced by 2 Cents for a one unit deviation from the expected score. Thus, there is always an incentive to solve as many tasks as possible. Moreover, even if managers tried to reach their predicted score by counting the calculated tasks (which is not easy because the tasks are not numbered), they would not be sure of their score, because they would not know with certainty if their calculations are correct. The strategy to predict to solve no task correctly and not to work on the task is not profitable either because solving just one task correctly would lead to a higher payoff.

a better or worse collaborator in relation to the manager's *actual* performance. The payments for stage 2 are calculated with regard to the *actual* (and not the predicted) performance. This regard is necessary because otherwise managers could strategically assess themselves very high in order to win the bonus of 7.50 €.

3.2.3 Design of the Superiors Treatment (SUT)

In this treatment, each of 17 "superiors" gets associated to a manager of MAT and has to choose a collaborator for him. By a comparison of MAT and SUT, I investigate whether collaborator choices differ when a decision-maker (superior) selects a collaborator for another person (in this case for the manager) and not for himself. As an example, one can imagine a superior in an enterprise who recruits a new employee for his department. The superior allocates the new collaborator to a certain other employee in his department with whom he shall work together. Because the superior chooses the collaborator for his manager, the *manager's actual achievement* is also relevant in this treatment. As will be explained more precisely below, a superior's payoff due to his collaborator choice depends on the actual achievement of his associated manager, the selected collaborator's result, and the relation of the result of his team to those achieved by the teams of the reference group. In this treatment the superior's team consists of the associated manager and the superior's chosen collaborator. As in MAT, the manager's achievement is not known previously to the choice of a collaborator so that assessing his performance is of considerable importance in SUT.

The SUT is divided into two stages: At stage 1, a superior gets associated with a manager of MAT and has to predict this manager's performance.[52] At stage 2, the superior chooses a collaborator for his manager.[53]

Stage 1:

At stage 1, the superior is informed that a manager, who solved multiplication tasks for 10 minutes without using any auxiliary means except for pen and paper, has been associated

[52] Any superior is matched to a different manager, which yields 17 matched pairs. In my statistical analysis, I have to account for the fact that a superior and his associated manager are not independent of each other from a statistical point of view. The most important argument for this is that the manager's performance influences the behavior and payoff of the matched superior. When comparing the superiors with each other, tests for independent observations can be used, because superiors do not interact and have nothing to do with each other.

[53] The instructions of SUT can be found in the appendix (section 3.8).

to him. He is told that a typical task was "6 · 79" and that the manager got 4 Cents for each correct solution. Afterwards, the superior has to assess how many tasks the manager solved correctly. For his assessment he is paid according to the same scheme applied in MAT.[54] After stating his first assessment of the manager's performance (A1), the superior is told that his manager also had to assess how many tasks he would solve correctly. Besides, the superior is told how the manager was paid in dependence of his self-prediction. After being informed about the manager's first self-prediction (SP1), the superior can replace his assessment by a new one (A2). Then only the new, adjusted assessment is used for calculating his payoff (applying the same scheme). Thereafter, the superior is informed about the four subjects of BT who constituted the reference group for the managers in MAT and now also constitute the reference group for the superiors in SUT. Furthermore, he is informed about their performance. After having received this additional information, the superior can again consider his stated assessment and – if he wants – replace it by a new one (A3). Then, again, the payoff is solely calculated on the basis of this new assessment. In a last period of adjusting the assessment, the superior is informed about the revised self-prediction (SP2) the manager stated after he had learned the reference group's results. Again the superior can anew his assessment (A4). With the help of these four assessments, it can exactly be investigated to what degree a certain kind of information influences the superior's assessment. Additionally, the superior is always asked how sure he is of his respective assessment.

Stage 2:

At stage 2, a superior chooses a collaborator for his manager. The superior's payoff in SUT is calculated in the same way as for the managers in MAT. The payoffs consequently depend on the manager's actual achievement and the chosen collaborator's result. A superior's team consists of his associated manager and the superior's chosen collaborator. A superior is paid 7.50 € if his team scores the best or second best result of the five teams (as in MAT, the other four teams are composed of the reference subjects and their chosen collaborators). If the superior chooses a collaborator for his manager, whose result is at least as high as that of the manager, he obtains 2.50 € + 1 Cent · CR. Analogously to MAT, 5 € + 1 Cent · CR are paid if he chooses a collaborator who is worse than his manager. Before the superior chooses a collaborator for his manager, he has to assess which

[54] 1.80 € – {|actual number of the manager's correct solutions – assessed number of correct solutions| · 2 Cents}

collaborator was chosen by each of the four reference subjects. For each correct belief the superior obtains 50 Cents. Besides, the superior is asked how sure he is of his beliefs.

At the end of each treatment, subjects have to answer a questionnaire. The questionnaire includes questions concerning risk aversion. These questions are taken from the German Socio-Economic Panel (GSOEP) and deal with the overall risk behavior of subjects. I elicit subjects' risk aversion because their attitude towards risk can affect their collaborator choices: Winning the bonus is uncertain, while the payoff due to a worse collaborator is certain. If a subject is risk averse, he may prefer to voluntarily give up his chance of winning the bonus and employ a less qualified staff. A figure that gives an overview of the experimental design and the connection between my treatments can be found in the appendix (section 3.8).

3.2.4 Hypotheses

In this subsection I present my hypotheses. The first hypothesis deals with the self-predictions of managers in MAT and the second one refers to the performance assessments of superiors in SUT. Moreover, I state hypotheses concerning the collaborator choices of managers and superiors.

If people have to assess their own performance, overconfidence is generally considered to be a robust finding as is claimed by experimental studies in psychology and economics.[55] Moreover, some psychological studies show that subjects are overconfident in self-predictions – in the sense of being too optimistic: People have optimistic predictions about the time for completing a task (Buehler et al. 1994, Buehler and Griffin 2003), and overestimate the likelihood that they would engage in desirable behaviors (Epley and Dunning 2006). This aspect of overconfidence is most interesting in view of my experiment because the managers have to assess their performance before working on the task.[56] Because of the previous results I expect that managers overestimate their subsequent performance also in my real effort task and state a too high self-prediction SP1 compared to their subsequent performance.

[55] See, e.g., Camerer and Lovallo (1999), Fellner et al. (2004), and Weinstein (1980).
[56] In my study I use the term "overconfidence" in the sense that subjects overestimate their subsequent performance.

After getting information about the reference group's performance, managers can renew their self-predictions. The question arises whether this information helps to reduce their overconfidence. There are some psychological studies which investigate methods to de-bias subjects' self-assessments, for example by giving them feedback about their performance or train them in self-assessing.[57] These studies show ambiguous results: Some studies find that subjects get de-biased after getting information; others find that subjects ignore information and hence do not improve their self-assessments.[58] In these studies subjects get valuable information, i.e. information that can indeed help them with their self-assessments. In my experiment, it is not clear whether the information about the performance of the reference group can help managers to improve their self-predictions because this information is not specific to their own ability. As even some of the other studies find that subjects are not de-biased although obtaining valuable information, I expect that managers are still overconfident when stating their second self-prediction in my experiment. Therefore I state the following behavioral hypothesis concerning SP1 and SP2:

Hypothesis 3.1: Managers are overconfident when predicting their subsequent performance.

Superiors in SUT have to assess the managers' performances and can renew their assessments three times after getting information about SP1, the performance of the reference group, and SP2. Also in this treatment, it is interesting whether superiors regard the information about the reference group's performance as helpful when predicting the performance of their managers. Next to this question, I investigate how superiors evaluate the self-predictions of their managers. Do superiors believe that the self-predictions of their managers are accurate? Superiors may anticipate that the self-predictions of the managers are too high because overconfidence is a common bias. This idea is underlined by the results of Cesarini et al. (2006), who experimentally find that subjects anticipate the overconfidence of their peers. In my experiment I therefore expect that superiors think that managers are overconfident and thus do not take their self-predictions for sure and state lower performance predictions.

---

[57] See, e.g., Adams and Adams (1961), Lichtenstein and Fischhoff (1977), Pulford and Colman (1997), and Sharp et al. (1988).

[58] While for example Pulford and Colman (1997) and Sharp et al. (1988) find that feedback does mostly not influence overconfidence, Adams and Adams (1961) and Lichtenstein and Fischhoff (1977) find that feedback reduces overconfidence or can improve calibration, respectively.

Hypothesis 3.2: Superiors believe that managers are overconfident.

Concerning the collaborator choices in my treatments, I have to differentiate between those subjects who state a SP2 respectively an A4 that is higher than the performance of the second best reference person and those who state one that is lower.

If a manager thinks that he can correctly solve more tasks than the reference group's person with the second highest result, he will probably choose the best collaborator (with a result of 110). Then, in the eyes of the manager, his team will belong to the two best teams and he will get the bonus independent of the factual collaborator choices of the reference persons. This choice assures winning the bonus if the manager's subsequent performance indeed turns out to be higher than the result of the second best reference person. If the manager chooses a worse collaborator than he himself, he may obtain the bonus, too. However, this is combined with uncertainty because the manager does not know which collaborator the reference persons selected. Therefore I think that most managers, whose self-assessments are higher than the result of the second best reference person, choose the best collaborator. This is analogously true for the choices of the superiors if their A4 is higher than the result of the second best reference person. Therefore, I state the following hypotheses:

Hypothesis 3.3a: (Most) Managers whose self-predictions are above the performance of the second best reference person choose the best collaborator.

Hypothesis 3.3b: (Most) Superiors whose performance assessments are above the performance of the second best reference person choose the best collaborator.

Let me now consider the case in which the self-prediction of a manager or the assessment of a superior is below the result of the second best reference person. Here, the choice of a collaborator with a lower result than SP2 respectively A4 yields a sure payment of 5 € + 1 Cent · CR (instead of 2.50 € + 1 Cent · CR) independent of the collaborator choices of the reference group's persons. A subject with a low predicted performance might not think to have a chance to win the bonus. Therefore I expect that most of these selectors choose a

collaborator with a lower result than their own performance prediction. Hence, I state the following hypotheses:

Hypothesis 3.4a: ***(Most) Managers whose self-predictions are below the performance of the second best reference person choose a worse collaborator.***

Hypothesis 3.4b: ***(Most) Superiors whose performance assessments are below the performance of the second best reference person choose a worse collaborator.***

Of course, when choosing a worse collaborator, it is optimal to choose one whose result is only one point lower than the actual result of the manager. However, subjects cannot be sure that their performance predictions are correct. They may worry that the performance turns out to be lower than SP2 respectively A4. Therefore I expect that they want to make sure to *really* choose a worse collaborator by choosing one whose result is more than one point lower than their SP2 respectively A4.

What do the so far stated hypotheses suggest concerning a comparison between the collaborator choices of managers and superiors? If managers are indeed overconfident and superiors anticipate this, the average assessment of superiors will lie below that of the managers. If most of the superiors' assessments are below the second highest result of the reference group and most of the managers' predictions are above, I expect that fewer superiors than managers choose the best collaborator, given that most subjects believe that the two best reference persons have chosen the best collaborator. Therefore, I state my last hypothesis:

Hypothesis 3.5: More managers than superiors choose the best collaborator.

3.3 Results of BT

I conduct the BT to determine a typical reference group to which subjects in the main treatments are compared. In the group which I select as reference group the subjects correctly solved 43, 52, 64 respectively 81 multiplication tasks (mean 60). This group is the most "typical" one, because the average number of correctly solved tasks of the chosen reference group corresponds to the average number of correctly solved tasks over all

subjects in BT. Moreover, the result of the reference group's subject with the (second) highest result is very close to the average result of the subjects belonging to the (second) highest quartile. This is analogously true for the performances of the two worst subjects in the reference group. When investigating the collaborator choices of this group, I find that the best and the worst subject choose the best collaborator, while the other two subjects choose a collaborator who is one point worse than they themselves.

3.4 Results of MAT

In this section I start with the investigation of the managers' performance predictions. I compare the successive performance predictions, analyze the reactions to new information and consider the accuracy of the predictions. Then I examine the beliefs about the reference persons' choices. Both, performance predictions as well as beliefs are decisive for the collaborator choices of the respective selectors. Afterwards, I investigate the selection decisions in detail and study if they are rational from an ex ante point of view as well as from an ex post point of view.

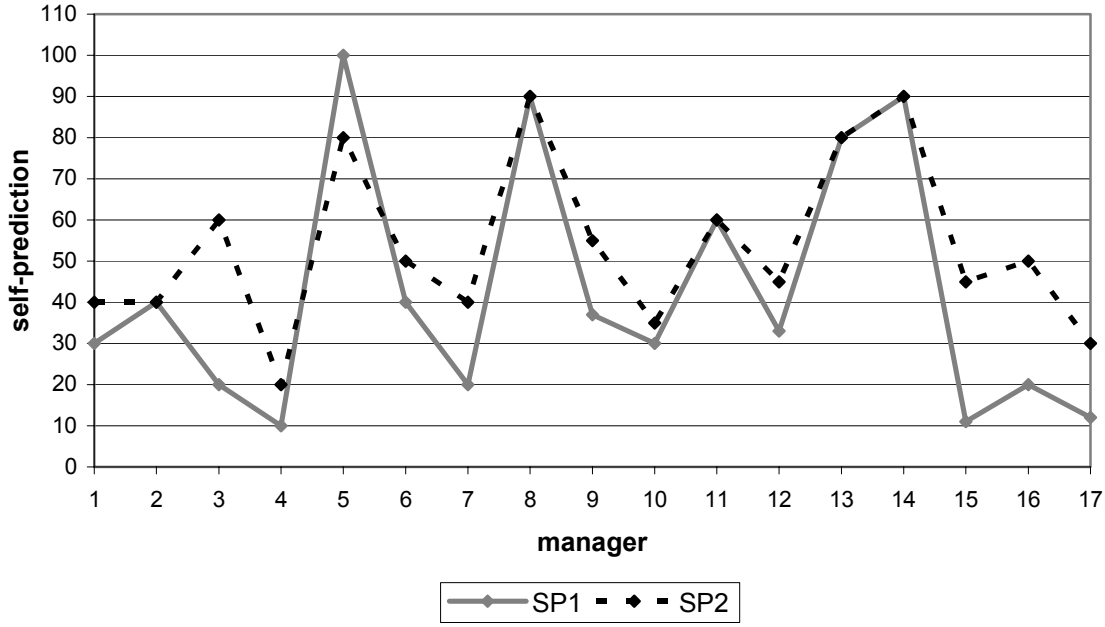
Self-Predictions

First, each manager is only informed about the kind of a typical multiplication task (namely "6 · 79") and states his first self-prediction (SP1). After being informed about the number of correctly solved tasks of the reference persons, managers state their second self-predictions (SP2). Figure 3.1 presents the self-predictions SP1 and SP2 of each manager. The average reported SP1 is 43, while the average reported SP2 is 54.[59] A Wilcoxon-Signed-Rank test for dependent observations shows that self-predictions 1 and 2 are significantly different from each other ($p < 0.01$, two-sided)[60], so managers renew their self-predictions in the light of new information. I observe that 70.59% of all managers adjust their first self-predictions into the direction of the average number of correctly solved tasks of the reference group (which is 60).

When comparing the absolute value of the difference 60 – SP2 with the absolute value of the difference 60 – SP1 for all managers, a significant difference can be found (Wilcoxon-Signed-Rank test, $p < 0.01$, two-sided)[61]. Thus managers seem to perceive the information

[59] The medians are 33 and 50; the modes are 20 and 40.
[60] This p-value is corrected for the 5 zero differences.
[61] This p-value is corrected for the 5 zero differences.

Figure 3.1: Self-predictions SP1 and SP2 in MAT



about the reference group as indication for their own performance in the multiplication task and adjust their SP1 into the direction of 60. This is particularly pronounced by those managers whose SP1 is lower than the average number of correctly solved tasks of the reference group (70.59%). Interestingly, 91.67% of them increase their assessments after receiving information about the reference group. Out of those managers who have stated a SP1 higher than the average number of correctly solved tasks of the reference group (23.53%), only 25% decrease their prediction, while the others do not change their predictions.

Apart from the mean, managers may consider the distribution of results of the reference group. Note that 70.59% of the managers state a SP1 that is lower than the worst result of the reference group. In contrast to this, only 17.65% state a SP1 which is higher than the best result of the reference group. After being informed about the results of the reference group, these managers know that their self-predictions are rather extreme. It is plausible that these managers adapt their SP2 accordingly, unless they are completely confident of their self-prediction. Indeed, out of 88.24% having an extreme SP1, 80% adjust their self-prediction. 46.67% of the managers with an extreme SP1 state an adjusted SP2 lying in the interval [43, 81], which is the range of the reference persons' results. The other 33.33% change their self-predictions in the direction of the average number of correctly solved tasks of the reference group but still have rather "extreme" SP2s.

To test Hypothesis 3.1 and to investigate how accurately managers predict their subsequent performance in the task, I compare their self-predictions SP1 and SP2 with their actual number of correctly solved tasks. A Wilcoxon-Signed-Rank test shows that I have to reject Hypothesis 3.1 since SP1 as well as SP2 are not significantly higher than the actual performance. Since I nevertheless want to analyze if managers' self-predictions deviate from their subsequent performance, I additionally apply a two-sided Wilcoxon-Signed-Rank test. While the test indicates that the first self-predictions SP1 are significantly different from the actual performance (Wilcoxon-Signed-Rank test, $p = 0.004$, two-sided), the second self-predictions SP2 are not. Contrarily to Hypothesis 3.1, the first self-predictions are lower than the subsequent performance: 76.47% of managers underestimate their performance without the information about the reference group. Thus, managers are biased (i.e. make systematical mistakes) in predicting their subsequent performance, but in the opposite direction as hypothesized. This finding can be underlined by the strand of psychological literature which deals with the so-called "hard easy effect" when assessing one's performance. As many authors find[62] absolute overconfidence is most extreme at tasks of great difficulty[63]. However, when tasks get easier, overconfidence is reduced. Indeed, subjects responding to very easy tasks are often underconfident. According to this literature, my findings can be explained by the fact that subjects are familiar with multiplication tasks and do not have severe problems to work on them. Thus, most managers reveal underconfidence when predicting their performance in this (easy) task.

Table 3.1: Average accurateness of self-predictions

		SP1	SP2
Difference between self-prediction and actual result	mean	– 14.94	– 3.94
	standard deviation	17.53	10.05
Absolute value of the difference between self-prediction and actual result	mean	19.29	8.53
	standard deviation	12.20	6.34

[62] See, for example, Clarke (1960), Lichtenstein and Fischhoff (1977), and Gigerenzer et al. (1991).
[63] Almost impossible tasks, which have been investigated, are for example the distinction between European and American handwritings and between Asian and European children's drawings.

Some information about the reference group's performance seems to really help subjects to accurately predict their performance. As Table 3.1 indicates, the average difference between the self-predictions and the actual results becomes smaller after the managers receive the information about the reference group.

Beliefs

If I want to interpret the choice of a collaborator, I have to pay attention to the managers' beliefs about the choices of the reference persons.[64] These beliefs are shown in Table 3.2. For example, row 1 pictures the managers' beliefs about the choice of the reference person with the lowest result in the multiplication task. "Score – 1" means that a collaborator, who has a one-point-lower score than the particular subject of the reference group, has been chosen. "110" specifies the choice of the best collaborator, and "other choice" indicates the choice of another collaborator. First of all, note that interestingly at most 17.65% of the managers do not have rational beliefs and/or think that a particular person of the reference group has not rationally chosen a collaborator. While most of the managers expect that the two persons with the lowest results have chosen a worse collaborator, the two persons with the highest results are expected to have chosen the best collaborator.

Table 3.2: Beliefs in MAT

	score – 1	**110**	**other choice**
belief1	82.35%	0	17.65%
belief2	70.59%	17.65%	11.76%
belief3	11.76%	76.47%	11.76%
belief4	17.65%	70.59%	11.76%

Choice of Collaborator

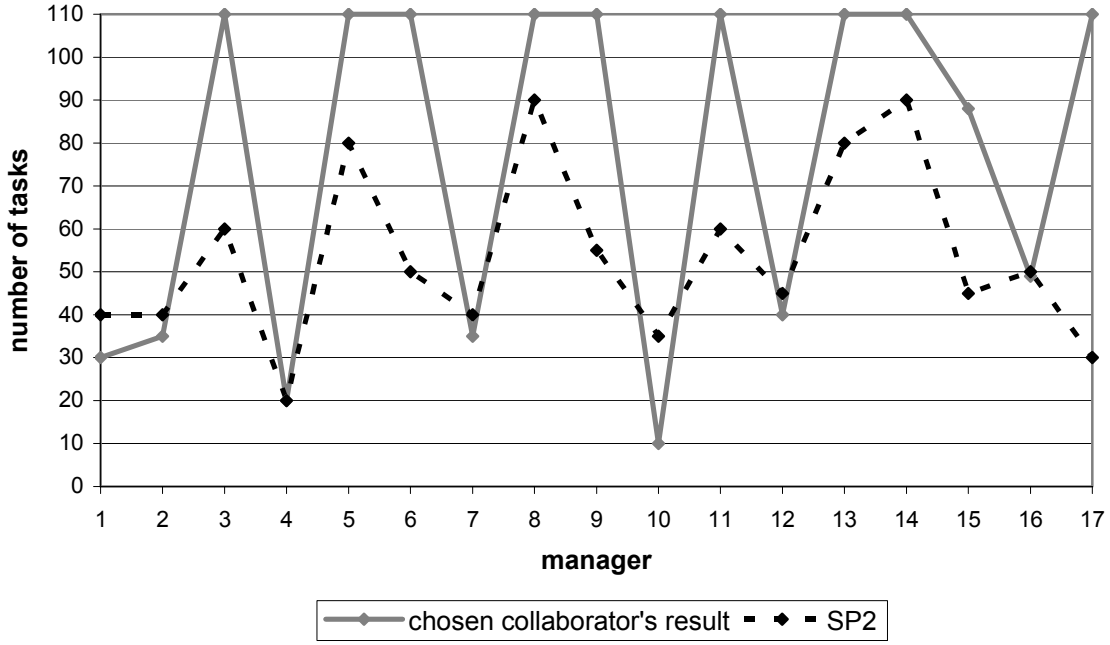
The managers' collaborator choices are shown in Figure 3.2. The best collaborator is chosen by 52.94% of the managers.[65] Interestingly, all managers who have a SP2 higher than the result of the second best reference person choose the best collaborator: A one-

[64] Beliefs are denoted belief1 till belief4, with belief1 denoting the expected choice of the worst subject of the reference group and belief4 denoting the expected choice of the best subject of the reference group.

[65] One may think that the collaborator choice is influenced by a manager's risk attitude. However, no significant correlation is found between the applied risk aversion measure and the collaborator choice.

sided Binomial test reveals that the probability of selecting the best collaborator is significantly higher than 0.5 ($p = 0.0625$). Therefore Hypothesis 3.3a cannot be rejected. As these managers believe that the two best reference persons have chosen the best collaborator they want to win a bonus by choosing the best collaborator, too.

Figure 3.2: Choice of collaborator and SP2 in MAT



There is one manager in MAT who chooses a collaborator with a higher result than his own self-prediction, but not the best collaborator. This might be explained by the fact that this manager fails to recognize that the choice of the best collaborator is most profitable when choosing a better collaborator. Another manager chooses a collaborator who is equal to his SP2. Thus his choice neither corresponds to a worse nor to a better collaborator. Because his SP2 is very low (just 20) I think that he actually hopes to solve more tasks correctly and therefore has the choice of a worse collaborator in mind.

A worse collaborator is chosen by 35.29% of the managers. Almost all of these managers choose a collaborator who is more than one point worse than their SP2. These managers want to assure that they really choose a worse collaborator compared to their subsequent performance. On average they choose a collaborator with a 10-point-lower result so that the average deduction for reasons of cautiousness is 9 points. Again I use a one-sided Binomial test to investigate whether most of the managers who have a SP2 lower than the result of the second best reference person choose a worse collaborator (Hypothesis 3.4a).

However, in this case I cannot reject the null hypothesis that the probability of choosing a worse collaborator is lower than or equal to 0.5. Indeed, only 46.15% of the managers who have a SP2 lower than 64 choose a worse collaborator.

It seems plausible that the proportion of managers who choose a better collaborator is higher for managers with SP2 > 64 than for managers with SP2 ≤ 64.[66] The results of a one-sided Fisher exact test show that the proportion of managers with SP2 > 64, choosing a better collaborator, is not significantly higher than the proportion of managers with an SP2 ≤ 64, choosing a better collaborator ($p = 0.115$, one-sided). However, there is a clear tendency in this direction. While all managers with SP2 > 64 choose a better collaborator, only half of the managers with SP2 ≤ 64 choose a better collaborator.

The so far conducted analysis of collaborator choices does not take note of the managers' beliefs. In order to incorporate their beliefs and to analyze the managers' choices in terms of rational behavior, I develop two rationality concepts: one from an ex ante and one from an ex post point of view. If a manager chooses a collaborator in line with his beliefs (belief1 to belief4) and SP2, I call his choice *ex ante rational*: When a manager states certain beliefs and a certain SP2, I can calculate whether he should choose a better or worse collaborator to maximize his payoff from an ex ante point of view. I call a manager *ex post rational*, if he chooses a collaborator in line with his *actual* performance and the *actual* collaborator choices of the reference persons. Clearly, this is a very theoretical approach because managers do neither know their actual performances nor the choices of the reference persons. But with this concept I can investigate whether the lack of information leads to a severe distortion of collaborator choices.

With both rationality concepts I investigate whether a manager chooses a better or a worse collaborator than SP2 and whether he should – from an ex ante or an ex post point of view – select a better or a worse one. Of course, I could also investigate the exact-point-choices and state whether these are ex ante (ex post) rational or not. For example, consider a manager who thinks that he will solve 50 calculations correctly and that the two best reference persons have chosen the best collaborator. Then he should – from an ex ante point of view – choose a 49-point-collaborator. But this choice cannot concretely be

[66] I exclude the manager whose collaborator choice equals his SP2 because he neither chooses a better nor a worse collaborator.

expected, because the manager is normally not totally sure that he will indeed correctly solve 50 calculations. This can be one reason for the choice of a collaborator who is a bit worse than 49. So, I think it would be misleading to describe this person as to be not ex ante rational. Thus, I use concepts of rationality which just control for the choice of a better or a worse collaborator and not for an exact-point-choice.

Examining the data reveals that only 37.5% of the managers behave ex ante rationally.[67] 33.33% of those choose a better collaborator. More than half of the ex ante irrational managers choose a better collaborator, although a worse collaborator would be rational, given their SP2 and beliefs. If managers state a relative low SP2 and irrationally choose a better collaborator, this may be explained by "wishful thinking". At the moment when managers choose a better collaborator, they may think that their stated self-prediction was too low. While hoping for a better performance, they may be tempted to try to win the bonus by choosing a better collaborator. If their subsequent performance turns out to be higher than 64, they will indeed obtain a bonus. This supposition can be corroborated by the observation that 66.67% of the managers, who irrationally choose a better collaborator and have a SP2 lower than 64, have self-predictions in the interval [50, 60]. These self-predictions are not far away from 64. Since these managers believe that the reference group's subjects with the two highest results choose the best collaborator, they try to win the bonus by choosing a better collaborator, too. Another explanation for irrationally choosing a better collaborator concerns managers with a very high SP2: Managers with an extreme high SP2 who should from an ex ante point of view choose a worse collaborator to maximize their payoff (given their statements they will win the bonus also by choosing a worse collaborator) are of course tempted to choose the best one to be sure to get the bonus. They may be afraid that their performance will turn out to be lower than predicted.

If I assess the collaborator choices from an ex post point of view, I observe that 41.18% of the managers behave ex post rationally. 60% of the managers, who do not behave ex post rationally, choose a better collaborator. They would have done better by choosing a worse one.

[67] I exclude the manager who chooses a collaborator equal to his own SP2 because the collaborator is neither better nor worse than SP2.

In summary, the results of MAT show that most managers underestimate their subsequent performance without any additional information which seems to be a consequence of the so-called "hard easy effect". Information about the reference group is very valuable to managers: Most of them adjust their self-predictions in "the right direction". Thus I do not find a significant difference between the second self-predictions and the actual performance anymore. As hypothesized, most managers whose SP2 is above the result of the second best reference person choose the best collaborator. In contrast to Hypothesis 3.4a, only 46.15% of those managers whose SP2 is lower than the result of the second best reference person choose a worse collaborator. Furthermore, only 37.5% of managers choose their collaborator ex ante rationally. Interestingly, many of the ex ante irrational managers choose a better collaborator, although there seems to be no chance to win the bonus.

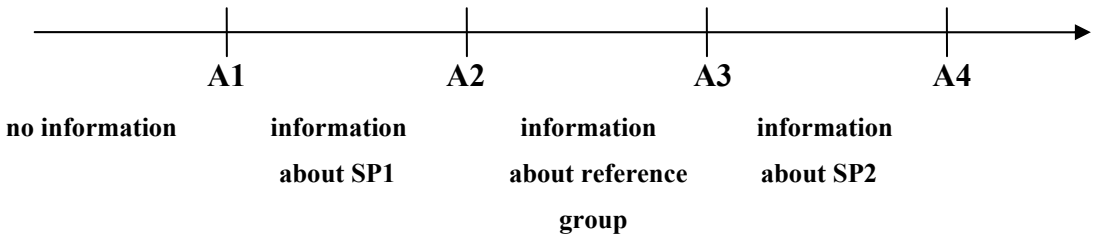
3.5 Results of SUT

In this section I first investigate the superiors' assessments of the managers' performances and the influence of information on these assessments. Afterwards, I analyze the superiors' beliefs and collaborator choices. Moreover, I examine whether the superiors' choices are ex ante rational, given their assessments and beliefs, and whether they are rational from an ex post point of view.

Superiors' Assessments

A superior has the opportunity to revise his assessments three times after getting more and more information (see Figure 3.3). Starting with an average assessment of 46, the second

Figure 3.3: Information and assessments in SUT



average assessment is somewhat lower (41). After receiving information about the performance of the reference group's subjects, the superiors' average assessment becomes higher (50), with the highest after the information about SP2 (51).[68]

Table 3.3: Average accurateness of assessments

		A1	A2	A3	A4
Difference between assessment and actual result	mean	– 11.41	– 16.59	– 7.29	– 6.59
	standard deviation	30.19	17.41	18.64	17.33
Absolute value of the difference between assessment and actual result	mean	26.94	19.53	16.12	14.47
	standard deviation	16.68	13.80	11.31	11.14

How accurately do superiors assess their managers' performance? To get an insight into this issue, Table 3.3 shows the average differences between superiors' assessments A1 to A4 and their managers' actual results. I make two different computations: One takes account of the signs; the other regards the absolute values of the respective differences. Both cases show that the last assessment, which incorporates all information, corresponds best to the actual result.

Moreover, the first row shows that superiors underestimate managers' performances on average. However, a two-sided Wilcoxon-Signed-Rank test reveals that there is only a significant difference between A2 and the actual performance ($p = 0.002$, two-sided). A2 is significantly lower than the actual performance. Keeping the payoff structure of my experiment in mind, the four assessments should only differ if the superiors believe to earn more if they renew their assessment after obtaining more information. To investigate the influence of information, I first examine whether the superiors' assessments differ from each other. I find significant differences between A2 and A3 and between A2 and A4, with A2 being lower in both comparisons (Wilcoxon-Signed-Rank test, $p < 0.01$ und $p < 0.01$, both two-sided)[69].

[68] Medians are always lower: 38, 35, 48, and 50, respectively. The modes are 57, 35, 45, and 50.
[69] These p-values are corrected for the 8 respectively 5 zero differences.

Reactions to Information about SP1

To investigate in more detail whether superiors adjust their assessments to the new information about SP1, I compare the absolute difference between SP1 and the former assessment A1 with the absolute difference between SP1 and the new assessment A2. If superiors take SP1 for sure, they should renew their assessment in the corresponding direction. A comparison between the absolute values of the differences SP1 – A1 and SP1 – A2 reveals a significant difference (Wilcoxon-Signed-Rank test, $p < 0.01$, two-sided)[70] with SP1 – A1 being higher. Most superiors take the information of SP1 into consideration and shift their second assessment in the direction of SP1.

It is an interesting aspect to analyze how much a superior trusts the self-prediction of his manager. On the one hand, SP1 is doubtful information because a superior does not know whether his manager is good in predicting his own performance. On the other hand, it is – at this stage of the experiment – the only information that is directly connected with the manager whom he has to assess. Examining the superiors' reactions to SP1 more closely shows that only 17.65% of them state an A2 equal to SP1. Moreover, 11.76% do not change their assessment and therefore ignore SP1. The others modify their assessment in the direction of SP1. Interestingly, none of these superiors changes his assessment in such a way that his new assessment A2 is "at the other side" of SP1 viewed from A1: If A1 is below SP1 (50% of these superiors), A2 is below SP1, too, but less distant. If A1 is above SP1, A2 is also above SP1 but less distant. As my data show, most superiors take the information SP1 for serious but do not completely take note of it and keep their first assessment in mind.[71] These superiors do not think that their managers are completely right with their self-prediction.

To test Hypothesis 3.2 regarding SP1, a one-sided Wilcoxon-Signed-Rank test shows that SP1 is not significantly higher than A2. Therefore, Hypothesis 3.2 can be rejected regarding SP1. Furthermore, also SP1 and A1 do not significantly differ. Hence, already the first predictions are quite similar.

[70] This p-value is corrected for the 2 zero differences.

[71] In psychology, there is much evidence that once people have formed an opinion, they cling to it too tightly and far too long (e.g. Lord et al. 1979). This phenomenon is commonly called "belief perseverance".

Reactions to Information about the Reference Group

To investigate superiors' reactions to the next information, I take the average performance of the reference group into consideration. Thus I take the average number of correctly solved tasks of the reference group as comparison value in my tests. One may argue that superiors consider the distribution of performances next to the average performance of the subjects (see MAT). This can be particularly relevant if a superior's A2 is higher (lower) than the result of the best (worst) subject of the reference group. But also in these cases, the superiors may adapt A2 into the direction of the mean of the reference group. Therefore, considering the average performance of the reference group in my tests is not misleading.

Comparing the absolute values of the differences 60 – A2 and 60 – A3 shows that they are significantly different (Wilcoxon-Signed-Rank test, $p < 0.01$, two-sided)[72], with 60 – A2 being higher. A lot of superiors seem to revise their assessment in the light of new information. Investigating their reactions in detail, I observe that 47.06% of the superiors move A3 in the direction of 60 but remain below it, as are all their A2s. These superiors regard the average result of the reference group but do not forget their former assessments. Moreover, one superior states an A3 of 60. This superior seems to assume that the reference group is a really typical group. Contrarily, 47.06% of the superiors do not change their assessments and retain A2. These superiors either feel their assessment to be supported by the information about the reference persons or do not think that the performance of the reference group is helpful information for predicting the manager's performance.

Reactions to Information about SP2

The last information a superior gets is the SP2 of his manager. Again, I observe a significant difference between the absolute values of the differences SP2 – A3 and SP2 – A4 (Wilcoxon-Signed-Rank test, $p < 0.05$, two-sided)[73], with SP2 – A3 being higher. Thus the information about SP2 is very often taken for serious; 41.18% of the superiors take this information into consideration and adjust their next assessment in the direction of the new information, or even state an A4 which equals SP2 (23.53% of these subjects). However, 52.94% of the superiors do not change their A3. Indeed 17.65% state an A3 that is already

[72] This p-value is corrected for the 8 zero differences.
[73] This p-value is corrected for the 9 zero differences.

equal to SP2. The others who do not change their A3 do not make use of SP2. To test Hypothesis 3.2, I conduct a one-sided Wilcoxon-Signed-Rank test that demonstrates that SP2 is not significantly higher than A4. Therefore, Hypothesis 3.2 can be rejected regarding SP2. Moreover, A3 and SP2 do not significantly differ.

An interesting finding can be detected if I divide the SP2s into two categories: SP2 < 60 and SP2 ≥ 60. Superiors who get to know a SP2 lower than the average performance of the reference group may assume that their managers' self-consciousness is too low. A two-sided Wilcoxon-Signed-Rank test reveals that A4 is significantly different from SP2 for the first category ($p < 0.1$)[74] but not significantly different for the second. If the SP2 of a manager is lower than 60, most of the matched superiors state an A4 that is higher than SP2. This hints at the fact that superiors, whose manager predicts a performance below average, expect that the manager is underconfident.

Beliefs

As can be seen from Table 3.4, most superiors believe that the worst and second worst reference subjects choose a collaborator with a one-point lower result. Moreover, most of them believe that the two best reference subjects choose the best collaborator. Nonetheless, there are also many superiors who believe that the best two reference subjects choose a one-point-worse collaborator.

Table 3.4: Beliefs in SUT

	score – 1	**110**	**other choice**
belief1	82.35%	5.88%	11.76%
belief2	70.59%	17.65%	11.76%
belief3	29.41%	52.94%	17.65%
belief4	41.18%	47.06%	11.76%

[74] This p-value is corrected for the 4 zero differences.

Choice of Collaborator

35.29% of the superiors choose the best possible collaborator, while 52.94% choose a worse collaborator compared to their stated A4.[75] There are two other superiors, who choose a better collaborator, but not the best one. This selection is not rational because the choice of the best collaborator would increase their payoffs. There are two different explanations for their behaviors. First, it may be that these superiors simply do not understand that the choice of the 110-point-collaborator is most profitable when choosing a better collaborator. Second, it may be that these superiors increase their assessments after stating A4 or forget what they exactly wrote down as A4 and have a higher value in mind. In these cases the collaborator choice of the two "irrational" superiors can be interpreted as if these superiors purposed the selection of a worse collaborator. This seems to be the case at least for one of these two superiors: This superior states an A4 of 50, while the chosen collaborator's result is 55. Moreover, this superior writes in the questionnaire that he has chosen a worse collaborator for his manager. This indicates that this superior indeed intended to choose a worse collaborator for his manager. Nevertheless, these superiors have chosen a better collaborator on their decision sheet, so I treat them as if they wanted to choose a better collaborator.

Examining the data in detail shows that 66.67% of the superiors who choose a worse collaborator select one who is more than one point worse than their stated A4. On average, the difference between the chosen collaborator's result and A4 is 18 for these superiors. Similar to some managers in MAT these superiors include a deduction of points for reasons of cautiousness. They really want to make sure to choose a collaborator who is worse than the actual performance of the associated manager. One of these superiors even chooses a 0-point-collaborator to ensure that his manager is better.

To examine whether I can find support for Hypotheses 3b and 4b, I separately analyze the choices of superiors whose performance assessments are above respectively below the performance of the second best reference person. Hypothesis 3.3b cannot be rejected: A Binomial test reveals that I can reject the null hypothesis that the probability for choosing the best collaborator is lower than or equal to 0.5 ($p = 0.0625$). All superiors who have an A4 that is higher than the performance of the second best reference person choose the best

[75] One may think that the collaborator choice is influenced by a superior's risk attitude. However, no significant correlation is found between the applied risk-aversion measure and the collaborator choice.

collaborator to assure to get the bonus if the manager's subsequent performance indeed turns out to be higher than the result of the second best reference person. Similar as in MAT, I have to reject Hypothesis 3.4b: Although most of the superiors (69.23%), whose performance assessments are below the performance of the second best reference person, choose a worse collaborator, a one-sided Binomial test indicates that I cannot reject the null hypothesis that the probability of choosing a worse collaborator is lower than or equal to 0.5.

It seems plausible that the proportion of superiors, who choose a better collaborator, is higher for superiors with $A4 > 64$ than for superiors with $A4 \leq 64$. Indeed, a one-sided Fisher exact test shows that the proportion of superiors with an $A4 > 64$, who choose a better collaborator, is significantly higher than the proportion of superiors with an $A4 \leq 64$, who choose a better collaborator ($p = 0.029$, one-sided). This shows that the performance assessments have a crucial influence on the choices of superiors.

However, note that this analysis neglects the beliefs of the superiors. Thus in a next step I incorporate beliefs and investigate whether superiors' collaborator choices are *ex ante rational* and therefore consistent with their stated beliefs (belief1 to belief4) and A4.[76] 58.82% of the superiors behave ex ante rationally. Only 28.57% of those superiors, who do not behave ex ante rationally, choose a better collaborator. Thus, in this treatment too many superiors choose a worse collaborator from an ex ante point of view.

Another interesting aspect is the evaluation of the superiors' choices with regard to the managers' actual performances and the actual choices of the reference persons. As observed in BT, the best and the worst subject of the reference group choose the 110-point-collaborator and the other two subjects a collaborator with a one-point-lower result. If these choices and the actual performance of the associated manager in MAT were known by the superior, he could precisely choose the collaborator who maximizes his payoff. The data show that 52.94% of the superiors behave ex post rationally. Of those who do not behave ex post rationally 50% choose a better collaborator although they should choose a worse one.

[76] Again I disregard the exact-point-choices and only take into account whether superiors choose a better or worse collaborator.

3.6 Comparison of MAT and SUT

In this section I compare the accurateness of performance predictions, the beliefs, and the collaborator choices across MAT and SUT.

Performance Predictions

First, I regard the managers' and superiors' performance predictions. I cannot find any significant differences between SP1 and A1 as well as between SP2 and A4. I also do not observe a significant difference if I compare the accurateness of performance predictions across treatments, neither when taking note of the signs nor when comparing the absolute values.[77] Thus, managers do not predict their subsequent performance systematically more accurately than superiors, neither before getting any information nor after obtaining information.

Beliefs

Next, I compare the beliefs about the reference persons' collaborator choices: Concerning belief1 and belief2 most managers and most superiors think that the two worst reference subjects choose a worse collaborator. But regarding belief4 there is a noticeable difference: Much more superiors than managers believe that the best reference person chooses a worse collaborator. This tendency can also be seen for belief3. Nevertheless, a Wilcoxon-Signed-Rank test shows that there is no significant difference when comparing the beliefs of managers and superiors about a certain reference person with each other.

Collaborator Choice

I now compare the collaborator choices across MAT and SUT. As has already been mentioned, 52.94% of the managers and 35.29% of the superiors choose the best collaborator. Contrarily, 35.29% of the managers and 52.94% of the superiors choose a collaborator with a worse result than SP2 respectively A4. The average result of the chosen collaborators in MAT is 76 and 64 in SUT. This shows that the collaborators in MAT have on average higher results. If I only compare the chosen collaborators, who are worse than the subject's SP2 respectively A4, their average results are very similar (33 in MAT, 31 in

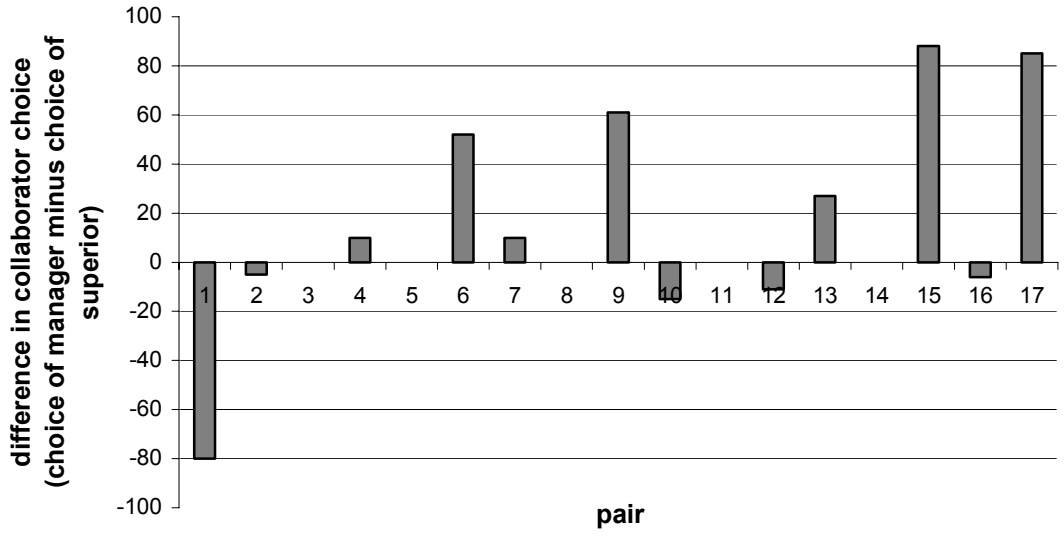
[77] While a sign test reveals no significant difference between |SP2 – actual performance| and |A4 – actual performance|, a Wilcoxon-Signed-Rank test does. However, one assumption of the Wilcoxon-Signed-Rank test is that the difference of the compared values has to be symmetrically distributed. The particular difference of the values of |SP2 – actual performance| and |A4 – actual performance| is rather asymmetrically distributed so that the results of the sign test are more meaningful and therefore presented here. For the other comparisons both tests do not show a significant difference.

SUT). Interestingly, the average deduction due to reasons of cautiousness is much lower for managers (9) than for superiors (17). This may reflect that superiors, who want to be sure to choose a worse collaborator, are less sure of their managers' performance than managers are of their own performance.

If I compare the collaborator choices of managers and superiors, I also have to account for the fact that their self-predictions and assessments differ, even if not significantly. The data show that 23.53% of the managers and also 23.53% of the superiors state a SP2 respectively A4 which is higher than 64. All of these managers and superiors choose the best collaborator. Thus, these subjects want to assure to win the bonus. 76.47% of all managers and superiors state a SP2 respectively A4 that is lower than 64. In these cases, more managers than superiors choose a better collaborator than SP2 respectively A4 (46.15% of managers and 30.77% of superiors).

When considering the matched pairs of managers and superiors, there are noticeable differences in collaborator choices. Figure 3.4 illustrates the difference in collaborator choices (choice of a manager minus choice of the matched superior) for each matched manager-superior pair. As can be seen, only in 29.41% of all pairs, managers and superiors exactly choose the same collaborator (which is the best collaborator in all these cases). In

Figure 3.4: Differences in collaborator choices between MAT and SUT for matched pairs



41.18% of the matched pairs a manager chooses a better collaborator than the matched superior. In 57.14% of these cases the manager chooses a better collaborator than SP2, while the superior chooses a worse collaborator than A4. Contrarily, in 29.41% of the matched pairs the superiors choose a better collaborator than the managers. In 60% of these cases, both choose a collaborator worse than SP2 respectively A4 but the superiors' collaborators have higher results than those chosen by the managers.

To test Hypothesis 3.5, I conduct a one-sided McNemar test. It reveals that I have to reject Hypothesis 3.5 because there are not significantly more managers than superiors who choose the best collaborator. Furthermore, there is no significant difference when testing the collaborator choices across treatments with a Wilcoxon-Signed-Rank test.

Next to these aspects, I am interested in the question whether there are differences between managers' and superiors' behaviors concerning ex ante and ex post rationality. While only 37.5% of the managers behave ex ante rationally, 58.82% of the superiors are ex ante rational. Investigating this phenomenon in more detail, 80% of those managers who do not behave ex ante rationally choose a better collaborator. They should however choose a collaborator whose result is worse than their SP2 from an ex ante point of view. Having a look at the superiors, 28.57% of those who do not behave ex ante rationally choose a better collaborator although they state a very low A4. Therefore, this kind of mistake seems to happen more often in MAT than in SUT. A possible reason for this may be that more managers than superiors have a kind of "wishful thinking". Perhaps more managers hope to win the bonus – in contrast to their self-predictions and beliefs. They may hope to be better in the multiplication tasks than predicted before, or hope to have the luck that all subjects in the reference group choose a worse collaborator. Superiors do not seem to be so hopeful. They seem to condition their behavior better on the relevant statements.

If I compare the proportions of managers and superiors whose collaborator choice is ex post rational, I find that more superiors (52.94%) than managers (41.18%) are ex post rational.

3.7 Summary and Conclusion

My results reveal that managers' self-predictions (SP1) are biased if their information only consists of an example of a typical multiplication task. In contrast to my Hypothesis 3.1

managers significantly underestimate their subsequent performance. Even about 70.59% of them assess themselves worse than the worst reference person (whose result is not known to them at that time). This result can be explained by the "easiness" of my task and is in line with the so-called "hard easy effect". Interestingly, the accuracy of the managers' final predictions is remarkably improved, after managers are informed about the reference group's results. The very most of them adjust their predictions in the direction of the reference group's average result so that no significant difference to the subsequent performance remains.

Similarly, a very good information processing and adjustment of assessments can be observed for the superiors. Superiors significantly adjust all their assessments in the direction of the respective new information. However, they do not completely take note of the new information and keep their former assessment in mind ("belief perseverance"). They are especially sensitive to the information about the reference persons' performances, apparently regarding those as really factual data. As expected, A4 is the most accurate assessment of the managers' performances. It is not significantly different from SP2 and from the managers' factual performance. However, considering only superiors who are matched with a manager who states a SP2 that is lower than the average result of the reference group reveals that superiors' last assessments are significantly higher. Thus these superiors suppose that their managers are underconfident.

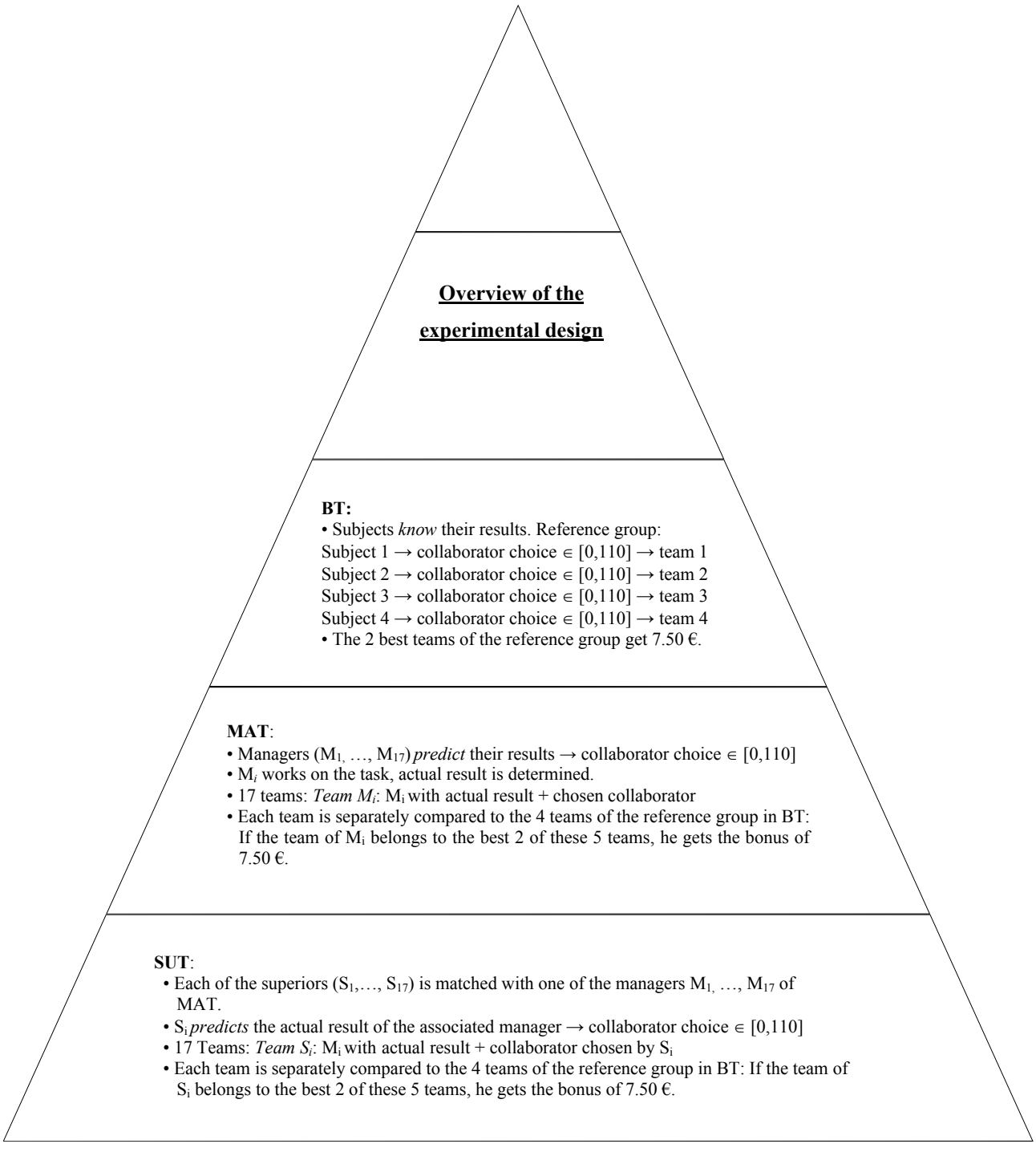
As hypothesized, most managers and most superiors whose SP2 respectively A4 is above the result of the second best reference person choose the best collaborator. Furthermore, only 46.15% of managers and 69.23% of superiors, whose predictions are lower than the result of the second best reference person, choose a worse collaborator.

Looking at the matched pairs shows that in 41.18% of the pairs the managers choose a better collaborator than the superiors but the differences are not statistically significant. Thus the collaborator choice does not seem to depend on the "authority" of the decision maker, i.e. whether he chooses a collaborator for himself or a collaborator selected by a third person is allocated to him. This result is a consequence of the excellent information processing and the excellent assessments of both, managers and superiors. Nevertheless, to a considerable degree I observe irrationality in choosing collaborators. Astonishingly, only 37.5% of the managers choose their collaborator ex ante rationally. In many cases, they

prefer the best possible one although they have almost no chance to win a bonus according to their self-predictions and beliefs. Such behavior is most likely caused by a kind of "wishful thinking": They hope to win the bonus in spite of a rather bad precondition. Likewise the superiors do not choose collaborators for their managers in a notably rational way. 41.18% of the choices are irrational in the light of their information and the given incentives. However, "wishful thinking" is less frequently observed for the superiors.

In my experiment, the collaborator choices depend on performance (self-)predictions, beliefs about the reference subjects' choices, and the given incentives. It would be interesting to examine in a further treatment, whether and how a change in incentives affects the collaborator choices. If the bonus is increased, I expect that more subjects try to win the bonus. Perhaps "wishful thinking" is more widespread in such an environment. Then even more subjects may choose the best collaborator, although their performance prediction and beliefs tell them to choose a worse one. This may also be the case if the bonus remains at 7.50 € but the amount of money, a subject gets for sure if choosing a worse collaborator, is decreased. More subjects may then compete harder by choosing the best collaborator, because the strategy to choose a worse collaborator does not lead to a comparably high payoff anymore.

3.8 Appendix


Overview of the experimental design
BT:
• Subjects *know* their results. Reference group:
Subject 1 → collaborator choice ∈ [0,110] → team 1
Subject 2 → collaborator choice ∈ [0,110] → team 2
Subject 3 → collaborator choice ∈ [0,110] → team 3
Subject 4 → collaborator choice ∈ [0,110] → team 4
• The 2 best teams of the reference group get 7.50 €.
MAT:
• Managers (M_1, …, M_{17}) *predict* their results → collaborator choice ∈ [0,110]
• M_i works on the task, actual result is determined.
• 17 teams: *Team* M_i: M_i with actual result + chosen collaborator
• Each team is separately compared to the 4 teams of the reference group in BT: If the team of M_i belongs to the best 2 of these 5 teams, he gets the bonus of 7.50 €.
SUT:
• Each of the superiors (S_1,…, S_{17}) is matched with one of the managers M_1, …, M_{17} of MAT.
• S_i *predicts* the actual result of the associated manager → collaborator choice ∈ [0,110]
• 17 Teams: *Team* S_i: M_i with actual result + collaborator chosen by S_i
• Each team is separately compared to the 4 teams of the reference group in BT: If the team of S_i belongs to the best 2 of these 5 teams, he gets the bonus of 7.50 €.

Instructions of SUT in German

Überblick über das Experiment

Sie nehmen an einem wirtschaftswissenschaftlichen Entscheidungsexperiment teil. Sämtliche Entscheidungen sind anonym. Um dies sicherzustellen, agieren Sie unter einem bestimmten Codenamen. Es ist nicht möglich, Entscheidungen und Angaben, die Sie unter diesem Codenamen machen, mit Ihrer wahren Identität in Verbindung zu bringen.

Bitte lesen Sie sich die Beschreibungen der einzelnen Stufen des Experiments jeweils sorgfältig durch. Wenn Sie etwas nicht verstehen sollten, schauen Sie bitte noch einmal in die entsprechende Anleitung. Sollten Sie dann noch Fragen haben, geben Sie uns bitte ein Handzeichen.

Das Experiment ist in verschiedene Stufen unterteilt: Auf Stufe 1 bekommen Sie als Manager einen Agenten zugeteilt, über den Sie bestimmte Informationen erhalten. Auf Stufe 2 werden Sie für den Agenten einen Mitarbeiter auswählen.

Für dieses Experiment sind zwei weitere Experimente relevant (*Experiment A* und *Experiment B*), die bereits stattgefunden haben. Aus *Experiment A* ist für Sie eine Vergleichsgruppe von vier Personen relevant. Aus *Experiment B*, das einige Wochen nach dem *Experiment A* stattgefunden hat, ist für Sie eine einzelne Person als Ihr Agent relevant.

Stufe 1: Ihr Agent

Ihr Agent aus *Experiment B* hat 10 Minuten lang ohne Hilfsmittel (nur mit Papier und Stift) Multiplikationsaufgaben gelöst. Für jede richtig gelöste Aufgabe erhielt er **4 Cent**. Ein typisches Beispiel für diese Aufgaben ist **6 • 79**.

Sie werden gleich einen Bogen erhalten, auf dem Sie gefragt werden, was Sie glauben, **wie viele** dieser **Aufgaben** Ihr Agent in 10 Minuten **richtig** gelöst hat.

Stimmt Ihre Einschätzung genau mit der tatsächlichen Anzahl richtig gelöster Aufgaben Ihres Agenten im *Experiment B* überein, erhalten Sie **1,80 €** für Ihre Einschätzung. Gibt es **keine genaue Übereinstimmung**, so berechnet sich Ihre Auszahlung für die Einschätzung wie folgt: Es wird die **Differenz** aus der **tatsächlich erreichten Anzahl richtiger Lösungen** Ihres Agenten im *Experiment B* und **Ihrer geschätzten Anzahl richtiger Lösungen** berechnet. Der *Betrag* der Differenz wird mit **2 Cent** multipliziert. Die positive Zahl, die sich daraus ergibt, wird von **1,80 €** abgezogen. Diesen **Restbetrag**, dessen untere Grenze wir auf Null festlegen, erhalten Sie als Auszahlung für Ihre Einschätzung.

Formal kann man Ihre Auszahlung in Abhängigkeit der Genauigkeit Ihrer Einschätzung folgendermaßen darstellen:

1,80 € - (|**tatsächliche** Anzahl richtiger Lösungen Ihres Agenten - **geschätzte** Anzahl richtiger Lösungen| · **2 Cent**)

Ihre Auszahlung ist also umso höher, je genauer Sie die Anzahl richtig gelöster Aufgaben Ihres Agenten einschätzen.

Codename:

Einschätzung Ihres Agenten

Geben Sie bitte an, was Sie denken, wie viele Aufgaben Ihr Agent im Experiment B richtig gelöst hat:

Ich denke, dass mein Agent _____ Aufgaben richtig gelöst hat.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die Selbsteinschätzung Ihres Agenten

Auch Ihr Agent wurde zu Beginn des *Experiments B* nach seiner Einschätzung gefragt, wie viele Multiplikationsaufgaben er in **10 Minuten** richtig lösen würde. Er erfuhr, dass er für jede richtig gelöste Aufgabe **4 Cent** erhalten würde, und dass **6 • 79** ein typisches Beispiel für diese Aufgaben sei. Für seine Einschätzung wurde er analog zu Ihnen entlohnt: Stimmte seine Einschätzung genau mit der später erreichten Anzahl richtig gelöster Aufgaben überein, erhielt er **1,80 €** für seine Einschätzung. Gab es **keine genaue Übereinstimmung**, so berechnete sich seine Auszahlung für die Einschätzung wie folgt: Es wurde die **Differenz aus der tatsächlich erreichten Anzahl richtiger Lösungen und der geschätzten Anzahl richtiger Lösungen** berechnet. Der *Betrag* der Differenz wurde mit **2 Cent** multipliziert. Die positive Zahl, die sich daraus ergab, wurde von **1,80 €** abgezogen. Diesen **Restbetrag** erhielt der Agent für seine Einschätzung (die untere Grenze des Restbetrages wurde auf Null festgelegt). Formal kann man die Auszahlung des Agenten in Abhängigkeit der Genauigkeit seiner Einschätzung folgendermaßen darstellen:

1,80 € -(|**tatsächliche** Anzahl richtiger Lösungen - **geschätzte** Anzahl richtiger Lösungen|· **2 Cent**)

Sie werden gleich die Selbsteinschätzung Ihres Agenten erfahren. Danach können Sie - falls Sie möchten - Ihre Einschätzung über die tatsächliche Anzahl richtig gelöster Aufgaben Ihres Agenten ändern. Im Fall einer **erneuten** Einschätzung wird nur Ihre **neue Einschätzung entlohnt**, die alte entfällt. Für die neue Einschätzung gilt dann wieder das Entlohnungsschema, das wir Ihnen bereits vorgestellt haben. **Verändern** Sie Ihre Einschätzung **nicht**, so wird Ihre **alte Einschätzung** wie gehabt **entlohnt**.

Die Selbsteinschätzung Ihres Agenten lautet:

Ihre erneute Einschätzung

Möchten Sie Ihre Einschätzung darüber ändern, wie viele Aufgaben Ihr Agent im Experiment B tatsächlich richtig gelöst hat?:

☐ Nein, ich ändere meine Einschätzung nicht.

☐ Ja. Ich denke, dass mein Agent _____ Aufgaben richtig gelöst hat.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die Vergleichsgruppe Ihres Agenten

Nachdem sich Ihr Agent im *Experiment B* selbst eingeschätzt hatte, erfuhr er von vier Personen aus *Experiment A*, die seine Vergleichsgruppe bildeten. Er erfuhr, dass diese Personen **dieselben** Multiplikationsaufgaben bearbeitet hatten, die er noch bearbeiten würde, und dass auch sie 4 Cent pro richtiger Lösung erhalten hatten. Außerdem erfuhr er, dass diese **vier Personen** der **Vergleichsgruppe** folgende Anzahl an Aufgaben richtig gelöst hatten:

43, 52, 64, 81.

Sie haben nun die Möglichkeit, Ihre Einschätzung, wie viele Aufgaben Ihr Agent im *Experiment B* wirklich richtig gelöst hat, zu ändern. Im Fall einer **erneuten** Einschätzung wird nur die **neue Einschätzung** nach dem bereits vorgestellten Schema **entlohnt**, die alte entfällt. **Verändern** Sie Ihre Einschätzung **nicht**, so wird Ihre **alte Einschätzung** wie gehabt **entlohnt**.

Füllen Sie nun bitte diesen Bogen aus:

Ihre erneute Einschätzung

Die Selbsteinschätzung Ihres Agenten lautet:

Die Personen der Vergleichsgruppe Ihres Agenten hatten die folgende Anzahl an Multiplikationsaufgaben richtig gelöst: **43, 52, 64, 81**

Möchten Sie Ihre Einschätzung darüber ändern, wie viele Aufgaben Ihr Agent im Experiment B tatsächlich richtig gelöst hat?:

☐ Nein, ich ändere meine Einschätzung nicht.

☐ Ja. Ich denke, dass mein Agent _____ Aufgaben richtig gelöst hat.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die erneute Selbsteinschätzung Ihres Agenten

Nachdem Ihr Agent die Anzahl richtiger Lösungen der Personen seiner Vergleichsgruppe erfuhr, hatte auch er die Möglichkeit, eine erneute Selbsteinschätzung abzugeben. Im Fall einer erneuten Selbsteinschätzung wurde nur die neue Selbsteinschätzung (wie oben beschrieben) entlohnt, die alte Angabe entfiel. Veränderte ein Agent seine Selbsteinschätzung nicht, wurde die alte Angabe wie gehabt entlohnt.

Sie werden gleich die erneute Selbsteinschätzung Ihres Agenten erfahren. Wenn Sie möchten, können Sie dann erneut einschätzen, was Sie glauben, wie viele Multiplikationsaufgaben Ihr Agent tatsächlich richtig gelöst hat. Wie gehabt, entfällt dann Ihre alte Einschätzung und nur die **neue Einschätzung** wird nach dem beschriebenen Schema **entlohnt**. **Verändern** Sie Ihre Einschätzung **nicht,** wird die **alte Einschätzung** wie gehabt **entlohnt**.

Füllen Sie nun bitte diesen Bogen aus:

Ihre erneute Einschätzung

Die **erneute Selbsteinschätzung** Ihres Agenten, nachdem er die Werte seiner Vergleichsgruppe erfuhr, lautet:

Wie Sie bereits wissen, lautet: - die erste Selbsteinschätzung Ihres Agenten:
- die Anzahl richtig gelöster Multiplikationsaufgaben der Personen der Vergleichsgruppe Ihres Agenten: **43, 52, 64, 81**

Möchten Sie Ihre Einschätzung darüber ändern, wie viele Aufgaben Ihr Agent im Experiment B tatsächlich richtig gelöst hat?:

□ Nein, ich ändere meine Einschätzung nicht, bleibe also bei _____ richtig gelösten Aufgaben.
□ Ja. Ich denke, dass mein Agent _____ Aufgaben richtig gelöst hat.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher										sehr sicher
	☐	☐	☐	☐	☐	☐	☐	☐	☐	
	-4	-3	-2	-1	0	1	2	3	4	

Bitte merken Sie sich für Stufe 2 Ihre hier angegebene Einschätzung.

Stufe 2: Wahl eines Mitarbeiters für Ihren Agenten und Einschätzung der Wahl der Vergleichspersonen

Auf dieser Stufe agieren Sie als **Manager eines Teams**. Sie werden **für Ihren Agenten** einen **Mitarbeiter auswählen**, der mit diesem zusammen ein Team bildet. Das Ergebnis dieses Teams wird mit vier anderen Teams, wie nachfolgend beschrieben, verglichen. Doch zunächst zu dem Mitarbeiter, den Sie für Ihren Agenten wählen können: Dieser ist keine reale Person, sondern er entstammt einer hypothetischen Menge von Mitarbeitern. Es wird angenommen, dass diese hypothetisch vorhandenen Mitarbeiter dieselben Multiplikationsaufgaben bearbeitet haben wie Ihr Agent und ein Ergebnis zwischen Null und (einschließlich) 110 richtigen Lösungen erreicht haben. Sie können also **einen** Mitarbeiter auswählen, dessen **Ergebnis** eine **ganze Zahl aus dem Intervall [0, 110]** ist.

Das **Teamergebnis des Teams, das Sie managen (Ihr Team),** ergibt sich aus der **Summe** der Anzahl **tatsächlich richtig gelöster Aufgaben Ihres Agenten** im *Experiment B* und dem **Ergebnis des hypothetischen Mitarbeiters**, den Sie auswählen. Die Mitarbeiterwahl für Ihren Agenten beeinflusst Ihre Auszahlung.

Auch die vier Personen der bereits erwähnten **Vergleichsgruppe** aus dem früheren *Experiment A* hatten jeweils einen Mitarbeiter mit einem Ergebnis aus dem Intervall [0, 110] ausgewählt. Hierdurch gelangten auch diese vier Personen zu bestimmten Teamergebnissen und Auszahlungen. Auch ihre Teamergebnisse addierten sich jeweils aus ihrer eigenen Anzahl richtig gelöster Multiplikationsaufgaben und dem Ergebnis desjenigen Mitarbeiters, den sie ausgewählt hatten. **Das Ergebnis Ihres Teams wird mit dem der vier anderen Teams der Vergleichsgruppe verglichen**.

<u>Überblick über die Teamzusammensetzungen:</u>

Ihr Team und die vier anderen Teams:

Ihr Agent	+ der von Ihnen für den Agenten gewählte hypothetische Mitarbeiter
erste Person der Vergleichsgruppe	+ der von dieser Person gewählte hypothetische Mitarbeiter
zweite Person der Vergleichsgruppe	+ der von dieser Person gewählte hypothetische Mitarbeiter
dritte Person der Vergleichsgruppe	+ der von dieser Person gewählte hypothetische Mitarbeiter
vierte Person der Vergleichsgruppe	+ der von dieser Person gewählte hypothetische Mitarbeiter

Die von den jeweiligen Personen der Vergleichsgruppe gewählten Mitarbeiter konnten gleiche oder verschiedene Ergebnisse haben.

Ihre Auszahlungen:

Je nachdem, welchen Mitarbeiter Sie für Ihren Agenten wählen, erhalten Sie unterschiedliche Auszahlungen:

- Wenn Sie für Ihren Agenten einen Mitarbeiter wählen, dessen Ergebnis **niedriger** als die **tatsächliche Anzahl** richtig gelöster Aufgaben Ihres Agenten aus *Experiment B* ist, erhalten Sie:

  5 € + die Hälfte vom Ergebnis des von Ihnen gewählten Mitarbeiters · 1 Cent

- Wenn Sie für Ihren Agenten einen Mitarbeiter wählen, dessen Ergebnis **höher oder genauso hoch** wie die **tatsächliche Anzahl** richtig gelöster Aufgaben Ihres Agenten aus *Experiment B* ist, erhalten Sie:

  2,50 € + die Hälfte vom Ergebnis des von Ihnen gewählten Mitarbeiters · 1 Cent

- Wenn das Ergebnis Ihres Teams das **beste oder zweitbeste** von den fünf Teamergebnissen ist, erhalten Sie **zusätzlich**

  7,50 €.

  (Sollte aufgrund mehrerer gleich hoher Teamergebnisse nicht eindeutig klar sein, ob Sie die 7,50 € erhalten oder nicht, wird gelost.)

Überblick über die Auszahlungen aufgrund der Mitarbeiterwahl für Ihren Agenten

Falls das Ergebnis des gewählten Mitarbeiters NIEDRIGER als die tatsächliche Anzahl richtig gelöster Aufgaben Ihres Agenten aus Experiment B ist:	**ODER**	**Falls das Ergebnis des gewählten Mitarbeiters HÖHER oder GENAUSO HOCH wie die tatsächliche Anzahl richtig gelöster Aufgaben Ihres Agenten aus Experiment B ist:**
5 € + die Hälfte vom Ergebnis des gewählten Mitarbeiters · 1 Cent		**2,50 €** + die Hälfte vom Ergebnis des gewählten Mitarbeiters · 1 Cent
zusätzlich 7,50 €, falls Ihr Teamergebnis zu den zwei höchsten Teamergebnissen gehört		**zusätzlich 7,50 €,** falls Ihr Teamergebnis zu den zwei höchsten Teamergebnissen gehört

Einschätzung über die Mitarbeiterwahl der Vergleichsgruppe:

Bevor Sie selbst einen Mitarbeiter für Ihren Agenten auswählen, **schätzen Sie bitte ein**, welche Mitarbeiter die vier Personen der **Vergleichsgruppe** jeweils gewählt haben. Auch diese wurden nach demselben Schema entlohnt, haben also **5 €** + *die Hälfte vom Ergebnis des gewählten Mitarbeiters · 1 Cent* erhalten, falls sie einen *schlechteren Mitarbeiter als sich selbst* gewählt haben, und **2,50 €** + *die Hälfte vom Ergebnis des gewählten Mitarbeiters · 1 Cent*, falls sie einen *besseren oder einen genauso guten Mitarbeiter* gewählt haben.

Diejenigen Personen der Vergleichsgruppe erhielten eine Zusatzzahlung von **7,50 €**, falls deren Teamergebnis zu den besten **zwei** der **vier** Teamergebnisse dieser Gruppe gehörte.

Beachten Sie, dass die vier Personen der Vergleichsgruppe **ihre eigene Anzahl richtig gelöster** Multiplikationsaufgaben bereits vor dem Zeitpunkt ihrer Mitarbeiterwahl **gekannt haben**. Ebenso kannten sie die Anzahl richtig gelöster Aufgaben der anderen drei Personen ihrer Gruppe. So erhielt z.B. die Person X aus der Vergleichsgruppe, die 52 Aufgaben richtig gelöst hat, die folgende Information:

Codename: X
Sie haben folgende Anzahl an Aufgaben richtig gelöst: 52

Die Personen Ihrer Gruppe haben folgende Anzahl an Aufgaben richtig gelöst: 43, 64, 81

Die Personen der Vergleichsgruppe mussten einschätzen, welchen Mitarbeiter die anderen Personen dieser Gruppe wählen würden. So musste z.B. die Person X die Mitarbeiterwahl der Personen mit 43, 64 und 81 richtig gelösten Aufgaben einschätzen. Pro richtiger Einschätzung der Mitarbeiterwahl erhielt eine Person 50 Cent, für eine falsche Einschätzung erhielt sie keine Auszahlung. Anschließend hat jede Person der Vergleichsgruppe selbst einen Mitarbeiter aus dem Intervall [0,110] ausgewählt.

Nun wieder zu dem heutigen Experiment: Sie werden **einschätzen**, welchen Mitarbeiter die Personen der **Vergleichsgruppe** gewählt haben. Hierzu werden wir Ihnen einen Bogen austeilen, auf dem die Ergebnisse der Vergleichspersonen bei den Multiplikationsaufgaben aufgelistet sind. Für jede **richtige** Einschätzung erhalten Sie **50 Cent**, für eine **falsche** Einschätzung erhalten Sie **keine Auszahlung**. Außerdem werden Sie gefragt, wie **sicher** Sie sich bei Ihren Einschätzungen sind. **Anschließend werden Sie selbst für Ihren Agenten einen Mitarbeiter wählen.**

Einschätzung der Mitarbeiterwahl der Vergleichsgruppe

und Ihre Wahl für Ihren Agenten

Was glauben Sie, welcher hypothetische Mitarbeiter mit einem Ergebnis aus [0, 110] jeweils von den Personen der Vergleichsgruppe gewählt wurde?

Die Person mit **43** richtigen Lösungen aus der Vergleichsgruppe hat einen hypothetischen Mitarbeiter mit dem **Ergebnis** _____ gewählt.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die Person mit **52** richtigen Lösungen aus der Vergleichsgruppe hat einen hypothetischen Mitarbeiter mit dem **Ergebnis** _____ gewählt.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die Person mit **64** richtigen Lösungen aus der Vergleichsgruppe hat einen hypothetischen Mitarbeiter mit dem **Ergebnis** _____ gewählt.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Die Person mit **81** richtigen Lösungen aus der Vergleichsgruppe hat einen hypothetischen Mitarbeiter mit dem **Ergebnis** _____ gewählt.

Kreuzen Sie bitte an, wie sicher Sie sich bei Ihrer Einschätzung sind:

nicht sehr sicher	☐	☐	☐	☐	☐	☐	☐	☐	☐	sehr sicher
	-4	-3	-2	-1	0	1	2	3	4	

Welchen Mitarbeiter mit einem Ergebnis aus dem Intervall [0, 110] wählen Sie für Ihren Agenten?

Ich wähle für meinen Agenten einen Mitarbeiter, der ein **Ergebnis von** _____ hat.

Zur grafischen Veranschaulichung des Zusammenhangs von den Experimenten A und B mit dem heutigen Experiment:

Experiment A		**Experiment B**		**heutiges Experiment**
Die vier Personen der Vergleichsgruppe bearbeiteten die Multiplikationsaufgaben. Sie erfuhren ihre eigene Anzahl richtig gelöster Aufgaben und die der anderen drei Personen der Vergleichsgruppe.	→	Ihr Agent schätzte ein, wie viele Multiplikationsaufgaben er richtig lösen würde, und bearbeitete später die Aufgaben.	→	Sie schätzen ein, wie viele Aufgaben Ihr Agent tatsächlich richtig gelöst hat.
→ danach wählte jeder einen hypothetischen Mitarbeiter aus dem Intervall [0,110].				→ wählen für Ihren Agenten einen hypothetischen Mitarbeiter aus dem Intervall [0,110].
→ somit entstanden vier Teams, die sich aus einer Person der Vergleichsgruppe und dem jeweils gewählten Mitarbeiter zusammensetzten.				→ Ihr Agent und der ausgewählte hypothetische Mitarbeiter bilden ein Team.

Diese Teams werden miteinander verglichen.

Chapter 4 [78]

The Effects of Feedback on Self-Assessment

4.1 Introduction

The self-assessment of individuals is often wrong. In particular, numerous empirical studies and experiments confirm that overconfidence is a widespread feature. For example, people overestimate their relative driving abilities or their chances of obtaining a preferred job (see, e.g., Svenson 1981, Weinstein 1980). In addition, people perceive themselves more favorably than other people perceive them and they show an illusion of control (see, e.g., Taylor and Brown 1988, Weinstein 1980).

Such behavioral patterns have far reaching implications in economic settings: Overconfidence can lead to excessive market entry (see Camerer and Lovallo 1999). Standard incentive contracts do not work for overconfident managers (see, e.g., De la Rosa 2005). Moreover, overconfident people behave differently in Bertrand or Cournot competition, or choose different investment strategies than unbiased individuals (see, e.g., Eichberger et al. 2005). Also in contests the behavior of overconfident players differs from the one of unbiased players (see, e.g., Ando 2004). In auctions, where many people compete for an object with a common but uncertain value, it is a well-known phenomenon that bidders *overpay* when they win the object, i.e. they are prone to the winner's curse.[79] People who are overconfident are even more prone to the winner's curse because they *systematically* overestimate the value of the object or their ability to assess the value of the object correctly. Overconfidence can also lead to wrong financial decisions when, for example, investors overestimate the precision of private relative to public signals (see Barberis and Thaler 2003 for an overview). One prominent example in this context is the "hubris hypothesis" by Roll (1986) which can be interpreted as an extreme kind of the winners curse.[80]

---

[78] This chapter is based on Eberlein, Ludwig, and Nafziger (2008).

[79] For an overview on the winner's curse see Thaler (1988). See Biais et al. (2005) for a related experiment on financial markets.

[80] There are also studies that show that people are underconfident in certain situations. Such findings are much less discussed in literature. Underconfidence is almost only discussed in the context of the so called "hard easy effect". Several authors find that overconfidence is most extreme at very difficult tasks and is

Most experiments investigate overconfidence in one shot situations, while real life situations are characterized by repeated actions where individuals can learn about the tasks and their ability to solve these tasks over time. What are the effects of these learning opportunities? Does a wrong self-assessment prevail over time if learning is possible? Why does it (not) prevail?

I want to tackle these questions in my experiment. First, I ask whether *feedback* reduces overconfidence and whether it improves an individual's self-assessment. For instance, people who have never solved a specific task before, might have a too high prior about their ability in solving this task. When they, however, solve the task repeatedly and receive feedback, they could update their prior and learn about their true ability. The consequence of such an updating process would be that overconfidence vanishes.

It is not only important to know *whether* information influences people's (biased) self-assessment. The second aim of my experiment is to analyze *how* and *when* information influences self-assessments: Do individuals use information in the right way? Is adjusting one's self-assessment to feedback always the best thing to do or could ignorance sometimes be better? How do different kinds of feedback influence subjects' reactions?

The design of my experiment is roughly as follows. First, subjects answer 90 multiple-choice questions, which are divided into six blocks (15 questions per block). After each block of questions, subjects have to assess their ability in the just accomplished block relative to a reference group.[81] A subject's relative ability is measured by his ability (i.e. the number of correct answers in the considered block of questions) being higher or lower than the median ability of the reference group in the same block.

In each of the six rounds of questions and self-assessments (except for the first round), subjects receive feedback about their achievement in the preceding round. More precisely, before making the self-assessment regarding the current block, subjects are told whether they were better or worse than the median in answering the questions of the preceding block. Furthermore, they are informed about the number of questions they answered

---

reduced when tasks get easier (e.g. Clarke 1960, Lichtenstein and Fischhoff 1977, Gigerenzer et al. 1991). Indeed, subjects responding to very easy tasks are often underconfident.

[81] The reference group consists of subjects in another experiment who answered the same multiple-choice questions.

correctly in the preceding block. After having answered all the questions and after the self-assessment in the last block of questions, subjects also assess their "total ability" over all blocks relative to the reference group.

My data show that the reduction of wrong self-assessments after five rounds of feedback is just moderate. To examine this result in more detail, I investigate the influence of different kinds of feedback. I distinguish between robust/ varying feedback, between good/ bad news and between confirming/ non-confirming feedback. First, I observe that subjects, who almost always get to know to belong to the better respectively worse 50% (robust feedback) are less often wrong with their self-assessments than subjects who receive varying feedback. Second, feedback messaging the good news to belong to the better 50% is followed more frequently than when bad news (to belong to the worse 50%) is messaged. Interestingly, subjects make more mistakes when not following bad news than when not following good news. Third, following confirming feedback (i.e. the inherent message that one's preceding assessment is right) leads to a decrease in mistakes over rounds. Surprisingly, non-confirming feedback shows no appropriate reaction. Contrarily, the number of mistakes increases over rounds. Such feedback might lead to confusion and thus to mistakes.

Related Literature

Psychologists study the effects of feedback on overconfidence in studies where subjects have to perform so called calibration tasks (e.g. stating confidence intervals for their answers). These studies yield ambiguous results: While Pulford and Colman (1997) and Sharp et al. (1988), for example, find that feedback mostly has no effect on overconfidence, Adams and Adams (1961) and Lichtenstein and Fischhoff (1977) find that feedback reduces overconfidence (or improves the calibration). These studies provide subjects with different forms of feedback: Pulford and Colman give outcome feedback (the correct answers to the posed questions), whereas the others give performance feedback (information about realism of confidence) or statistical feedback (information about calibration).

My experiment differs considerably from these psychological studies: First, I provide subjects with precise feedback about their absolute and relative ability. This can help to avoid that subjects confound a high (low) absolute ability with being better (worse) than

the median. Second, I introduce performance related payments. Third, subjects in my experiment do not have to carry out a calibration task, but simply estimate their relative ability: Compared to assessing one's (relative) ability, a calibration task is quite difficult and may be a reason why feedback does not necessarily help if subjects are not familiar enough with it. Furthermore, assessment of one's *relative* ability (compared to absolute ability which most psychologists consider) seems most relevant for economic settings: For example, in a tournament not the absolute ability matters but the relative abilities of a player and his opponent(s). Fourth, I use a neutral wording in the instructions, i.e. I do not explicitly ask subjects whether they think they are better than the median. Instead, subjects have to make a choice between two alternatives which is a more neutral and indirect way to find out their self-assessment as the wording does not influence people in their self-assessments.

In economics, the field experiment by Ferraro (2005) is most closely related to my experiment. He tests whether people assess their own ability correctly or rather overestimate it. In his experiment, students are asked after an exam to estimate how many questions they answered correctly in the exam and what they think their rank is. This procedure is repeated in two subsequent exams. Feedback in the form of a student's grades and the distribution of grades in the first and second exam, respectively, shows little impact on the accuracy of the self-estimation and overconfidence. Contrarily to my study, subjects cannot perfectly deduce their rank (i.e. their relative ability) from the information, which is thus an additional source of uncertainty. Moreover, since some time passes between the exams, subjects have the opportunity to prepare better or worse for the next exam. This preparation may influence their self-assessment and cannot be disentangled from the effects of feedback.

In contrast to my study, some experiments investigate whether feedback improves subjects' decisions by learning when it is not a wrong self-assessment that leads to mistakes, but reasoning errors. One example is the study by Budescu and Maciejovsky (2005) who analyze whether reasoning errors prevail even in a competitive situation. They find that payoff feedback improves the decisions of subjects.

This chapter is structured as follows. In section 4.2, I describe the experimental design. In section 4.3, I present and discuss the results. In the last section, I sum up and conclude.

4.2 Procedure and Experimental Design

The computerized experiment was conducted at the University of Bonn and programmed with z-Tree (Fischbacher, 1999). Subjects were recruited via the internet by using ORSEE software (Greiner, 2004) announcing the possibility to earn an amount of money dependent on their behavior. During the experiment, subjects earned Talers, which were converted into Euros in the end (210 Talers = 1 €). Average hourly earnings were 10 €. The instructions were read out loudly before the experiment started. Subjects also answered control questions to make sure that they understood the experimental procedure.

In *Experiment F(eedback)*, 15 subjects answer 90 general knowledge questions (multiple-choice, 4 answer possibilities), which are divided into six blocks (15 questions per block). I choose the specific questions of each block very carefully to make sure that a subject's number of correct answers does not considerably vary over blocks.[82]

After each block of questions, subjects have to state whether they think that they belong to the better or worse 50% in this block, compared to the subjects in the reference experiment (*Experiment Q(uestions)*) which was conducted prior to *Experiment F*.[83] Subjects know that in *Experiment Q*, 15 other subjects answered the same questions, divided into the same six blocks, and were given the same incentives to answer a question correctly as in *Experiment F*.[84] The better or worse 50% are determined as follows: Subjects are ranked according to their number of correctly answered questions, a higher number implying a higher rank. If two or more subjects have the same number correct, the subject who answered the correct questions faster receives the higher rank. The worse 50% (of the 16 individuals composed of the considered subject and the 15-person-reference-group) are the subjects with ranks 1 to 8, the better 50% the ones with ranks 9 to 16. Thus, the relative ability of an individual depends on whether it belongs to the better or worse 50%. Note that

---

[82] In each block there is one question of each of the following 15 topics: biota, history, geography, astronomy, sports, literature, chemistry, music, biology, arts, physics, movies, maths, religion, and politics. Moreover, the level of difficulty is the same in all blocks (I controlled for this already before my experiment by conducting a pre-test).

[83] Note that there is no strategic interaction between subjects in *Experiment F*: There is no influence of the abilities or self-assessments of a subject in *Experiment F* on the payoffs of another subject in *Experiment F*. Hence, I have 15 independent observations.

[84] In *Experiment Q*, subjects stated after each block, how many questions they think they answered correctly in this block. They received small monetary incentives to make a correct statement. In this experiment, subjects do not get any feedback. Subjects show severe difficulties with correctly assessing their absolute ability. Biases do not differ significantly across subsequent rounds or between the first and the last round according to two-sided Wilcoxon-Signed-Rank tests. Thus, there is no "learning" effect just by completing the task several times when subjects receive no information in between.

I avoid phrases like "the better 50%" during the whole experiment, but call the better 50% (worse 50%) group A (group B).

After answering the second block of questions but before the self-assessment in round 2, subjects receive the information to which group (A or B) they belonged to - i.e. their relative ability - and how many questions they answered correctly - i.e. their absolute ability - in the first round. I apply the same procedure for the consecutive rounds. By telling the subjects how many questions they answered correctly, I avoid an additional source of uncertainty. After their self-assessment in round 6, subjects finally assess to which group they belong to when *all* the questions of *all* blocks are considered.

The payoffs for the answers and self-assessments are as follows. I make two independent random draws out of the six rounds: one draw that determines which of the six blocks of questions is rewarded in the end of the experiment and another draw that determines which of the six block-wise self-assessments is rewarded. During the experiment, subjects do not know which rounds will determine their payment. This motivates subjects during the experiment. Without this procedure, subjects might not exert much effort in later rounds when they think to have already earned enough money in the experiment and the payments for further correct answers or self-assessments are relatively low.

For each correctly answered multiple-choice question in the chosen block, subjects earn 270 Talers (roughly 1.30 €) minus 0.9 Talers for every second a subject needs for his correct answer.[85] The second randomly drawn round determines the payment for the assessment task: A subject receives 1500 Talers (roughly 7.15 €) if he has placed himself in the correct group A or B, respectively, in this round, otherwise he gets 180 Talers (roughly 0.85 €). Moreover, subjects receive 300 Talers for a correct total assessment after the last round and 20 for a wrong one.[86] Finally, subjects get 2.50 € for showing up. Subjects can earn a lot when their self-assessment is correct, but I want to ensure that subjects with a wrong assessment still earn a reasonable amount of money. To keep the

[85] Because I need the time as a tie-breaking rule, I set some small monetary incentives to be fast.
[86] Note that, in principle, subjects could use their assessments for hedging: In not answering any question (which counts as a wrong answer in each case) and stating to belong to group B in the block-wise assessments and in the final assessment, subjects may want to assure earning 1800 Talers. Hedging will, however, not be a driving force for their behavior: When neglecting the time that a subject needs to answer a question correctly, he gets a payoff of 1820 Talers when correctly answering only 6 of the 15 questions in the chosen block and when he makes a wrong assessment in the chosen block and a wrong final assessment.

show-up fee rather low, subjects receive 200 Talers (180 plus 20 Talers) for sure in the assessment tasks if their assessments are wrong.[87]

4.3 Results

In this section, I investigate the subjects' general and individual reactions to feedback as well as their reactions to robust versus varying feedback, to good news versus bad news, and to confirming versus non-confirming feedback. Moreover, I examine the assessment of their relative total ability.

In *Experiment F* the average ability is 6.6.[88] Abilities do not differ significantly over subsequent rounds or between the first and the last round according to two-sided Wilcoxon-Signed-Rank tests.[89] [90]

Considering the relative self-assessments, I observe that in all rounds many subjects are wrong (see below for more details). Thus, subjects have severe problems with their relative self-assessments. Out of all *wrong self-assessments*, in 61% one's relative ability is overestimated, which means that a subject states to belong to the better group although he actually does not.[91]

Reactions to Feedback and Consequences

As a starting point I observe that roughly half of the self-assessments (46.67%) are wrong in the first round, while a third are wrong in the last round. There is, however, no

---

[87] The instructions of *Experiment F* can be found in the appendix (section 4.5).

[88] According to two-sided Mann-Whitney U tests there is no significant difference between abilities when comparing *Experiment F* with *Experiment Q*: p=0.976 when comparing average abilities over blocks of both experiments, p=0.146/0.472/0.571/0.880/0.475/0.819 when comparing single blocks of both experiments.

[89] When comparing subsequent blocks p=0.584/1/0.241/0.287/0.191, when comparing the first with the last block p=0.754.

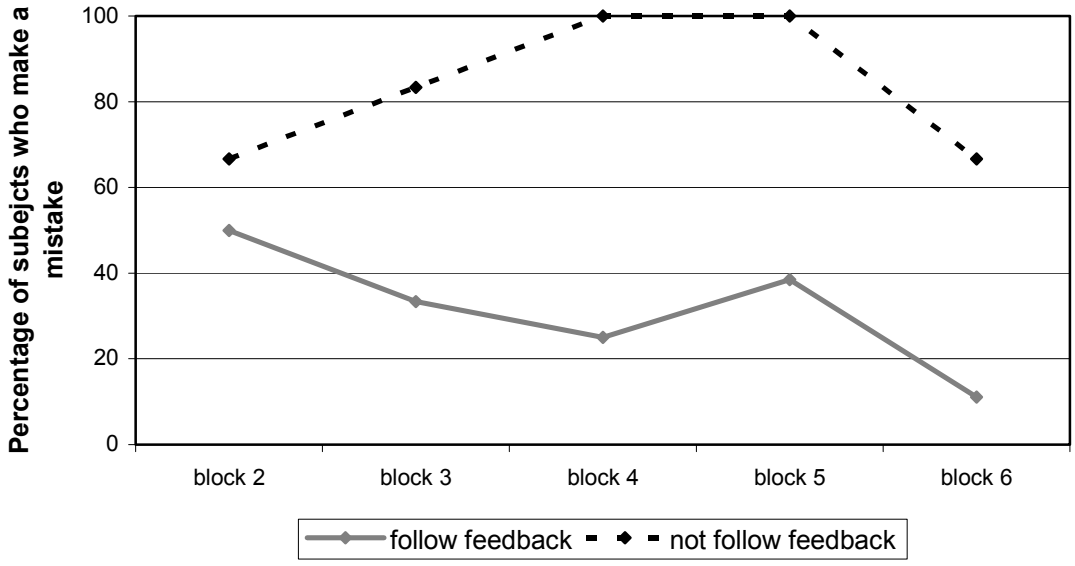
[90] I think that the number of correctly solved questions is mainly influenced by a subject's ability in my experiment. This is why I refer to the number of correctly solved questions as the subject's absolute ability. Clearly, a subject's effort does not play a crucial role when solving multiple-choice general knowledge questions. If a subject does not know the answer to a question, it does not help to exert much effort. But of course I motivate subjects to attentively think about their answers because of the implemented incentive scheme. Of course I cannot avoid that a subject does not know the answer to a certain question but makes a guess and has the luck that this guess is right. But as the number of correctly solved questions does not differ significantly across blocks due to Wilcoxon-Signed-Rank tests, this hints at the fact that the number of correctly answered questions indeed mainly mirrors a subject's ability.

[91] In some experiments (see, e.g., Maccoby and Jacklin 1974, Pulford and Colman 1997) it is observed that more men than women are overconfident. However, this is not the case in my experiment. Testing with a Fisher exact test in each round whether more men overestimate their relative ability, I find no significant difference.

significant improvement between the first and the last self-assessments according to a one-sided McNemar test (p=0.363).

How does feedback drive this result? To investigate this in more detail, I analyze whether reacting to feedback improves self-assessments, i.e. whether subjects who act in line with their feedback make fewer mistakes. To do so, I consider the self-assessments regarding blocks 2 to 6, i.e. the self-assessments where feedback is available. Out of all self-assessments that are in line with feedback on average 32.73% are wrong, whereas out of all self-assessments that are *not* in line with feedback on average 80% are wrong. Figure 4.1 illustrates the development over time: The mistake rate for subjects following feedback has a decreasing tendency. Contrarily, if subjects do not follow their feedback the percentage of subjects making a mistake increases over rounds (except for the last round). Furthermore, I find a negative correlation between the number of mistakes and the frequency with which subjects follow their feedback when assessing their relative ability (Spearman Correlation, rho=-0.763, p=0.001, two-sided). Taken together these results indicate that following feedback generally improves self-assessments and thus feedback accounts for the decrease in mistakes over rounds.

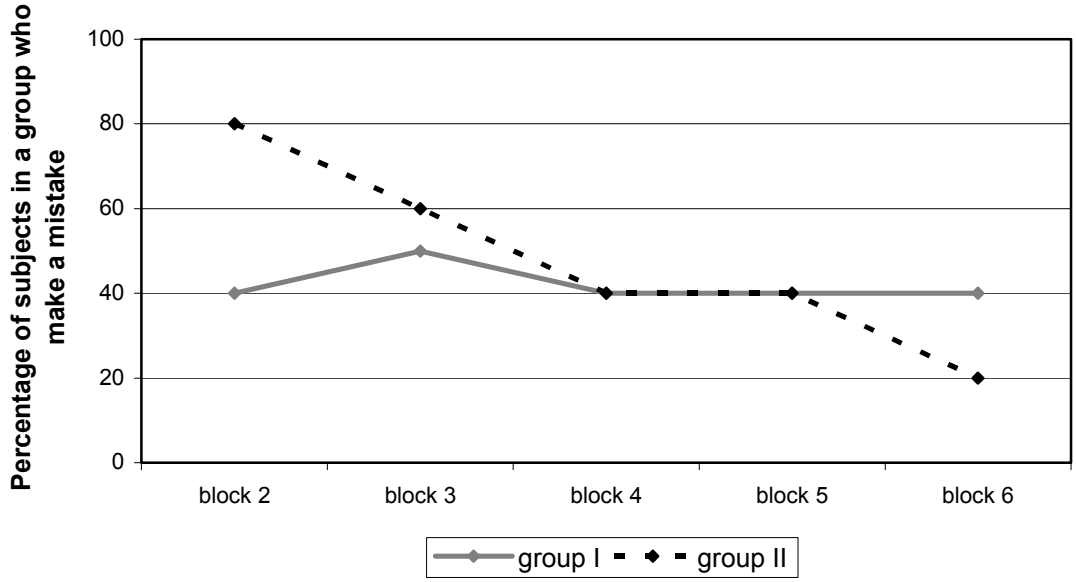
Figure 4.1: Mistakes in relative self-assessments by reactions to feedback over blocks



When investigating the data in more detail, two different kinds of noticeable behavior can be detected: there are subjects who change their belief (to belong to the better or worse 50%) rather frequently, while others rather stick to one belief. To understand the driving

forces and consequences of such differing behavior, I investigate those groups of subjects in detail. With group I, I refer to those subjects who change their belief never or only once over the five self-assessments with feedback (66.67% of subjects); with group II, I refer to those subjects who change it more often than once (33.33% of all subjects). On average, the percentage of wrong self-assessments in the five self-assessments with feedback is very similar across these groups: 42% of assessments are wrong in group I, 48% in group II. When considering the rounds separately, however, I see in Figure 4.2 that the number of mistakes in group II is actually decreasing over rounds while it is nearly constant in group I.

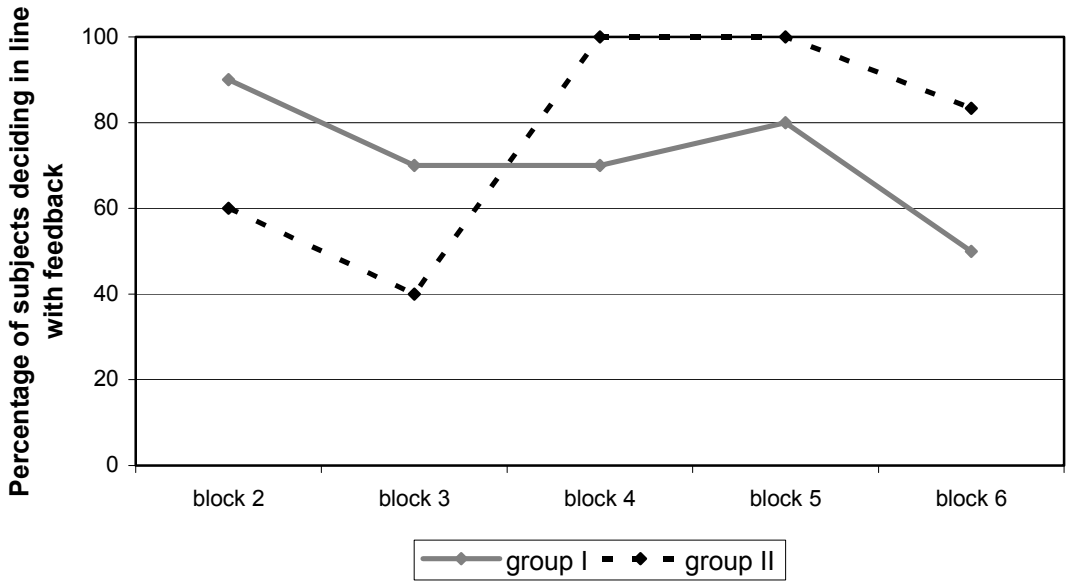
Figure 4.2: Mistakes in relative self-assessments by reactions to feedback over blocks by group



To understand whether feedback is the driving force behind the decrease in mistakes of subjects in group II, I ask which group decides more often in line with feedback. Altogether, a majority of subjects chooses their actions in line with their feedback about the previous round (73.33%). In group II, however, the percentage of subjects deciding in line with feedback is a bit higher than in group I and it shows an increasing tendency (see Figure 4.3). In group I, the percentage stays relatively constant over time. This finding can be interpreted in the following way: Subjects who change their belief often seem to learn from their feedback and make fewer and fewer mistakes after having received more and more information. Subjects who do not change their belief often, ignore feedback slightly more often than subjects in group II which does not lead to an improvement in their self-

assessments. In the end, group II, which had more problems with the self-assessment in the beginning, outperforms group I.

Figure 4.3: Subjects deciding in line with feedback in groups I and II

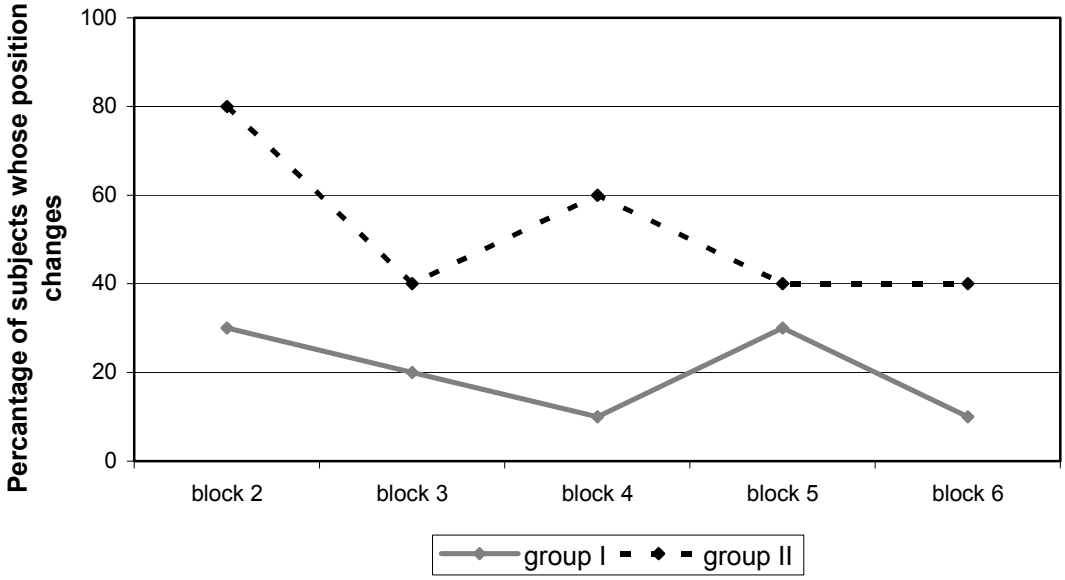


Next, I consider whether acting in line with one's feedback or not has different effects on making wrong or right self-assessments on the subjects in groups I and II. I observe that the average percentage of wrong self-assessments for subjects in group II is much larger than for group I when subjects decide *in line with their feedback* (47.37% versus 25%). When subjects, however, do *not decide in line with their feedback*, it is vice versa: While the percentage of wrong self-assessments in both groups rises, it does so *quite dramatically* for the group with beliefs that rarely change (92.86 % in group I, 50% in group II).

I see in Figure 4.3 that in blocks 2 and 3 less subjects in group II follow their feedback than in group I. This relationship reverses for the next three blocks. What causes this remarkable change between the first two (blocks 2 and 3) and the last three blocks for group II? As displayed in Figure 4.4, the relative position (being above or below the median) changes for 80% of subjects in this group when comparing their relative ability in blocks 1 and 2. This means, in one of the blocks they belong to the better 50%, in the other to the worse 50%. Hence, their first two feedbacks – the ones in blocks 2 and 3 – differ. Hence, those subjects might get confused. From block 4 on, subjects in group II seem to perceive correctly that they have problems with their self-assessment and try to improve their assessment by following the feedback in later rounds. This strategy proves successful:

The mistake rate in group II decreases continuously and is in the last round even lower than that of group I (compare Figure 4.2).

Figure 4.4: Change of position by group



Robust vs. Varying Feedback

As it is already seen, the feedback a subject receives is not necessarily robust over rounds, but can change. Could the robustness of feedback drive the results? Does varying feedback cause subjects to make more mistakes and to change their beliefs often?

I define feedback as robust when it changes at most once across rounds, i.e. subjects almost always get to know that they belong to the better respectively worse 50% of subjects in the question task regarding the reference group, with at most one exception.[92] Dividing subjects into those who get robust feedback (53.33%) and those who get varying feedback (46.67%), I can compare their number of mistakes. Of those subjects who have robust feedback, only 27.5% of assessments are wrong in blocks 2 to 6. Of those who have varying feedback, 65.71% of the assessments are wrong in these five blocks. When I account for this fact and look at subjects with robust feedback separately, I find that - in each round - they represent on average 70.73% of those who make a correct self-assessment. Interestingly, in the first block (where subjects have not yet received any feedback), these subjects represent only 37.5% of those being correct. A one-sided

[92] Changes happen particularly for those subjects whose ability is close to the median ability.

McNemar test shows that there is a significant improvement in self-assessments from the first to the last block for subjects with robust feedback (p=0.063).

To better understand the subjects' behavior, I again consider groups I and II: 60% of the subjects in group II and 40% of subjects in group I receive varying feedback. When analyzing the effect of varying feedback on wrong self-assessments in groups I and II, I find that 46.67% of self-assessments with varying feedback are wrong in group I while 60% of self-assessments with varying feedback are wrong in group II. The percentages of wrong self-assessments stay rather constant over blocks for both groups. When considering robust feedback, the percentage of wrong self-assessments is small (26.67%) and constant for subjects in group I. Concerning subjects with robust feedback in group II, 30% of self-assessments are wrong. Remarkably mistakes decrease over rounds for these subjects (except for block 3) until there is no wrong self-assessment in the last block. Thus, when subjects vary their belief often and feedback is robust, relative self-assessments strongly improve over time. Remember that I observed in Figure 4.2 that the number of mistakes in group II is decreasing over rounds. This result is hence driven by the improvement of relative self-assessments of those subjects who get rather robust feedback.

Figure 4.5: Subjects making a mistake after a change or no change of position

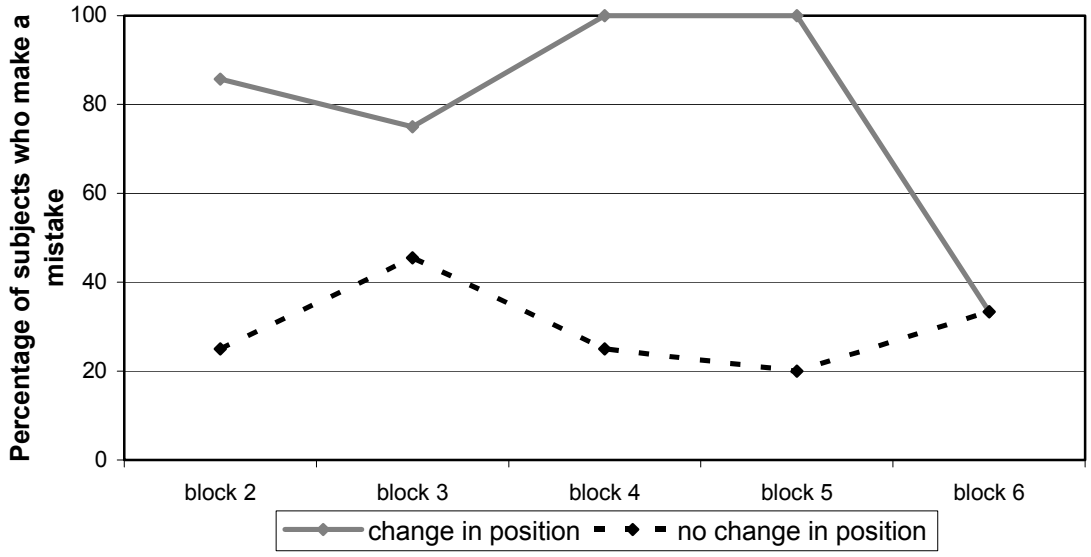


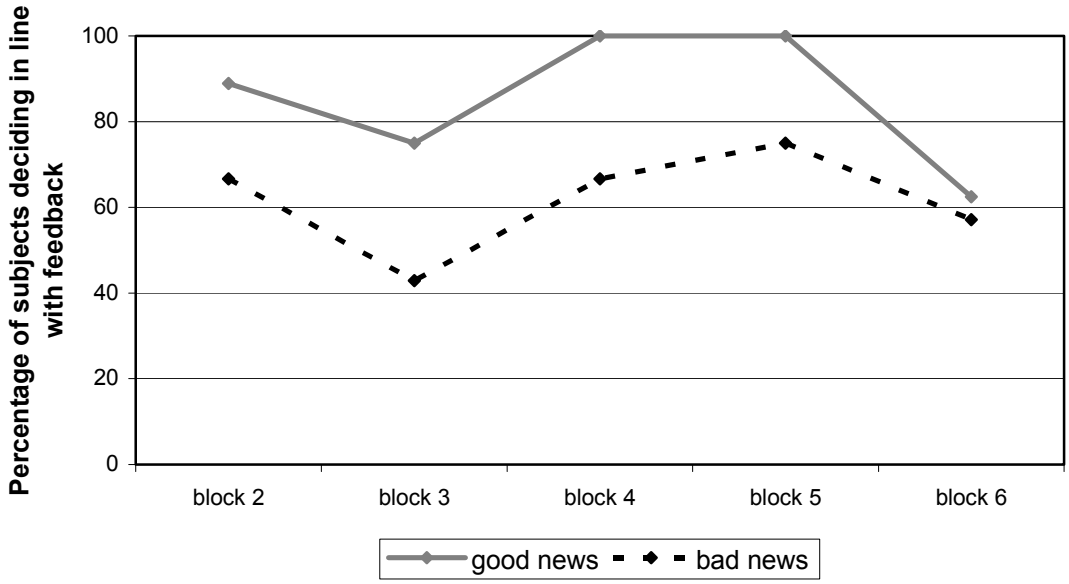
Figure 4.5 helps to understand why varying feedback can cause wrong self-assessments: It illustrates that the self-assessment is usually wrong in a round where a subject's position changes, in contrast to a round where it does not change. One can see that subjects whose

relative ability changes from the previous round to the current round ("change of position") make more often mistakes than subjects whose positions do not change. Hence, for a *single event* subjects do not perceive that their position changes and therefore feedback does not directly help in such a situation. If such changes occur *repeatedly*, there is, however, a learning effect, as it is indicated by the sharp drop in mistakes from block 5 to 6.

Good News vs. Bad News

Do subjects react differently to "good news" (i.e. feedback that indicates to the subject that he belongs to the better 50%) and "bad news" (i.e. feedback that indicates to the subject that he belongs to the worse 50%)? Dividing subjects into those who receive good news and those who receive bad news[93], I observe that subjects are more likely to follow good news (i.e. they decide in line with their "good" feedback), which is shown in Figure 4.6.

Figure 4.6: Following good news/ bad news



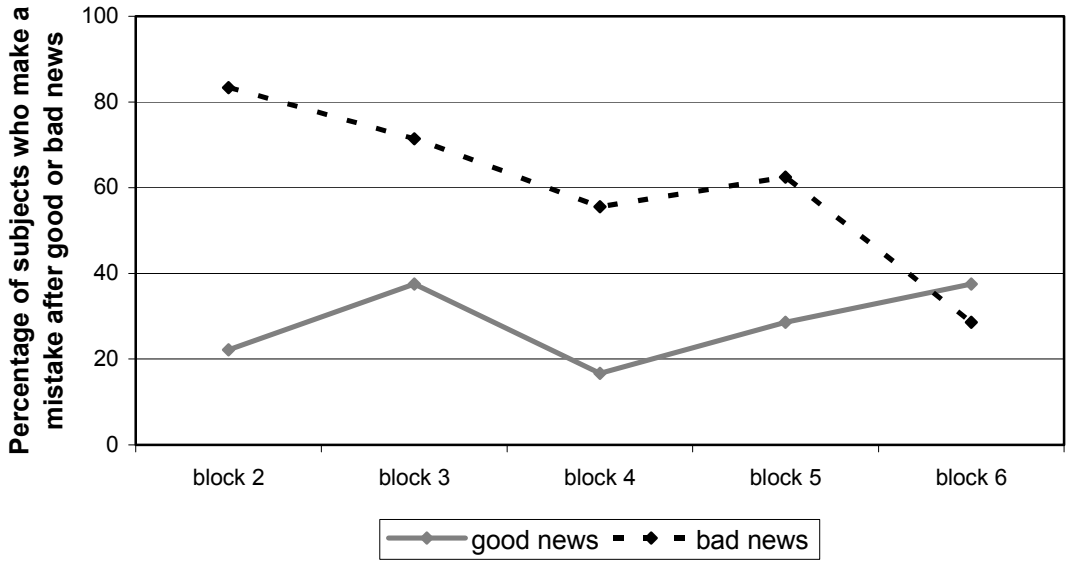
Nevertheless, out of the group that receives bad news still 49.33% follow this feedback on average. Thus, there is no evidence in my experiment that subjects generally ignore bad news, contrarily to what is claimed by psychologists.[94]

[93] The two groups of subjects that either receive good or bad news in a block are always similar in size. On average, 50.67% of subjects receive good news in each block.

[94] In psychological studies it is, for example, observed that for most subjects positive personality information is efficiently processed and easily recalled, whereas negative personality information is not processed or only poorly and is difficult to recall (see, for example, Kuiper and Derry 1992, and Kuiper and Mc Donald 1982). Moreover, it is found that most individuals attribute positive rather than negative outcomes for the self (see, for example, Bradley 1978, Miller and Ross 1975).

Furthermore, I find that the percentage of subjects making mistakes out of those who received bad news is, on average, larger than out of those who received good news (60.28% vs. 28.49%). Remarkably, mistakes after receiving bad news decrease while mistakes after good news stay rather constant, as can be seen in Figure 4.7. Note that these results are independent of whether subjects act in line with their feedback or not. The results are, however, much more pronounced when subjects do not follow their feedback. In this case, on average 23.33% of those subjects, who get good news and do not follow, make a mistake. Whereas on average 93.33% of those subjects, who get bad news and do not follow, make a mistake.

Figure 4.7: Mistakes after good news/ bad news

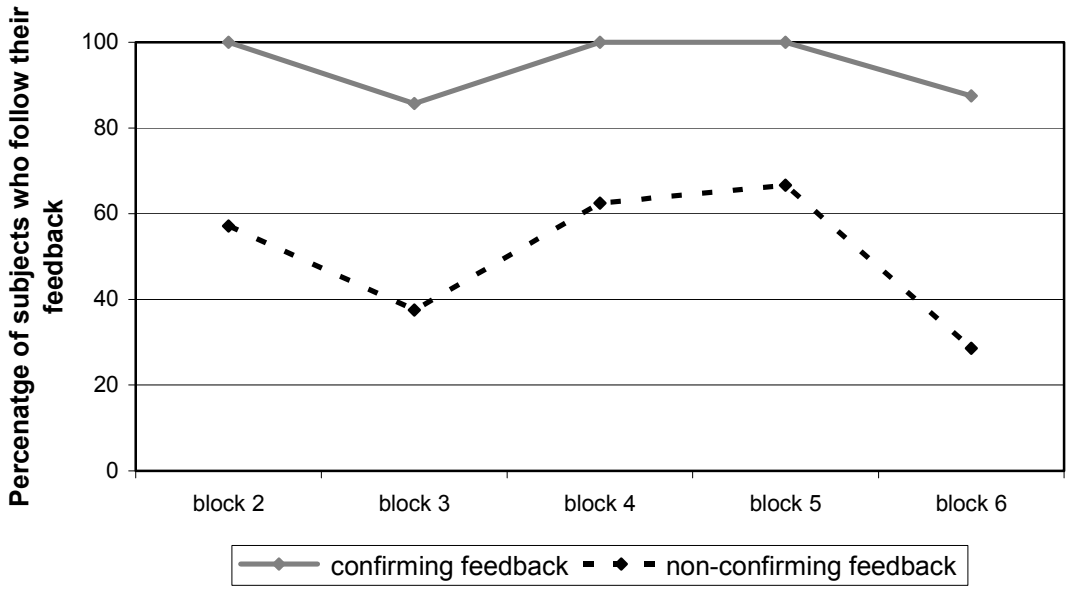


Confirming versus Non-Confirming Feedback

Confirming feedback – in the sense that a subject gets to know that he has been right with his self-assessment in the preceding round – can also be seen as good news, because the subject expects to receive the high payoff with at least a probability of 1/6. A subject receives bad news when he gets to know that his self-assessment in the preceding round was wrong (i.e. non-confirming feedback).

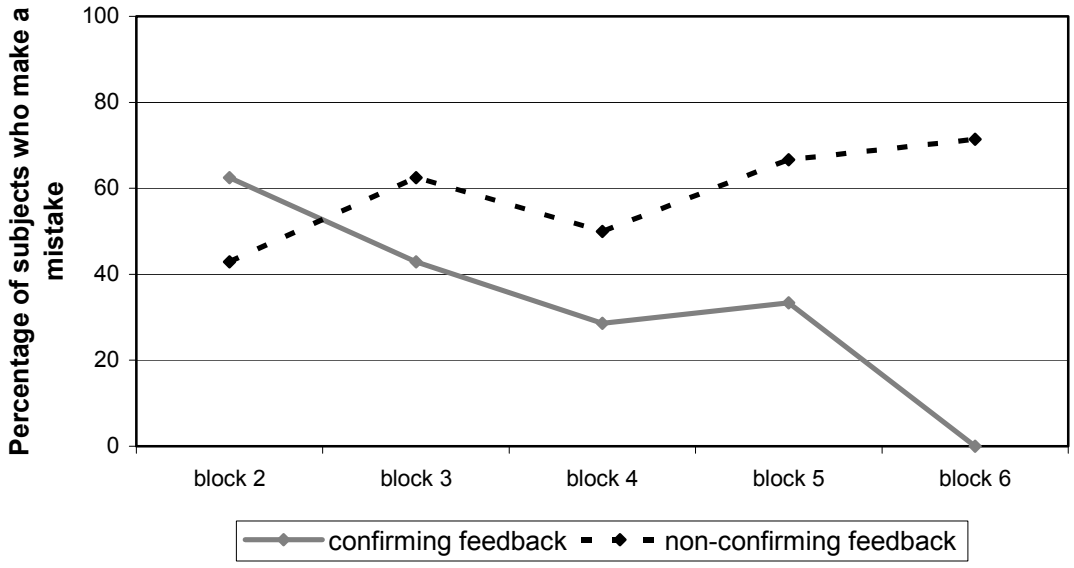
Figure 4.8 shows the percentage of subjects who follow confirming/ non-confirming feedback and Figure 4.9 shows the percentage of subjects who make a mistake after

Figure 4.8: Following confirming/ non-confirming feedback



receiving confirming / non-confirming feedback. As can be seen, the results in qualitative terms are the same as for good and bad news above: Subjects follow confirming feedback more than non-confirming feedback and mistakes are more frequent after non-confirming feedback than after confirming feedback. But remarkably, the frequency of mistakes decreases considerably after confirming feedback and increases after non-confirming feedback, both in sharp contrast to the reactions to the kind of good/ bad news discussed above.

Figure 4.9: Mistakes after confirming/ non-confirming feedback



Final Self-Assessment

For the total assessment in the end, almost every subject processes his information well. This task is relatively easy as subjects can use all feedback that they received previously. Here, only four subjects are wrong. Two of them got three times a good feedback and three times a bad one - hence their feedbacks do not deliver a clear advise for their total assessment. One of the four subjects who are wrong always said he belongs to the better 50%, although he was always told he does not. This is the only subject in my experiment who seemed to ignore all his failures. Remarkably, ignoring every failure *did not* occur more often, contrarily to observations by psychologists (see, e.g., Ross and Leppner 1980). This may be due to the fact that subjects are paid more for a correct than for an incorrect self-assessment in my experiment.

For the final self-assessment, I observe once more that those subjects who received robust feedback make up the majority (63.64%) of those with a correct overall assessment.

Individual Patterns of Reactions to Feedback

In this section, I want to have a closer look at individual reactions to feedback. There are many subjects, who have robust feedback and make most of the time a right self-assessment (33.33% of all subjects). These are subjects who are either good or bad in the question tasks and, moreover, correctly assess this. They have already made a right self-assessment in the first round. Their behavior is consistent with their feedback, perhaps because their belief about themselves is reinforced by the received feedback.

In contrast to these subjects, there are other subjects, who ignore feedback completely or at least mostly (13.33% of all subjects). These are also the subjects who most often make mistakes in evaluating themselves: They are never right. These subjects seem to have no sense of their relative ability and in addition do not react to their feedback.

Other subjects heed their feedback intensively (46.67% of subjects). Some of them are not sufficiently reflective and hence show peculiar reactions to feedback. For example, one subject *always* follows the direction of his changing feedback. Five subjects ignore bad news in the beginning, but after receiving repeatedly bad news, they start to change their self-assessment with some delay. Two of those subjects change their belief so dramatically

that they ignore good news in the end. Finally, one person ignores good news initially and rather follows bad news.

4.4 Summary and Conclusion

In this section, I sum up the results and conclude.

Summary

At the beginning of my experiment, 46.67% of subjects make a wrong relative self-assessment. After receiving feedback in five rounds, still 33.33% are wrong. Overall, the reduction of wrong self-assessments is just moderate and statistically not significant.

Feedback is generally followed by a majority of subjects. Moreover, following feedback seems to be a successful strategy to reduce mistakes. I find that the rate of mistakes is much lower and decreasing for subjects who follow feedback. In particular, subjects who change their belief often are able to considerably reduce their mistakes when following their feedback. When subjects do not follow their feedback, most of them wrongly assess whether they belong to the better or worse 50% in the current block.

I observe that subjects who receive robust feedback are less often wrong with their self-assessments than subjects who receive varying feedback. Subjects with robust feedback represent a majority of subjects making a correct self-assessment. Interestingly, they are not correct right from the beginning but their self-assessments considerably improve. Unsurprisingly, subjects have problems to assess themselves correctly in a round when their relative ability changed from the preceding to the current block. This is likely to cause confusion. In the last block, however, the rate of mistakes is not higher than for subjects with non-changing relative ability. This may indicate a learning effect.

Feedback messaging good news is followed more frequently than when bad news are messaged. This is analogously true for confirming and non-confirming feedback which can also be interpreted as good respectively bad news. Mistakes after bad news (in the former sense) are very high in the beginning but are considerably reduced over rounds. Interestingly, bad news in the sense of non-confirming feedback, i.e. the inherent message that one's preceding assessment is false, shows no appropriate reaction. Contrarily, the

number of mistakes increases over rounds. Such feedback might lead to confusion and thus to mistakes.

Concerning individual patterns of reactions to feedback, I identify three major groups: One group has rather robust feedback and mostly makes correct assessments. Subjects are either good or bad in answering the questions and know that. The subjects of another group mostly ignore feedback, make many mistakes, and have little sense of their true ability. Subjects of a third group heed their feedback intensively. Some of them are not sufficiently reflective so that sometimes peculiar reactions are observed.

Conclusion

At the end of my experiment still a third of the subjects wrongly assess their (relative) ability including those who are overconfident. It is not evident from my experiment whether this is caused by a real bias or just by mistakes which could be further reduced over time by learning. The decrease in mistakes is driven by subjects that get robust feedback, draw right conclusions and mostly follow their feedback. This indicates that at least some subjects are not really biased. But for those subjects who still make wrong self-assessments after five rounds of feedback it can hardly be identified whether these mistakes are due to a bias, to feedback that is difficult to process, or due to a lack of reasoning. For future research it seems therefore interesting to disentangle these effects.

4.5 Appendix

Instructions of *Experiment F* in German:

In dem wissenschaftlichen Experiment, an dem Sie heute teilnehmen, können Sie durch Ihre Entscheidungen Geld verdienen. Ihre Auszahlung aus diesem Experiment hängt von Ihren eigenen Entscheidungen sowie von den Entscheidungen von 15 anderen Experimentsteilnehmern ab, die hier nicht anwesend sind, sondern bereits letzte Woche an einem Experiment teilgenommen haben (*Experiment A*). Während des Experimentes wird Ihre Auszahlung in der fiktiven Währung „Taler" berechnet. Am Ende des Experimentes wird diese Summe an Talern zum Wechselkurs von **1 Euro pro 210 Taler** umgetauscht und bar an Sie ausbezahlt.

Ablauf des Experimentes:

Das Experiment besteht aus 6 Blöcken, die jeweils zwei Stufen umfassen: Auf Stufe 1 eines Blockes beantworten Sie 15 Multiple-Choice-Fragen. Auf Stufe 2 eines Blockes treffen Sie eine Entscheidung. Wenn Stufe 2 des sechsten und somit letzten Blockes vorbei ist, treffen Sie noch einmal eine Entscheidung (diese wird im Folgenden mit *Entscheidung **S*** bezeichnet).

Stufe 1:

- Ihnen werden **15 Multiple-Choice-Fragen** aus verschiedenen Themengebieten gestellt. Für jede Frage werden Ihnen **4 Antwortmöglichkeiten** zur Auswahl gegeben. Nur **eine** dieser vier möglichen Antworten ist jeweils korrekt. Eine mögliche Multiple-Choice-Frage könnte z.B. sein:

  Welches Festival findet nicht in Deutschland statt?
  - o Das Wacken Open-Air Festival
  - o Das Roskilde Festival
  - o Das Zillo Festival
  - o Das Hurricane Festival

- Sie wählen Ihre Antwort auf eine Frage aus, indem Sie auf den Kreis vor der entsprechenden Antwort klicken und danach auf „OK" drücken. Sobald Sie auf OK gedrückt haben, können Sie Ihre Antwort nicht mehr ändern und die nächste Frage wird eingeblendet.

- Für die Beantwortung jeder einzelnen dieser Fragen haben Sie maximal 20 Sekunden Zeit. Während dieser 20 Sekunden können Sie jederzeit Ihre Antwort abgeben. Die

noch verbleibende Zeit pro Frage wird auf dem Bildschirm eingeblendet. Wenn die Zeit abgelaufen ist, so zeigt der Computer automatisch die nächste Frage an.

Bitte beachten Sie:

Wenn Sie keine Antwort ankreuzen oder wenn Sie nicht auf „OK“ drücken, bevor die Zeit abgelaufen ist, so ist dies gleichbedeutend mit einer falschen Antwort.

- Sobald Sie die 15 Fragen beantwortet haben, beginnt Stufe 2 eines Blockes.

Stufe 2:

- Der Computer ermittelt für einen Block, wie viele Fragen Sie richtig beantwortet haben und wie viel Zeit Sie für die **richtig beantworteten** Fragen benötigt haben.
- Anhand der Anzahl richtig beantworteter Fragen in einem Block wird eine **„relative Rangordnung“** (Erläuterung folgt unten) von Ihnen und den 15 Teilnehmern des *Experimentes A*, welches bereits letzte Woche stattgefunden hat, ermittelt. In diesem *Experiment A* haben die Teilnehmer ebenfalls Stufe 1 unter den gleichen Bedingungen gespielt. D.h. die Teilnehmer wurden für jede richtige Frage nach dem gleichen Schema wie Sie entlohnt (Beschreibung folgt unten) und haben die gleichen Fragen, in den gleichen Blöcken unter den gleichen Zeitrestriktionen beantwortet. Für jeden Teilnehmer von *Experiment A* wurde ebenfalls ermittelt, wie viele Fragen richtig beantwortet wurden und welche Zeit hierfür benötigt wurde.
- Die **relative Rangordnung** dieser Gruppe von 16 Personen (die 15 Teilnehmer aus *Experiment A* und Sie) in einem Block wird nun folgendermaßen bestimmt: dem Teilnehmer, der die **meisten Fragen** in einem Block richtig beantwortet hat wird der **höchste Rang (Rang 16)** zugeordnet, dem Teilnehmer, der die **zweit meisten** beantwortet hat, der **zweit höchste Rang (Rang 15**) usw. bis zu dem Teilnehmer der die **wenigsten Fragen** beantwortet hat – dieser erhält **Rang 1**. Sollten zwei oder mehr Teilnehmer gleich viele Fragen in einem Block richtig beantwortet haben, so wird dem Teilnehmer der höhere Rang zugewiesen, der weniger Zeit für die richtigen Antworten in diesem Block benötigt hat. **Ihr Rang in der Rangordnung wird Ihnen nicht mitgeteilt.**
- Basierend auf der relativen Rangordnung werden zwei **Bereiche A und B** bestimmt. Hierbei umfasst Bereich B die Teilnehmer mit den Rängen 1 bis 8 und Bereich A die Ränge 9 bis 16.

Bereich A: umfasst Telnehmer mit Rängen 9-16
Bereich B: umfasst Teilnehmer mit Rängen 1-8

Ihre Entscheidung auf Stufe 2:

Sie treffen eine Entscheidung zwischen **2 Aktionen: Aktion A oder Aktion B.** Sie treffen Ihre Entscheidung, indem Sie das zugehörige Feld anklicken und danach auf „OK" drücken. In Abhängigkeit von Ihrer Aktionswahl und zu welchem der zwei Bereiche Sie gehören wird Ihre Auszahlung bestimmt. Dies wird unten detailliert erläutert.

Information auf Stufe 2:

In allen Blöcken bis auf Block 1 erhalten Sie **bevor** Sie Ihre Entscheidung auf Stufe 2 treffen die folgenden Informationen:

- Sie erfahren, wie viele Fragen Sie im **vorangegangenen Block** richtig beantwortet haben. Das heißt, vor Ihrer Entscheidung in Block 2 erfahren Sie wie viele Fragen Sie in Block 1 richtig beantwortet haben. Vor Ihrer Entscheidung in Block 3 erfahren Sie, wie viele Fragen Sie in Block 2 richtig beantwortet haben usw.
- Sie erfahren, zu welchem Bereich (A oder B) Sie im **vorangegangenen Block** gehören. Das heißt, vor Ihrer Entscheidung in Block 2 erfahren Sie zu welchem Bereich Sie in Block 1 gehören usw.

Sobald Sie Ihre Entscheidung auf Stufe 2 getroffen haben, beginnt Stufe 1 des nächsten Blockes. Nachdem Sie Ihre Entscheidung auf Stufe 2 von Block 6 getroffen haben, treffen Sie noch *Entscheidung **S***.

Ihre Auszahlung für Stufe 1 und 2:

Sie werden **nicht** für alle 6 Blöcke entlohnt: Für Stufe 1 und 2 wählt der Computer **jeweils einen** der sechs Blöcke **zufällig** aus. Stufe 1 und 2 werden dann für den zufällig gezogenen Block entlohnt.

Auszahlung Stufe 1:

Der Computer wählt zufällig einen der 6 Blöcke aus, auf dem Ihre Entlohnung für Stufe 1 basiert. Sie erhalten für jede **richtig** beantwortete Frage in diesem Block eine Auszahlung in Höhe von **270 Talern**, für jede **falsch beantwortete 0 Taler**. Wenn Sie eine Frage richtig beantwortet haben, so werden Ihnen **für jede Sekunde**, die Sie hierfür benötigt haben **0.9 Taler** von den 270 Talern **abgezogen.**

Auszahlung Stufe 2:

Der Computer wählt wiederum zufällig einen der 6 Blöcke aus. Ihre Auszahlung hängt davon ab, zu welchem Bereich (A oder B) Sie in diesem Block gehören und welche Aktion (A oder B) Sie wählen. Die folgende **Auszahlungstabelle** gibt Ihnen für jede mögliche Bereich- und Aktion- Kombination Ihre Auszahlung in Talern an:

		Aktion	
		A	***B***
Bereich	***Bereich A (umfasst Teilnehmer mit Rängen 9-16)***	1500	180
	Bereich B (umfasst Teilnehmer mit Rängen 1-8)	180	1500

Anmerkung: Die zufällige Auswahl eines Blockes für die Auszahlung auf Stufe 1 und 2 erfolgt so, als würde für jede Stufe einmal gewürfelt werden und die resultierende Zahl bestimmt den für diese Stufe relevanten Block.

Nachdem Stufe 2 des sechsten Blockes beendet ist, treffen Sie *Entscheidung S*:

***Entscheidung S*:**

- Zunächst wird Ihre **relative Rangordnung** in der Gruppe von 16 Personen (die 15 Teilnehmer aus *Experiment A* und Sie) über **alle sechs Blöcke** bestimmt. Das heißt, die Rangordnung wird über die **Gesamtanzahl** richtig beantworteter Fragen **in allen 6 Blöcken zusammen** bestimmt. Ansonsten wird die Rangordnung nach dem gleichen Prinzip wie oben erläutert bestimmt: Der Teilnehmer, der **insgesamt** die **meisten Fragen** richtig beantwortet wird der **höchste Rang (Rang 16)** zugeordnet usw.

  Ihr Rang in dieser gesamten Rangordnung über alle 6 Blöcke wird Ihnen nicht mitgeteilt.

- Sie treffen eine Entscheidung zwischen A und B. Die folgende **Auszahlungstabelle** gibt Ihnen für jede mögliche Bereichs- und Entscheidungs- Kombination Ihre Auszahlung in Talern an:

		Entscheidung S	
		A	***B***
Bereich	***Bereich A (umfasst Teilnehmer mit Rängen 9-16)***	300	20
	Bereich B (umfasst Teilnehmer mit Rängen 1-8)	20	300

Ihre Gesamtauszahlung:

Ihre Gesamtauszahlung aus dem Experiment ergibt sich aus der Summe der Auszahlungen aus Stufe 1 und Stufe 2 (gemäß den zufällig bestimmten Blöcken), aus der Auszahlung für *Entscheidung* ***S*** sowie einer zusätzlichen Zahlung von 525 Talern. Ihre Gesamtauszahlung wird zum Wechselkurs 1 Euro = 210 Taler in Euro umgerechnet.

Wenn Sie noch irgendwelche Fragen haben, wenden Sie sich bitte direkt an uns!

Chapter 5 [95]

Solidarity and Performance Differences

5.1 Introduction

Solidarity is defined as an act of providing support for a person in need (Bierhoff and Fetchenhauer, 2001). An important feature of solidarity is that the helping person implicitly assumes that it will be supported if it itself gets in a situation of distress. This is why the philosopher Jon Elster (1989) refers to solidarity as conditional altruism. One example for expressing solidarity with people in need is the support of unemployed (when neglecting the unemployment insurance). In some countries, e.g. Germany, the government has implemented a national unemployment compensation system. It is interesting to investigate if people's willingness to show solidarity with unemployed is influenced by the information whether the unemployed want to work or not. It may intuitively be expected that the willingness to support voluntary unemployed is rather low. People may infer that these unemployed did not work hard and invoked their situation of distress. In contrast, involuntary unemployed want to work and are not comfortable with their situation. Therefore, treating both groups equally may be considered as unfair.

Among investigations in sociology and political science, solidarity has also become a research topic in experimental economics. A seminal experimental study is conducted by Selten and Ockenfels (1998) who introduce the *solidarity game*. The participants of this experiment are randomly divided into groups of three and take part in a lottery. Each of the three subjects has a probability of 2/3 to win 10 DM and a probability of 1/3 to get nothing. These probabilities are independent of each other, so there can be no winner, one, two or three winners in a group. Before playing the lottery, each participant has to specify an amount of money, which he is willing to pay for each loser conditional on being a winner (conditional gifts). Thus, subjects have to specify their gifts to the loser(s) before winner(s) and loser(s) are determined.[96] Next to fairness arguments of distributional concerns (e.g.

[95] This chapter is based on Eberlein and Przemeck (2008a).
[96] This method corresponds to the strategy method (Selten 1967).

inequity aversion[97]), implicit reciprocity plays an important role. Donators might assume to be supported if they were losers themselves. This is different from pure reciprocity since actually gifts cannot be reciprocated.[98] Moreover, implicit reciprocity constitutes the crucial difference between the solidarity and the dictator game.[99]

Selten and Ockenfels (1998) show that most subjects donate positive gifts. Particularly, it is interesting that a bulk of persons are willing to transfer a fixed amount of money, independent of the number of losers. Selten and Ockenfels (1998) call this behavior *fixed total sacrifice*. So each person first decides how much money to keep and then distributes the remaining amount between the losers without taking into account the number of losers. Furthermore, females show more solidarity than males. Additionally, Selten and Ockenfels (1998) find an education effect: Male economic students transfer significantly less than male non-economic students. For females this education effect cannot be found.

Further experimental studies concerning the solidarity game are conducted by Ockenfels and Weimann (1999), Büchner et al. (2007) and Trhal and Radermacher (2006). Ockenfels and Weimann (1999) show that subjects from East Germany donate significantly less than West German subjects. The analysis of Büchner et al. (2007) investigates whether the results found by Selten and Ockenfels (1998) change if a person already knows to be a winner of the lottery at the time of specifying the gift for the loser(s). They postulate that implicit reciprocity does not influence solidarity. Moreover, Büchner et al. (2007) consider the influence of a constant group endowment depending on the different outcomes of the lottery (none, one, two or three winners). In contrast to Selten and Ockenfels (1998), they find that each loser gets a *fixed relative gift*. This means that a winner donates the same relative amount to each loser, independent of the number of losers. Consequently, a loser can expect to get the same income in each case, independently of the number of winners. Thral and Radermacher (2006) investigate gift giving behavior in the solidarity game when subjects can choose between a secure payment and a lottery including a probability of becoming needy. In a "within subjects" design they find that subjects give less if losing is

[97] Seminal works on inequity aversion in social psychology are Festinger (1954), Homans (1958, 1960, 1968) and Adams (1963, 1965). A detailed overview of theoretical economic models on inequity aversion is given by Eberlein and Grund (2006).
[98] There is a huge experimental literature concerning fairness, inequity aversion, and reciprocity, see, e.g., Forsythe et al. (1994), Güth et al. (1982) and Fehr et al. (1997). For reciprocity based models see, e.g., Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006).
[99] In the dictator game, an allocator can divide a certain amount of money between himself and a receiver. Then both players receive their payoffs according to this allocation.

self-inflicted by choosing the lottery. This result cannot be found in a "between subjects" analysis.

In the solidarity game experiments, subjects get an amount of money without having "worked" for it. I design an experiment to study the impact of real effort and performance differences on donating behavior in the solidarity game.[100] Intuitively, donations may differ when a real effort task is introduced to the solidarity game: If a person has to work to get an income and then loses this income in the lottery, donators may give a higher gift to this person than to another person who did not have to bear effort costs. Considering donators who had to work and other donators who did not have to work it may be the case that the latter donate higher gifts because they saved effort costs.

When real effort is introduced to the solidarity game, not only effort costs may play a crucial role but also performance differences between the subjects in a group. In the real working life, employers select applicants mainly according to their education, training, ability, and motivation. Thus performance can influence working position, income and status. This is just one reason why performance is regarded as very important in our society. Therefore, it is reasonable to assume that prosocial behavior and particularly solidarity are influenced by living in a performance-oriented society. Applied to the solidarity game, it may be the case that donators regard performance differences between losers. It is possible that losers, who performed the real effort task very well before the lottery is played, get higher gifts than other losers who performed worse. This is even more pronounced when subjects' performances influence their probability to lose the lottery, with a better performance implying a lower losing probability. A subject who performed very poor and loses the lottery may then be more blamed to be responsible for his situation of distress than someone who performed very well and therefore had only bad luck to lose the lottery. Therefore gifts to the latter one may be higher. Next to these aspects, performance differences may have another influence on donators' gifts that may arise because of a comparison between a donator's performance and the loser's performance.

[100] The experimental literature that investigates the influence of real effort (and asset origin) in other experimental games is broad: *Dictator game*: e.g. Cherry (2001), Cherry et al. (2002), Hoffman et al. (1994), Ruffle (1998); *ultimatum game*: e.g. Ruffle (1998), Hoffman et al. (1996); *trust game*: e.g. Fahr and Irlenbusch (2000), Oxoby and Friedrich (2006); *public good game*: e.g. Cherry et al. (2005); *auction*: e.g. Ball et al. (2001); *bargaining game*: e.g. Hoffman and Spitzer (1985), Burrows and Loomes (1994), Rutström and Williams (2000).

The remainder of the chapter proceeds as follows. Section 5.2 describes the procedure and the design of the experiment. Section 5.3 states the hypotheses for the different treatments, while section 5.4 presents and discusses the results of the experiment. The final section concludes.

5.2 Procedure and Experimental Design

5.2.1 Procedure

I conducted a pen and paper experiment at the Experimental Economics Laboratory of the University of Bonn. One session lasted approximately one hour and the average earning was 8.80 €. I ran two sessions per the first two treatments and one session for the last treatment that combined the other two treatments. 24 subjects participated in each session of my first two treatments, while 48 subjects participated in the session of the last treatment. In each treatment, subjects were organized into groups of three and did not know their group members. Thus, I had 16 three-person-groups per treatment.

Subjects were recruited via the internet by using ORSEE software (Greiner 2004) announcing the possibility to earn an amount of money dependent on their behavior. I had a balanced proportion of economic and non-economic students and males and females.

To ensure credibility of anonymous gift giving decisions, I used a double blind procedure. Subjects were paid by a neutral person who neither knew the content of the experiment nor the instructions. The payments were computed by two other persons who knew the instructions but could not attribute decisions to individual participants. The instructions were written in a neutral language. For example, I avoided expressions like “gift” or “donation” and used the word *transfer* instead.

5.2.2 Experimental Design

I run three treatments that differ with respect to the assignments of incomes and losing probabilities before subjects participate in a monetary lottery.

In the first treatment, the real effort treatment (RET), subjects have to work on a real effort task before taking part in a lottery.[101] They have to bead little plastic pearls on a string for ten minutes. I choose this task because of the following advantages: Subjects suffer from effort costs, subjects do not need to exhibit extraordinary abilities, and outcomes (i.e. the number of beaded pearls) are very unlikely to be exactly the same for the subjects in a group. I think that every subject is basically able to undertake this task. Furthermore, a high outcome mostly depends on high effort and not on ability. This is important for my experiment because subjects shall be able to evaluate whether someone is motivated and works hard to be successful. Moreover, I want to rank subjects according to their performance. Thus subjects' outcomes in a group should be different.

I organize a rank-order tournament in each three-person-group. After performing the task, subjects are ranked according to their task performance: The person who beaded the most pearls becomes *rank 1*, the second best becomes *rank 2*, and the worst becomes *rank 3*.[102] The rank determines the provisional income and the probability to lose the subsequent lottery. The connection between provisional task income and rank is as follows: *rank 1* gets 9 €, *rank 2* obtains 5 €, and *rank 3* gets 1 €. These amounts are only paid if subjects do not lose the lottery. The income of a loser of the lottery is reduced to zero. The winners get 7 € in addition to their provisional income. The probabilities of losing the lottery are the higher the worse the rank: *rank 1* loses with probability 1/6, *rank 2* with 2/6, and *rank 3* with probability 3/6. I choose these losing probabilities to picture the aspect of provoking a situation of distress. The intuition is that someone who is motivated and works hard is exposed to a lower risk to suffer from financial distress. Moreover, I choose these specific probabilities to enhance subjects' comprehension. They know from the instructions that the loser of a group is selected by rolling a dice for each group. *Rank 1* loses in case of a "one", in case of a "two" or "three", *rank 2* loses, and in case of a "four", "five" or "six", *rank 3* loses. The lottery assigns two winners and one loser in each group.[103]

A rank-order tournament is well suited for my experiment because of several reasons: First, subjects are intensively motivated to exert effort. Second, the conditions for various achievements are obvious and easily comprehensible for the subjects. Third, I want to

[101] The instructions of RET can be found in the appendix (section 5.6).
[102] In case of a tie I dice the better rank and inform the subjects about the random selection via a note on their decision sheet. However, a tie never occurred in my treatments.
[103] My focus is not the comparison of donating behavior in case of one or two losers. This is different from the study of Selten and Ockenfels (1998) in which the probabilities of losing the lottery are independent.

implement a competitive situation to test if subjects show solidarity even in such an achievement-oriented environment. Fourth, my main research question is whether performance differences influence solidarity: Thus I want to create an environment which enables subjects to compare their performance with the performance of other group members. Finally, the tournament ensures that a subject with a particular rank gets the same income as another subject with the same rank in another three-person-group. Hence, donations of subjects with the same rank are well comparable across three-person-groups.

In my second treatment, the non-real effort treatment (the NRET), subjects do not have to perform a real effort task. Instead of achieving ranks, subjects are randomly assigned to particular *type categories* labeled by *a*, *b*, or *c*. *Type a* equals *rank 1* because he has a provisional income of 9 € and a losing probability of 1/6. Analogously, *type b* (*type c*) gets a provisional income of 5 € (1 €) and has a losing probability of 2/6 (3/6). As in RET, winners of the lottery get 7 € in addition to their provisional income while the income of the loser is reduced to zero.

In both treatments, RET and NRET, each subject is informed about his rank respectively type before the lottery is played. Before the lottery starts, each subject has to announce an amount of money which he wants to donate to the loser of his group. The subjects' donations are elicited by applying the strategy method: Each subject has to state two transfer decisions: one decision for each possible loser rank respectively type.[104] After actual winners and losers are determined, one winner of each three-person-group is randomly chosen for the actual transfer to the actual loser. The other donation decision of this winner is not implemented. Moreover, the donation decisions of the other winner who is not chosen for the transfer and the loser's decisions are not implemented. Thus subjects know that their indicated donation to the factual loser of their group is conditioned on winning the lottery and on being chosen for the transfer.

The third treatment, the matched treatment (MT), combines RET and NRET. In this treatment, subjects are randomly assigned to two different rooms. In each room subjects are organized into groups of three and matched with another three-person-group from the other room. In one room subjects have to work on the real effort task. I call this sub-treatment MT-R. The procedure of MT-R is the same as in RET but subjects have to state

[104] With "loser rank" and "loser type" I refer to the potential losers of the lottery.

five transfer decisions (see below). In the other room, subjects do not have to perform a real effort task. This sub-treatment is called MT-N. Analogously, the procedure of MT-N is the same as in NRET but again subjects have to state five transfer decisions (see below). Thus, in MT a three-person-group of MT-R, in which subjects undertake the real effort task, is matched with a three-person-group of MT-N, in which subjects do not have to work. The lotteries are played separately for each group and determine two winners and one loser in each three-person-group, as it is the case in RET and NRET.

Before the lottery is played, each subject is informed about his rank respectively type and has to announce an amount of money which he wants to donate to the two potential losers of his group and to the three members of the matched group. Hence, each subject has to state five transfer decisions: one decision for each possible loser rank and one decision for each possible loser type.[105] After actual winners and losers are determined, one winner of each three-person-group is randomly chosen for the actual transfer. Then the allocation of the two relevant donations is randomly determined: One winner's donation is given to one of the two actual losers (either in his own or in the matched group), the donation of the chosen winner from the other group is transferred to the other loser. The other donation decisions of these winners are not implemented. Moreover, the donation decisions of winners who are not chosen for the transfer and the losers' decisions are not implemented. Therefore subjects know that their indicated donation to the factual loser of their group or the matched group is conditioned on winning the lottery and being chosen for the transfer, as in treatments RET and NRET.

5.3 Experimental Hypotheses

When regarding the three-person-groups in RET and MT-R, the subject with the best rank in a group has worked harder than his group members. This is rewarded with a higher provisional income and a lower losing probability. Nevertheless, this rank can lose all his money and end up in a situation of distress when losing the lottery. Considering achievement justice, donators may think that this potential loser deserves an exceptionally high donation. In particular, he deserves a higher donation than a worse potential loser rank

[105] For example, *rank 1* states transfer decisions for the potential losers *rank 2* and *rank 3* of his own group and for *type a*, *type b*, and *type c* of his matched group.

who is more responsible for losing the lottery because he only achieved a high losing probability.

This idea is strengthened by aspects of attribution theory. Among other things, attribution theory deals with the perceived causes of success and failure in achievement related situations which result in particular emotions (Weiner, 1986).[106] Weiner argues that emotions are related to causal dimensions: He distinguishes between the causal locus (internal versus external), the causal controllability (controllable versus uncontrollable), and the causal stability over time (stable versus unstable). In this context, there are some studies which deal with attribution theory and achievement evaluation and the influence of emotions on helping behavior (e.g. LePine and Van Dyne 2001). It is shown that causes, which are perceived as controllable by a needy person (like low effort), lead to a decrease in the probability to help. In contrast, causes which are perceived as uncontrollable by a needy person lead to an increased probability of help. In my experiment, I expect that subjects receive lower donations the lower their efforts and therefore the worse their ranks because being a loser in the lottery is then more self-inflicted, although not deterministically caused. Therefore I state the following hypothesis:

Hypothesis 5.1: Given a particular donator rank, gifts to the better of the two potential losers are higher than to the worse one.

In NRET and MT-N, I expect that donations of a particular winner type to various loser types do not systematically differ because income and probability assignments are random. In this case, arguments considering achievement justice do not apply since subjects do not work at all. No one can be blamed for having invoked a situation of distress because the assignment to the different types (and therefore to the losing probabilities) is not controllable by a subject himself. Therefore, I state the following hypothesis:

Hypothesis 5.2: Donations to various loser types do not differ, given a particular winner type.

I now consider my third treatment, in which a three-person-group of MT-R is matched with a three-person-group of MT-N. How do donations of a particular donator vary if the

[106] See Graham (1991) for a review of attribution theory in achievement contexts.

recipient is either a type or a rank? In this context, achievement justice and attribution theory do not deliver clear implications. While the type allocation is not controllable by a subject, the rank allocation mirrors the subjects' efforts. As described above, losing the lottery will be perceived as self-inflicted if *rank 3* loses but as bad luck if *rank 1* loses. But it is not obvious from these theories whether donators implicitly evaluate these performance differences when comparing a particular loser rank with a particular loser type. However, I think that donators may take into account that each rank has worked prior to the lottery and had to incur effort costs. Then donators might feel sorrier for a rank, who loses the lottery, than for a type. I so suppose that donations to a particular loser rank of MT-R are higher than donations to the corresponding loser type of MT-N.

Hypothesis 5.3: Donations to loser ranks are higher than donations to the corresponding loser types, given a particular donator.

When comparing the donations of ranks and those of the corresponding types to a particular loser, I expect that donations of potential winner types are higher than those of the corresponding potential winner ranks because ranks have to bear effort costs while types do not have to exert effort.[107]

Hypothesis 5.4: Donations of winner ranks are lower than donations of the corresponding winner types, given a particular loser.

5.4 Experimental Results

When analyzing the experimental data, I start with RET and MT-R, afterwards I present the results of NRET and MT-N, then I compare donations with and without real effort. Lastly, I discuss expected conditional gifts.

First of all, I examine whether I can pool the data of RET and MT-R (only donation decisions of a particular rank to a particular rank) and the data of NRET and MT-N (only donation decisions of a particular type to a particular type). Two-sided Mann-Whitney U tests for independent observations reveal that there are neither significant differences

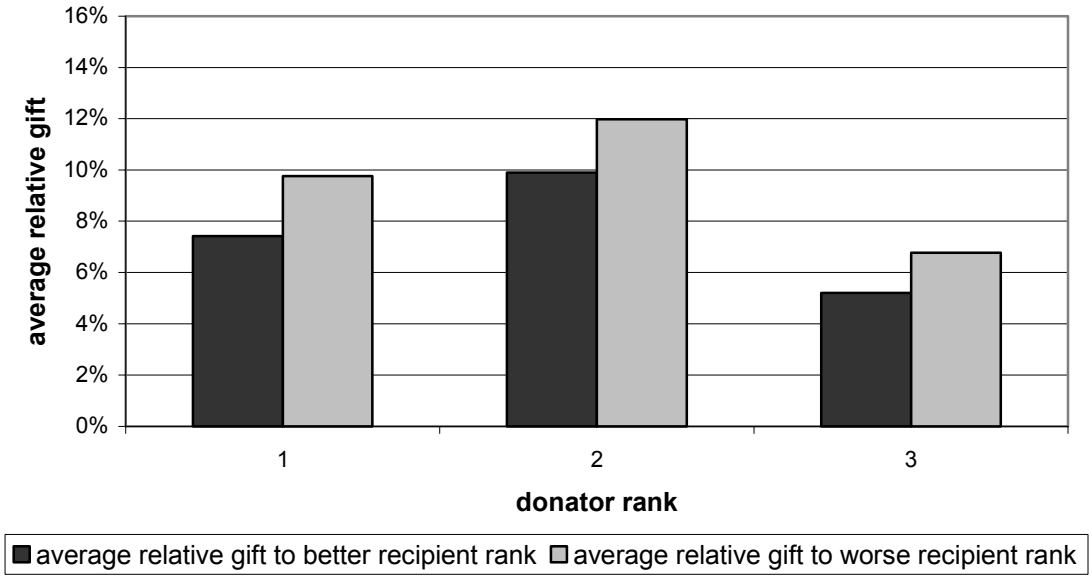
[107] Considering *dictator games*, it is found that allocators, who earned their money, behave more egoistically than randomly assigned allocators (see, e.g., Cherry 2001, Cherry et al. 2002, and Hoffman et al. 1994).

between the absolute gifts regarding RET and MT-R[108] nor between the absolute gifts of NRET and MT-N.[109] In the following I present the results of the pooled dataset when investigating donations of winner ranks to loser ranks and of winner types to loser types. Thus I have a considerable greater number of observations (24 independent observations of a particular donator to a particular recipient).

5.4.1 Donations in a Real Effort Environment

Figure 5.1 gives an impression of the subjects' average gifts relative to their final income (i.e. provisional task income + 7 €). It illustrates the average relative donation of a particular rank to a particular rank in the pooled dataset of RET and MT-R. As can be seen, subjects donate positive gifts on average. It can be observed that *rank 2* donates the highest

Figure 5.1: Average relative gifts of donator ranks



percentage of his final income in case of winning the lottery: He donates 9.9% of his final income to *rank 1* and 11.98% to *rank 3*. *Rank 1* donates the second highest percentage of his final income (7.42% to *rank 2* and 9.77% to *rank 3*), while *rank 3* donates the lowest

[108] Comparison of *rank 1* RET with *rank 1* MT-R: p = 0.393 regarding the gift to *rank 2*; p = 0.476 regarding the gift to *rank 3*. Comparison of *rank 2* RET with *rank 2* MT-R: p = 0.415 regarding the gift to *rank 1*; p = 0.151 regarding the gift to *rank 3*. Comparison of *rank 3* RET with *rank 3* MT-R: p = 0.731 regarding the gift to *rank 1*; p = 0.731 regarding the gift to *rank 2*.

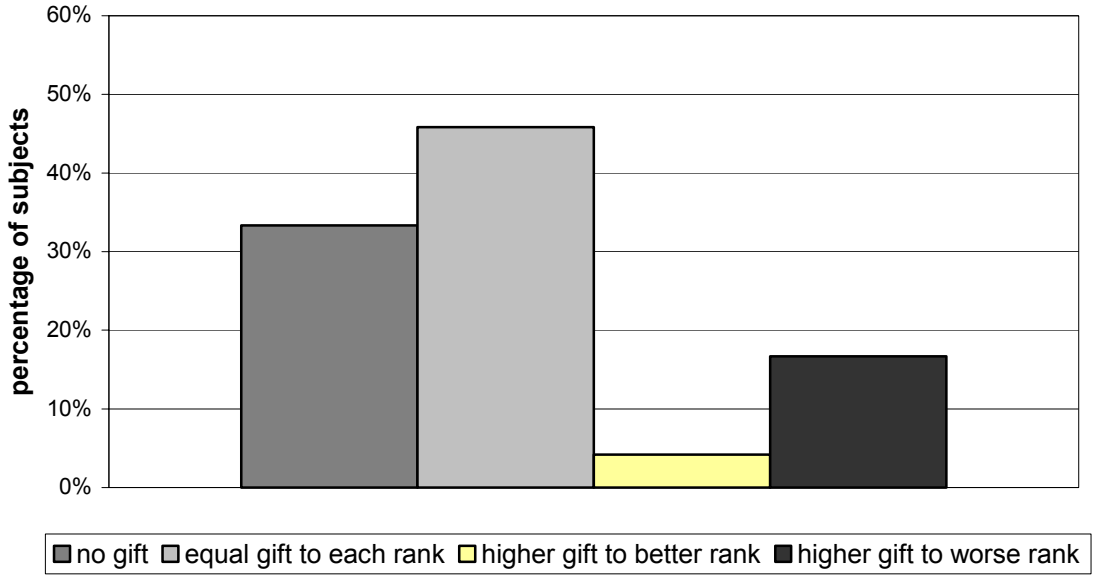
[109] Comparison of *type a* NRET with *type a* MT-N: p = 0.258 regarding the gift to *type b*; p = 0.451 regarding the gift to *type c*. Comparison of *type b* NRET with *type b* MT-N: p = 1 regarding the gift to *type a*; p = 0.244 regarding the gift to *type c*. Comparison of *type c* NRET with *type c* MT-N: p = 0.596 regarding the gift to *type a*; p = 0.602 regarding the gift to *type b*.

percentage (5.21% to *rank 1* and 6.77% to *rank 2*). Interestingly, on average all donator ranks donate more to the respective worse recipient rank than to the better recipient rank.

To investigate whether a potential winner of the lottery donates different gifts to the two potential losers in his group, I compare the two donation decisions of each donator. I test with one-sided Wilcoxon-Signed-Rank tests for dependent observations whether the donators' gifts to the better of the two potential loser ranks are higher (Hypothesis 5.1). The results reveal that a loser with a higher rank does not receive a significantly higher gift. Thus, for example, *rank 1* does not donate more to *rank 2* than to *rank 3*. This shows that potential winners do not systematically support subjects more who have shown a better performance than the other potential loser. This is surprising because the consequence of bad performance and low motivation is a high losing probability. The data contradict Hypothesis 5.1 which can therefore be rejected.

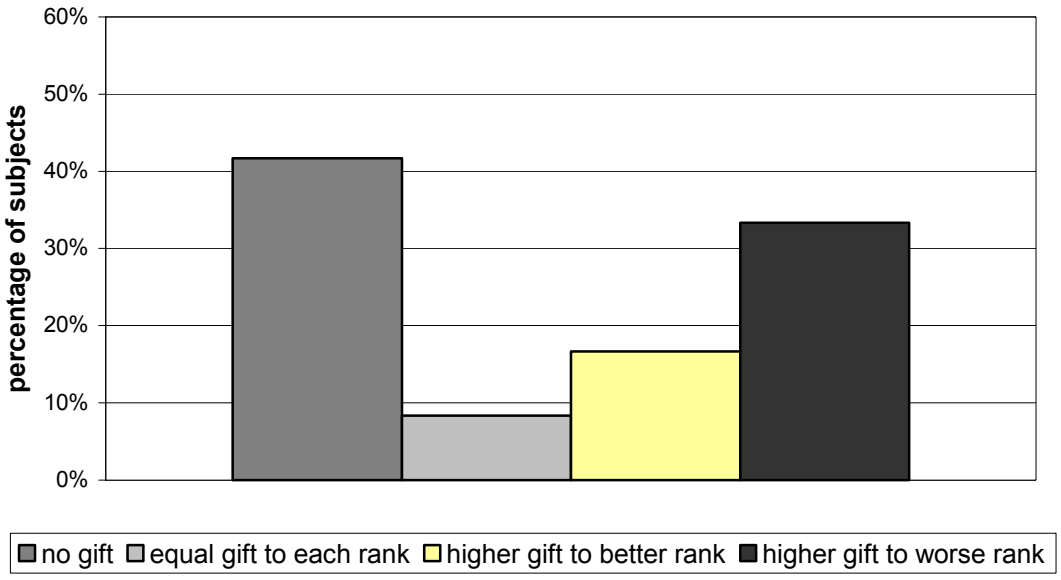
When investigating the data at the individual level, four different kinds of donating behavior can be found (see Figures 5.2 to 5.4): There are a lot of egoists who give nothing to anyone. Moreover, many donators give the same gift to both potential losers. Furthermore, there are donators who give more to the better potential loser. Finally, one can observe potential winners who behave the other way round by giving more to the worse potential loser. As Figure 5.2 shows, the most common behavior of *rank 1* is an equal gift (45.83%). These subjects only seem to consider that each loser's income is

Figure 5.2: *Rank 1*'s donating behavior



reduced to zero. The amount of the lost income[110] and the size of the losing probability are neglected. This hints at the fact that these subjects only pay attention to states and outcomes and not to the factors that lead to these outcomes. A high percentage of *rank 1* subjects (33.3%) donate nothing, which is the most common behavior of *ranks 2* and *3* (see Figures 5.3 and 5.4). It is more common among *rank 2* subjects to give different gifts than to donate equal positive gifts. Out of the non-egoistical subjects most of the subjects take into account a loser's performance but their assessment is ambiguous: 16.67% of *rank 2* subjects give more to *rank 1*, while 33.33% give more to *rank 3*.

Figure 5.3: *Rank 2*'s donating behavior



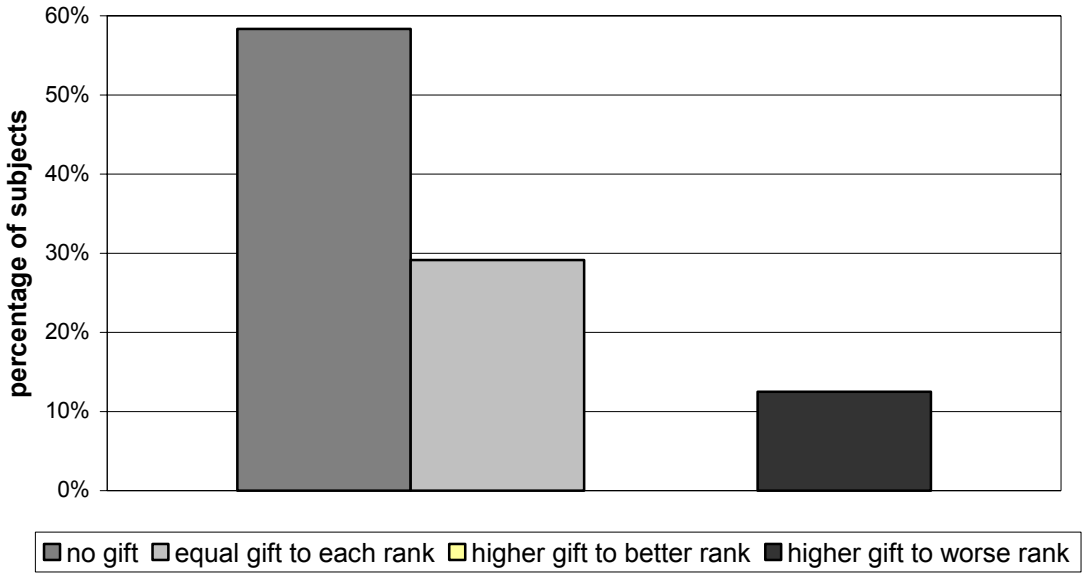
Furthermore my data show that none of the *rank 3* subjects donates more to *rank 1*. Interestingly, most of all non-egoistical *rank 3* subjects donate the same amount of money to the loser ranks. Thus, across all ranks, only 27.77% distinguish between different loser ranks and among those 75% donate more to the worse rank.

I am not only interested in the question whether particular donator ranks make differences between various losers but also whether relative gifts from various ranks to the same loser differ. Given a particular potential loser rank, two-sided Wilcoxon-Signed-Rank tests show that the relative donations of the various winner ranks do not differ significantly.

[110] With respect to neglecting the lost income of a loser, one can imagine a natural disaster. Imagine people of different income classes and professions who all have lost their holdings because of a hurricane. Then it should be plausible to give everybody the same to survive. In such a situation the relative performance of the victims before the disaster should not matter.

Therefore, performance differences between winners and between winners and losers do not seem to matter and do not systematically influence gift giving.

Figure 5.4: *Rank 3*'s donating behavior

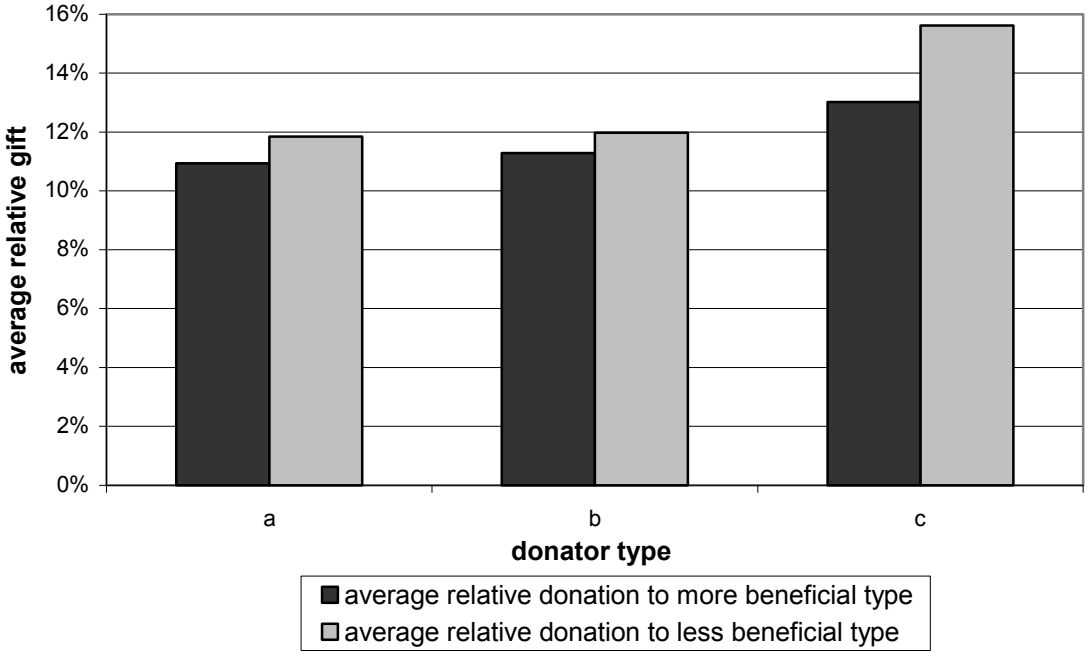


5.4.2 Donations in a Non-Real Effort Environment

Having a look at the relative donations of particular types in the pooled dataset NRET and MT-N, it can be seen from Figure 5.5 that *type c* donates the highest share of his final income. He donates 13.02% to *type a* and 15.63% of his final income to *type b*. Average relative donations of *type a* and of *type b* are slightly smaller and very similar to each other. On average, subjects do not behave egoistically. Each type donates on average a slightly higher amount to the potential loser with the less beneficial type[111] than to the other loser. I conduct two-sided Wilcoxon-Signed-Rank tests to find out whether types donate differently to the two potential losers. As hypothesized, none of these tests shows a significant difference (Hypothesis 5.2). Hence, most donators do not evaluate the loser's type. Income differences and probability allocations arising by chance do not have an obvious influence.

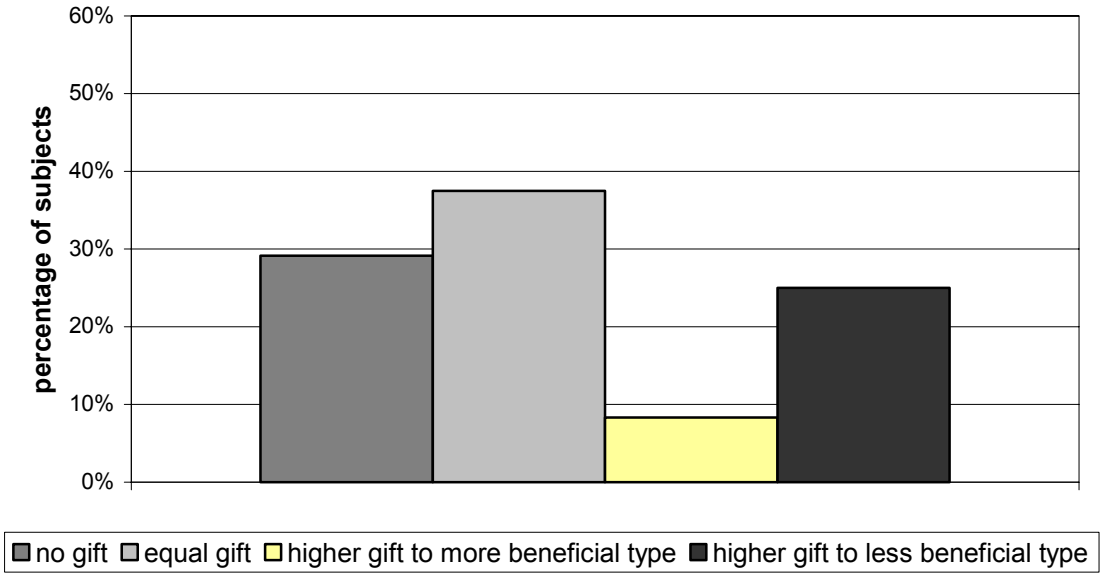
[111] A "more (less) beneficial type" is the type with the higher (lower) income and the lower (higher) losing probability.

Figure 5.5: Average relative gifts of donator types



Investigating the data at the individual level shows that different kinds of donating behavior exist. Figures 5.6 to 5.8 picture the donation behaviors of various types. The most common behavior shown by *type a* is an equal gift to each loser type (37.5%). There is a high percentage of egoistical behavior, too (29.17%). Concerning *type b*, the most common

Figure 5.6: *Type a*'s donating behavior



behavior is to donate nothing (33.33%). This is also the most common behavior of *type c* (37.5%). The most uncommon behavior of each type is to give more to the more beneficial loser type than to the less beneficial loser type. Therefore, potential winners seem to care less about subjects who already had good luck with the type allocation.

Figure 5.7: *Type b*'s donating behavior

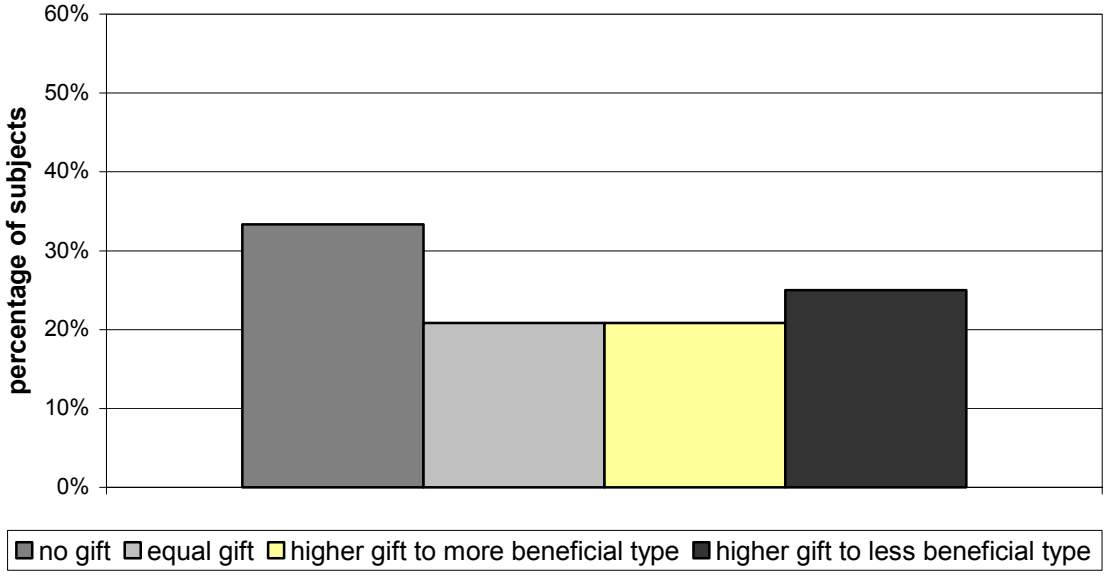
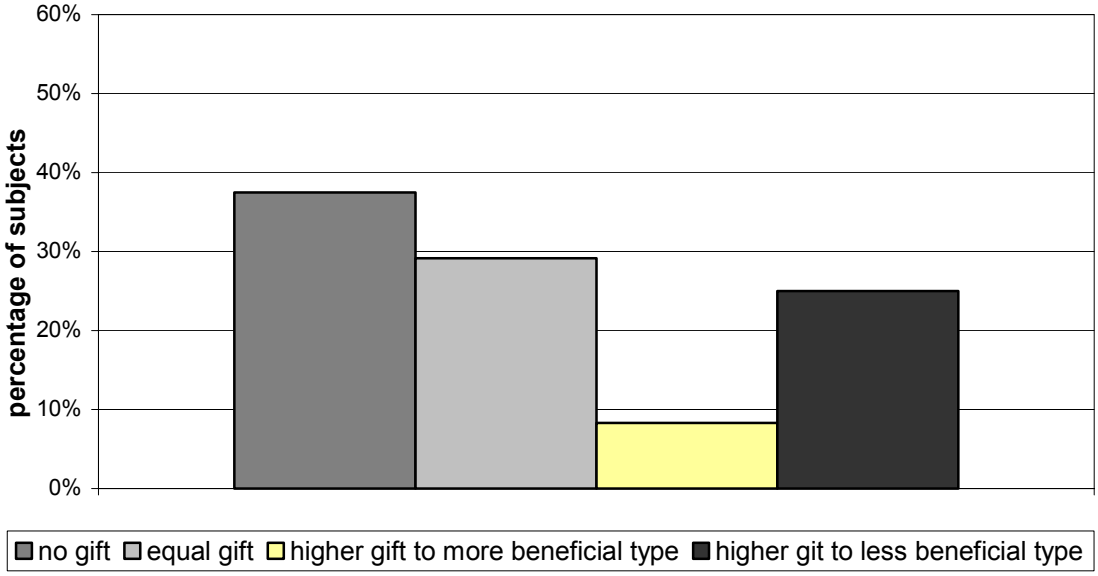


Figure 5.8: *Type c*'s donating behavior



I examine whether various types donate different relative gifts to the same loser type. The results of two-sided Wilcoxon-Signed-Rank tests show that there are no significant differences. Hence, subjects' donating behavior is not influenced by their own "luck" to

become a particular type. Subjects with a higher income do not donate different relative gifts than subjects with a lower income.

5.4.3 Comparison of Donations with and without Real Effort

In this subsection I first investigate whether an environment in which all subjects – potential winners and losers – have to work leads to different donations than an environment in which no subject has to work. Hence, I compare donations from a particular potential winner rank to a particular potential loser rank with the corresponding donation of a particular potential winner type to a particular potential loser type (e.g., I compare the donation of *rank 1* to *rank 2* with the donation of *type a* to *type b*). Two-sided Mann-Whitney U tests mostly show no significant differences. The only significant difference can be found by comparing the donations of *rank 3* to *rank 2* with the ones of *type c* to *type b* ($p = 0.047$). Here, *type c* gives a higher gift than *rank 3* to the respective loser. This might be explained by the fact that *rank 3* had to bear effort costs but *type c* not, which guides his behavior to be more generous. Remember that the average relative donation of *type c* to *type b* is the highest (15.63%) when considering the donation decisions of winner types to loser types. Interestingly, the second highest relative donation is also made from *type c*, but to *type a* (13.02%). A plausible explanation for the high relative donations of *type c* is that implicit reciprocity is strongly perceived by *type c*. His chance to lose the lottery is high. As he already had bad luck with the type allocation he might suffer most from the idea to lose the lottery. Contrarily, the average relative gifts of *rank 3* are the smallest when considering donations of winner ranks to loser ranks. This can be explained by the fact that he suffered from exerting effort but nevertheless has been worse than the other two persons in his group. Maybe he is frustrated and does not want to give high donations to the persons who outperformed him in the tournament.

However, with these comparisons I am not able to clearly identify the influence of real effort and performance. Because winners and losers either have to work or not, it might be that there are influences of real effort, but that these cancel each other out: On the one hand, a loser, who had to work, might get higher donations than a loser who did not have to work. On the other hand, the associated winner also had to work in my so far conducted analysis and might therefore give less. I can control for these different effects by additionally using the donation decisions in MT of winners, who worked and donate to losers that did not work, and of winners, who did not work and donate to losers that worked. Thus, I can examine whether donators give higher gifts to a loser that had to work

than to someone who did not have to work (Hypothesis 5.3). Additionally, I can investigate whether a particular loser receives more from a particular winner type than from a particular winner rank (Hypothesis 5.4).

To investigate Hypothesis 5.3, I examine the donations of a particular winner rank to a particular loser rank and to the corresponding loser type.[112] Furthermore, I investigate whether donator *types* give more to losers that had to work than to losers that did not have to work.[113] The data show that according to one-sided Wilcoxon-Signed-Rank tests donators (ranks as well as types) do not give higher gifts to loser ranks than to loser types.[114] Therefore, Hypothesis 5.3 can be rejected. Donator ranks as well as donator types only seem to perceive whether someone loses the lottery and do not evaluate whether he influenced his losing probability by working (or not).

Investigating Hypothesis 5.4, a one-sided Mann-Whitney U test shows that *rank 1* receives a higher gift from *type c* than from *rank 3* ($p = 0.087$).[115] Moreover, *type a* gets higher donations from *type c* than from *rank 3* ($p = 0.095$) and *type b* gets higher donations from *type c* than from *rank 3* ($p = 0.053$). However, I do not find significant differences between the donations of *rank 3* and *type c* to *rank 2*. The other comparisons also do not show any significant differences. These subjects' solidarity is not influenced by bearing effort costs and working for their income. Thus my results are ambiguous and I cannot reject Hypothesis 5.4 for all comparisons. One reason for the difference in donating behavior between *rank 3* and *type c* can be that *rank 3* gives less to *type a* and to *type b* than *type c* does because these losers did not suffer from effort costs but nevertheless have lower losing probabilities than *rank 3* himself. Perhaps *rank 3* perceives this as very unfair, especially when regarding the fact that the performance of these types could have been worse than his own performance on the task if they had to work on it.

In order to get a better insight into the driving forces behind donating behavior, I conduct a probit regression analysis. In my regression I model the donation as the dependent variable

---

[112] E.g., I compare the gift from *rank 2* to *rank 3* with the gift from the same *rank 2* to *type c*.
[113] E.g., I compare the gift from *type b* to *type c* with the gift from the same *type b* to *rank 3*.
[114] For these comparisons I only use observations from MT, because I consider a particular winner and two of his donation decisions.
[115] For the investigation of Hypothesis 5.4, I pool the data if possible: E.g. I pool observations from RET and MT-R concerning the donations of *rank 1* to *rank 2*, but have only observations from MT-R concerning the donations from *type a* to *rank 2* if I investigate whether *rank 1* gives more to *rank 2* than *type a* does.

which can either be "1" (positive donation) or "0" (no donation). Hence, I investigate the influence of the independent variables on the probability to donate. Thus I enrich my so far conducted analysis in which I turned my attention to the magnitude of a gift.

I regress the donation on the dummy variables *NoWork*, *donator2*, *donator3*, *GiveLowerRecipient*, *female*, *econmajor*, on a constant, and on the interaction term *NoWork* • *GiveLowerRecipient*. According to my research agenda, I investigate whether working on the real effort task influences donations. Therefore I include a dummy variable *NoWork* in my regression model that is 0 if subjects had to work on the real effort task (subjects in RET and MT-R) and 1 if they did not have to work on it (subjects in NRET and MT-N). Moreover, I want to tackle the question whether the rank respectively type of a donator influences the probability to donate. I therefore include the variables *donator2* and *donator3* in my regression model. If subjects have to work, *rank 2* (*rank 3*) is the donator if the dummy variable *donator2* is 1 (0) and the dummy variable *donator3* is 0 (1). If both donator dummy variables are 0, *rank 1* is the donator. If subjects do not have to work, *type b* (*c*) is the donator if *donator2* is 1 (0) and *donator3* is 0 (1). If both donator dummy variables are 0, *type a* is the donator.

Moreover I want to investigate the question whether the rank respectively type of the recipient influences the probability to donate. If subjects have to work and the dummy variable *GiveLowerRecipient* takes the value 1 (0), donations to the potential loser with the worse rank (better rank) are described. If subjects do not have to work and *GiveLowerRecipient* takes the value 1 (0), donations to the loser who is the less beneficial type (more beneficial type) are described.

I include a dummy variable for sex (1 denotes *female*), and one that indicates whether subjects' major is economics (1 denotes *econmajor*) or not. I expect that the probability to donate is higher for females than for males. Moreover I think that the probability to donate is higher for non-economists than for economists. As is found by Büchner et al. (2007) the educational background (i.e. the major) is decisive for gift giving behavior. They find that egoistical behavior is more common between economists than between students of other fields of studies. Table 5.1 shows the estimated coefficients and their effect on a subject's probability to donate.

Table 5.1: Probit estimates for dependent variable donation

Independent variable	Coefficient (Standard error)
donator 2	-.14135
	(.25313)
donator 3	-.46249*
	(.24772)
female	.56146***
	(.20485)
econmajor	-.08047
	(.20424)
GiveLowerRecipient	.18708**
	(.09677)
NoWork	.19465
	(.21371)
NoWork•GiveLowerRecipient	.11449
	(.14964)
const	-.13366
	(.26522)
N	288
clusters	144
(Pseudo) R^2	0.0672
LR-χ^2	22.90
Prob > chi2	0.0018

Asterisks indicate variables as being significant at 1%*, 5%**, and 10%*.**

Table 5.1 shows that the probability of females to donate is significantly higher than that of males. Moreover, someone who is *rank 3* respectively *type c* donates with a significant lower probability than someone who is *rank 1* respectively *type a*. Furthermore, subjects donate with a significant higher probability to the potential loser with the worse rank respectively less beneficial type than to the potential loser with the better rank respectively more beneficial type.

5.4.4 Expected Conditional Gifts

As already described, I apply the strategy method to elicit donation decisions in my experiment. Every subject is instructed to imagine a situation in which he has already won the lottery and has been chosen for the transfer decision. Keeping this in mind, subjects are asked what amount of money they would like to give to the potential losers. These potential losers differ in their rank respectively type and therefore in the specific losing probabilities. The income of the actual loser is reduced to zero. Thus, donators know that

the only income source of the actual loser will be their indicated donation in this situation. When specifying their gifts, it is possible that they take the updated probabilities of the others to lose the lottery into account.

Using Bayes Rule, they can calculate the conditional probability that a particular subject will lose the lottery in case they have won it and were chosen for the transfer in their three-person-group: The probability that *rank 2* (*rank 3*) loses, given *rank 1* is the winner, is 2/5 (3/5). The probability that *rank 1* (*rank 3*) loses, given *rank 2* wins the lottery, is 1/4 (3/4). Given *rank 3* wins the lottery, the probability that *rank 1* (*rank 2*) loses, is 1/3 (2/3). Analogously, the probabilities for types in a three-person-group can be calculated. With these probabilities, subjects can calculate an expected conditional gift to a particular loser. Thus I can investigate whether donator ranks regard performance differences if taking the conditional losing probabilities into consideration. Therefore, I examine the expected conditional gifts.

Comparing expected conditional donations of a particular donator rank, I find significant differences according to two-sided Wilcoxon-Signed-Rank tests: Each rank donates a higher expected conditional gift to the worse of the two potential losers.[116] This shows that subjects consider performance differences when thinking about expected conditional gifts. But again Hypothesis 5.1 can be rejected because donators give a higher expected conditional gift to the worse potential loser rank. Also when investigating the expected conditional donations of types, two-sided Wilcoxon-Signed-Rank tests reveal that types donate more to the loser with the less beneficial type than to the one with the more beneficial type.[117] So Hypothesis 5.2 can be rejected in this specification since donators give different expected conditional gifts to various loser types.

If subjects consider the expected conditional value of their donations, their support is higher for a lower rank respectively a less beneficial type.

---

[116] The significance levels (corrected for zero differences) of the two-sided Wilcoxon-Signed-Rank tests are: $p < 0.001$ concerning the expected conditional gifts of *rank 1* to *rank 2* and to *rank 3*, $p < 0.005$ concerning the expected conditional gifts of *rank 2* to *rank 1* and to *rank 3*, and $p < 0.005$ regarding the expected conditional gifts of *rank 3* to *rank 1* and to *rank 2*.

[117] The significance levels (corrected for zero differences) of the two-sided Wilcoxon-Sign-Rank tests are: $p < 0.005$ concerning the expected conditional gifts of *type a* to *type b* and to *type c*, $p < 0.0001$ concerning the expected conditional gifts of *type b* to *type a* and to *type c*, and $p < 0.001$ regarding the expected conditional gifts of *type c* to *type a* and to *type b*.

5.5 Concluding Remarks

My data show that solidarity is mainly not influenced by performance differences between needy persons or between a winner and a needy person. Because of living in a performance oriented society, it is very surprising that solidarity as prosocial behavior is unaffected by relative performance. Several subjects, who are not purely egoistical, evaluate that subjects are in a situation of distress but neither achievement justice nor inequity aversion seem to influence their donating behavior.

When investigating whether real effort per se influences donating behavior, I find that all potential recipients, except for *rank 2*, get significant higher donations from *type c* than from *rank 3*. One possible explanation for this is that *rank 3* had to bear effort costs while *type c* did not. Moreover, the fact that *types a* and *b* receive lower donations from *rank 3* than from *type c* may reflect that *rank 3* perceives it as unfair that *type a* and *b* have lower losing probabilities than he himself, although they did not have to work. Furthermore, relatively high donations of *type c* may mirror that he hopes that his gifts will implicitly be reciprocated: Because he had already bad luck with the type allocation, he may worry most to have bad luck again in losing the lottery.

To pronounce aspects of achievement justice to a greater extent in my experiment, the instructions could be framed to stress a more realistic situation of a self-inflicted distress. To investigate further possible criteria for donating decisions, one could run an extension of my experiment by implementing a higher group size and assigning ranks several times. It could then be investigated if a potential winner rank shows more solidarity with a loser of the same rank.

5.6 Appendix

Instructions of RET in German:

In dem wissenschaftlichen Experiment, an dem Sie heute teilnehmen, können Sie durch Ihre Entscheidungen Geld verdienen.

Sämtliche Entscheidungen und die Auszahlung sind anonym. Um dies sicherzustellen, agieren Sie unter einem bestimmten Codenamen, den Sie später zufällig ziehen werden. Es ist nicht möglich, Entscheidungen und Angaben, die Sie unter diesem Codenamen machen, mit Ihrer wahren Identität in Verbindung zu bringen. Die Person, die Ihnen Ihren Geldbetrag auszahlt, kennt den Ablauf des Experiments und die Instruktionen nicht. Somit kann kein Rückschluss auf Ihr Verhalten gezogen werden.

Während des gesamten Experiments ist keine Kommunikation mit den anderen Teilnehmern erlaubt. Haben Sie während des Experiments Fragen, so wenden Sie sich bitte direkt an uns!

Ablauf des Experiments:

Der erste Teil des Experiments besteht aus einer Aufgabe. Während diese Aufgabe ausgewertet wird, erhalten Sie einen Fragebogen, den Sie bitte ausfüllen. Im zweiten Teil des Experiments müssen Sie eine Entscheidung treffen. Danach wird eine Lotterie gespielt, in der Sie einen Betrag gewinnen oder verlieren können. Anschließend erhalten Sie erneut einen Bogen, den Sie bitte ausfüllen.

Ihre Entscheidung und Ihre weiteren Angaben tüten Sie bitte in einen Briefumschlag ein, den Sie in eine dafür vorgesehene Box einwerfen, mit der wir zu Ihren Kabinen kommen.

1. Teil:

Aufgabe:

Ihre Aufgabe besteht darin, so viele Plastikperlen wie möglich auf einen Faden aufzufädeln. Hierfür haben Sie 10 Minuten Zeit. **Bitte beginnen Sie erst, wenn wir Ihnen ein Zeichen geben und bitte beenden Sie die Aufgabe sofort, wenn wir Sie dazu anweisen!** Damit die aufgefädelten Perlen nicht vom Faden rutschen, sind an einer Seite des Fadens eine Büroklammer und ein Schildchen mit Ihrem Codenamen befestigt. Sie finden auch einen Aufkleber in Ihrer Kabine. Bitte kleben Sie nach Beendigung der Aufgabe diesen so um das andere Ende des Fadens,

dass die Perlen nicht herunterrutschen können. Wir werden dann die Perlenschnüre einsammeln und auswerten.

Die Ergebnisse der Aufgabe bestimmen Ihre Entlohnung aus der Aufgabe und haben Einfluss auf die Lotterie im zweiten Teil des Experiments. Bevor Sie mit der Aufgabe beginnen, werden zufällig 3er Gruppen gebildet, die während des gesamten Experiments bestehen bleiben. Sie erfahren nie die Identität der anderen Gruppenmitglieder.

Anhand der Ergebnisse der Aufgabe werden innerhalb einer Gruppe Ränge verteilt. Hat eine Person die meisten Perlen aufgefädelt, belegt sie Rang 1, hat sie am wenigsten Perlen aufgefädelt, belegt sie Rang 3. Belegen Sie in Ihrer Gruppe Rang 1, erhalten Sie **9 Euro** aus der Aufgabe. Belegen Sie Rang 2, erhalten Sie **5 Euro**. Belegen Sie Rang 3, erhalten Sie **1 Euro**.

Diese Auszahlungen erhalten Sie jedoch nur, wenn Sie nicht die Person sind, die bei der anschließenden Lotterie verliert.

Sollten mehrere Personen in einer Gruppe gleich viele Perlen aufgefädelt haben, wird gelost, welche Person den höheren Rang belegt. Sollte das der Fall sein, wird Ihnen dies auf Ihrem Entscheidungszettel mitgeteilt. Finden Sie keine Information hierüber auf Ihrem Entscheidungszettel, hat niemand in Ihrer Gruppe dieselbe Perlenanzahl aufgefädelt und es musste nicht gelost werden.

Der von Ihnen belegte Rang beeinflusst neben Ihrer Entlohnung aus der Aufgabe die Wahrscheinlichkeit, mit der Sie in der anschließenden Lotterie verlieren (Verlustwahrscheinlichkeit).

Belegen Sie Rang 1, beträgt Ihre Verlustwahrscheinlichkeit $\frac{1}{6}$

Rang 2, beträgt Ihre Verlustwahrscheinlichkeit $\frac{2}{6}$

Rang 3, beträgt Ihre Verlustwahrscheinlichkeit $\frac{3}{6}$.

Fragebogen:

Während wir die Ergebnisse der Aufgabe auswerten, bekommen Sie einen Fragebogen in einem offenen Briefumschlag. Bitte tragen Sie auf dem Bogen zunächst Ihren Codenamen ein und füllen Sie dann den Bogen aus. Wenn Sie fertig sind, legen Sie bitte den Bogen wieder zusammengefaltet in den Briefumschlag. Wir werden dann die Bögen einsammeln, indem wir mit einer Box zu Ihnen kommen, in die Sie bitte Ihren Briefumschlag einwerfen.

2. Teil:

Lotterie:

Sie nehmen an der oben bereits erwähnten Lotterie teil. In dieser können Sie **zusätzlich** zum Einkommen aus der Aufgabe **7 Euro** gewinnen. Verlieren Sie, wird Ihr Einkommen aus der Aufgabe auf **Null** reduziert!

Die Lotterie bestimmt immer zwei Gewinner und einen Verlierer pro Gruppe. Dabei verliert Rang 1 einer Gruppe mit Wahrscheinlichkeit $\frac{1}{6}$, Rang 2 mit Wahrscheinlichkeit $\frac{2}{6}$ und Rang 3 mit Wahrscheinlichkeit $\frac{3}{6}$.

Hierbei wird durch Würfeln bestimmt, wer in der Lotterie verliert. Wird eine 1 gewürfelt, verliert Rang 1. Wird eine 2 oder eine 3 gewürfelt, verliert Rang 2. Wird eine 4, 5 oder 6 gewürfelt, verliert Rang 3.

Ihre Entscheidung:

Bevor die Lotterie gespielt wird, erhalten Sie einen weiteren Briefumschlag mit einem Entscheidungszettel. Auf diesem wird Ihnen Ihr Rang aus der Aufgabe und Ihr Gesamteinkommen im Gewinnfall mitgeteilt.

Sie müssen entscheiden, wie viel Sie *im Falle eines Gewinns* bereit sind, freiwillig an *den Verlierer* Ihrer Gruppe zu transferieren. Sollten Sie zu den Gewinnern der Gruppe gehören, ist einer der beiden anderen Ränge der Verlierer. Bitte geben Sie deshalb für beide anderen Ränge an, wie viel Sie an den entsprechenden Rang transferieren möchten, falls dieser Rang in der Lotterie verliert und Sie die Lotterie gewinnen werden. Der Transfer kann zwischen Null Euro und dem Einkommen, das Sie durch die Entlohnung der Aufgabe und durch die Lotterie verdient haben, liegen. Nach der Lotterie wird einer der beiden Gewinner zufällig für die Transferentscheidung ausgewählt, der dem Verlierer tatsächlich etwas transferiert!

Werden Sie hier zufällig nicht für die Transferentscheidung ausgewählt, so werden Ihre angegebenen Transferentscheidungen nicht ausgeführt. Werden Sie zufällig für die Transferentscheidung ausgewählt, so wird Ihre Transferentscheidung für den Rang, der *tatsächlich verloren* hat, überwiesen. Ihre andere Transferentscheidung für den Rang, der nicht verloren hat, tritt nicht in Kraft.

Verlieren Sie selbst in der Lotterie, so transferieren Sie keinen der angegebenen Beträge.

Ihr angegebener Transfer wird also nur in dem Fall überwiesen, in dem Sie gewinnen und zufällig für die Transferentscheidung ausgewählt werden. Außerdem wir in diesem Fall nur der Transfer an denjenigen Rang ausgeführt, der tatsächlich verloren hat. Die Entscheidungen des anderen Gewinners und Ihre Entscheidung bezüglich des Ranges, der nicht verloren hat, werden in diesem Fall nicht ausgeführt.

<u>Der Entscheidungsprozeß</u>:

Der Entscheidungszettel sieht folgendermaßen aus:

Codename:

Sie belegen innerhalb Ihrer Gruppe Rang …

Falls Sie gewinnen, beträgt Ihr Einkommen aus der Aufgabe und der Lotterie … Euro.

Stellen Sie sich folgende Situation vor:

Angenommen, Sie gewinnen in der Lotterie und der Verlierer ist Rang …

Wie viel Euro möchten Sie an Rang … transferieren?

_____ Euro

Angenommen, Sie gewinnen in der Lotterie und der Verlierer ist Rang …

Wie viel Euro möchten Sie an Rang … transferieren?

_____ Euro

Nachdem Sie den Entscheidungszettel ausgefüllt haben, legen Sie ihn bitte wieder in den Briefumschlag und kleben Sie den Umschlag fest zu. Wir kommen anschließend mit einer weiteren Box zu Ihnen und sammeln den Umschlag ein.

<u>Auszahlung</u>:

Ihre Auszahlung aus dem Experiment hängt davon ab, welchen Rang Sie in Ihrer Gruppe belegen und ob Sie in der Lotterie gewinnen oder verlieren. Für den Fall, dass Sie gewinnen, hängt Ihre Auszahlung außerdem von Ihrer Transferentscheidung für den Verlierer der Gruppe ab und davon, ob Sie zufällig für die Transferentscheidung ausgewählt werden oder nicht. Verlieren Sie in der Lotterie, so hängt Ihre Auszahlung von der Transferentscheidung desjenigen Gewinners ab, der zufällig für die Transferentscheidung ausgewählt wurde.

Die folgende Tabelle dient der Veranschaulichung:

	Sie werden zufällig für die Transferentscheidung ausgewählt	*Sie werden zufällig nicht für die Transferentscheidung ausgewählt*
Sie sind ein Gewinner der Lotterie	Entlohnung aus der Aufgabe anhand Ihres Ranges + 7 Euro – Ihr Transfer an den Verlierer Ihrer Gruppe	Entlohnung aus der Aufgabe anhand Ihres Ranges + 7 Euro
Sie sind der Verlierer der Lotterie	Transfer des zufällig für die Transferentscheidung ausgewählten Gewinners	

Während die Lotterie durchgeführt wird und die Auszahlungen berechnet werden, bekommen Sie einen weiteren Briefumschlag mit einem Bogen. Tragen Sie bitte zunächst wieder Ihren Codenamen ein, und füllen Sie den Bogen aus. Nachdem Sie ihn wieder in den Briefumschlag gesteckt haben, wird dieser wieder mit einer Box eingesammelt.

References

Adams, J. and Adams P. (1961) “Realism of Confidence Judgments”, *Psychological Review*, 68, 33-45.

Adams, J. (1963) “Toward an Understanding of Inequity”, *Journal of Abnormal and Social Psychology*, 67, 422-436.

Adams, J. (1965) “Inequity in Social Exchange”, 267-299. In: Berkowitz L. (Ed.): *Advances in Experimental and Social Psychology*, New York.

Alpert, M. and Raiffa, H. (1969) “A Progress Report on the Training of Probability Assessors”, 294-305. Reprinted in: Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982): *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge.

Anderson, L. and Holt, C. (1997) “Informational Cascades in the Laboratory”, *The American Economic Review*, 87, 847-862.

Ando, M. (2004) “Overconfidence in Economic Contests”, *mimeo*.

Arkes, H.R., Dawes, R.M., and Christensen, C. (1986) “Factors Influencing the Use of a Decision Rule in a Probabilistic Task”, *Organizational Behavior and Human Decision Processes*, 37, 93-110.

Ball, S., Eckel, C., Grossman, P.J., and Zame, W. (2001) “Status in Markets”, *The Quarterly Journal of Economics*, 116, 161-188.

Barberis, N. and Thaler, R. (2003) “A Survey on Behavioral Finance", 1051-1121. In: Constantinides, G.M., Harris, M., and Stulz, R.M. (Eds.): *Handbook of the Economics of Finance*, Amsterdam.

Beniers, K. (2005) “Party Governance and the Selection of Parliamentarians”, *Tinbergen Institute Discussion Papers No. 05-080/1*.

Biais, B., Hilton, D., Mazurier, K., and Pouget, S. (2005) “Judgmental Overconfidence, Self-Monitoring and Trading Performance in an Experimental Financial Market”, *Review of Economic Studies*, 72, 287-312.

Bierhoff, H.-W. and Fetchenhauer, D. (2001) *Solidarität. Konflikt, Umwelt und Dritte Welt*, Opladen.

Bradley, G.W. (1978) “Self-Serving Biases in the Attribution Process: A Reexamination of the Fact or Fiction Question”, *Journal of Personality and Social Psychology*, 36, 56-71.

Brewer, B. (1979) “In-group Bias in the Minimal Intergroup Situations: A Cognitive-Motivational Analysis”, *Psychological Bulletin*, 86, 307-324.

Buchan, N., Johnson, E., and Croson, R. (2006) “Let’s Get Personal: An International Examination of the Influence of Communication, Culture and Social Distance on Other Regarding Preferences”, *Journal of Economic Behavior and Organization*, 60, 373-398.

Büchner, S., Coricelli, G., and Greiner, B. (2007) “Self-Centered and Other-Regarding Behavior in the Solidarity Game”, *Journal of Economic Behavior and Organization*, 62, 293-303.

Budescu, D. and Maciejovsky, B. (2005) “The Effect of Payoff Feedback and Information Pooling on Reasoning Errors: Evidence from Experimental Markets”, *Management Science*, 51, 1829-1843.

Buehler, R. and Griffin, D. (2003) “Planning, Personality, and Prediction: The Role of Future Focus in Optimistic Time Predictions”, *Organizational Behavior and Human Decision Processes*, 92, 80-90.

Buehler, R., Griffin, D., and Ross, M. (1994) “Exploring the “Planning Fallacy”: Why People Underestimate Their Task Completion Times”, *Journal of Personality and Social Psychology*, 67, 366-381.

Burrows, P. and Loomes, G. (1994) “The Impact of Fairness on Bargaining Behaviour”, *Empirical Economics*, 19, 201-221.

Camerer, C. and Lovallo, D. (1999) “Overconfidence and Excess Entry: An Experimental Approach”, *The American Economic Review*, 89, 306-318.

Cesarini, D., Sandewall, Ö., and Johannesson, M. (2006) “Confidence Interval Estimation Tasks and the Economics of Overconfidence”, *Journal of Economic Behavior and Organization*, 61, 453-470.

Charness, G., Rigotti, L., and Rustichini, A. (2007) “Individual Behavior and Group Membership”, *The American Economic Review*, 97, 1340-1352.

Chen, Y. and Li, X. (2007) “Group Identity and Social Preferences”, *mimeo.*

Cherry, T.L. (2001) “Mental Accounting and Other-Regarding Behaviour: Evidence from the Lab”, *Journal of Economic Psychology*, 22, 605-615.

Cherry, T.L., Frykblom, P., and Shogren, J.F. (2002) “Hardnose the Dictator”, *The American Economic Review*, 92, 1218-1221.

Cherry, T.L., Kroll, S., and Shogren, J.F. (2005) “The Impact of Endowment Heterogeneity and Origin on Public Good Contributions: Evidence from the Lab”, *Journal of Economic Behavior and Organization*, 57, 357-365.

Clarke, F.R. (1960) “Confidence Ratings, Second-Choice Responses, and Confusion Matrices in Intelligibility Tests”, *Journal of the Acoustical Society of America*, 32, 35-46.

Crisp, R.J. and Hewstone, M. (2000) “Crossed Categorization and Intergroup Bias: The Moderating Roles of Intergroup and Affective Context”, *Journal of Experimental Social Psychology*, 36, 357-383.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998) “Investor Psychology and Security Market under Overreactions”, *Journal of Finance*, 53, 1839-1885.

De la Rosa, L. (2005) “Overconfidence and Moral Hazard", *mimeo*.

De Long, J., Shleifer, A., Summers, L., and Waldmann, R. (1991) “The Survival of Noise Traders in Financial Markets”, *Journal of Business*, 64, 1-19.

Doise, W. and Sinclair, A. (1973) “The Categorization Process in Intergroup Relations”, *European Journal of Social Psychology*, 3, 145-157.

Downing, L.L. and Monaco, N.R. (1986) “In-Group/Out-Group Bias as a Function of Differential Contact and Authoritarian Personality”, *The Journal of Social Psychology*, 126, 445-452.

Dufwenberg, M., and Kirchsteiger, G. (2004) “A Theory of Sequential Reciprocity”, *Games and Economic Behavior*, 47, 268-298.

Eberlein, M. and Grund, C. (2006) “Ungleichheitsaversion in Prinzipal-Agenten-Beziehungen”, *Journal für Betriebswirtschaft*, 56, 133-153.

Eberlein, M., Ludwig, S., and Nafziger, J. (2008) “The Effects of Feedback on Self-Assessment“, *mimeo*.

Eberlein, M. and Przemeck, J. (2008a) “Solidarity and Performance Differences“, *mimeo*.

Eberlein, M. and Przemeck, J. (2008b) "Whom will you Choose? Collaborator Selection and Selector's Self-Prediction", *Bonn Econ Discussion Papers No. 12/2008*.

Eberlein, M. and Walkowitz, G. (2008) "Positive and Negative Team Identity in a Promotion Game", *Bonn Econ Discussion Papers No. 13/2008*.

Eckel, C. and Grossman, P. (2005) "Managing Diversity by Creating Team Identity", *Journal of Economic Behavior and Organization*, 58, 371-392.

Eichberger, J., Kelsye, D., and Schipper, B. (2005) "Ambiguity and Social Interactions", *mimeo*.

Elster, J. (1989) *The Cement of Society. A Study of Social Order*, Cambridge.

Epley, N. and Dunning, D. (2006) "The Mixed Blessings of Self-Knowledge in Behavioral Prediction: Enhanced Discrimination but Exacerbated Bias", *Personality and Social Psychology Bulletin*, 32, 641-655.

Fahr, R. and Irlenbusch, B. (2000) "Fairness as a Constraint on Trust in Reciprocity: Earned Property Rights in a Reciprocal Exchange Experiment", *Economics Letters*, 66, 275- 282.

Falk, A. and Fehr, E. (2003) "Why Labor Market Experiments?", *Labour Economics*, 10, 399-406.

Falk, A. and Fischbacher, U. (2006) "A Theory of Reciprocity", *Games and Economic Behavior*, 54, 293-315.

Fehr, E., Gächter, S., and Kirchsteiger, G. (1997) "Reciprocity as Contract Enforcement Device: Experimental Evidence", *Econometrica*, 65, 833-860.

Fellner, G., Güth W., and Maciejovsky, B. (2004) "Illusion of Expertise in Portfolio Decisions: An Experimental Approach", *Journal of Economic Behavior and Organization*, 55, 355-376.

Ferguson, C.K. and Kelley, H.H. (1964) "Significant Factors in Overvaluation of Own-Group's Product", *Journal of Abnormal and Social Psychology*, 69, 223-228.

Ferraro, P.J. (2005) "Know Thyself: Incompetence and Overconfidence", *Experimental Laboratory Working Paper Series 2003-001*, Georgia State University.

Festinger, L. (1954) "A Theory of Social Comparison Processes", *Human Relations*, 7, 117-140.

Fischbacher, U. (1999) "Z-Tree. Toolbox for Readymade Economic Experiments", *IEW Working Paper 21*, University of Zurich.

Forsythe, R., Horowitz, J.L., Savin, N.E., and Sefton, M. (1994) "Fairness in Simple Bargaining Experiments", *Games and Economic Behavior*, 6, 347-369.

Friebel, G. and Raith, M. (2004) "Abuse of Authority and Hierarchical Communication", *RAND Journal of Economics*, 35, 224-244.

Friedman, D. and Sunder, S. (1994) *Experimental Methods. A Primer for Economics*, Cambridge.

Gervais, S. and Odean, T. (2001) "Learning to Be Overconfident", *The Review of Financial Studies*, 14, 1-27.

Gigerenzer, G., Hoffrage, U., and Kleinbolting, H. (1991) "Probabilistic Mental Models: A Brunswikian Theory of Confidence", *Psychological Review*, 98, 506-528.

Glazer, A. and Segendorff, B. (2005) „Credit Claiming", *Economics of Governance*, 6, 125-137.

Goette, L., Huffman, D., and Meier, S. (2006) "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups", *IZA Discussion Papers No. 06-7*.

Graham, S. (1991) "A Review of Attribution Theory in Achievement Contexts", *Educational Psychology Review*, 3, 5-39.

Greiner, B. (2004) "An Online Recruitment System for Economic Experiments", 79-93. In: Kremer, K. and Macho, V. (Eds.): *Forschung und wissenschaftliches Rechnen*. GWDG Bericht 63. Ges. für Wiss. Datenverarbeitung, Göttingen.

Grund, C. and Sliwka, D. (2005) "Envy and Compassion in Tournaments", *Journal of Economics and Management Strategy*, 14, 187-207.

Gürtler, O. (2008) "On Sabotage in Collective Tournaments", *Journal of Mathematical Economics*, 44, 383-393.

Güth, W., Schmittberger, R., and Schwarze, B. (1982) "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization*, 3, 367-388.

Harbring, C. and Irlenbusch, B. (2008) "How Many Winners are Good to Have? On Tournaments with Sabotage", *Journal of Economic Behavior and Organization*, 65, 682-702.

Harbring, C., Irlenbusch, B., Kräkel, M., and Selten, R. (2007) "Sabotage in Corporate Contests – An Experimental Analysis", *International Journal of the Economics Business*, 14, 367-392.

Hoffman, E., McCabe, K.A., Shachat, K., and Smith, V. (1994) "Preferences, Property Rights, and Anonymity in Bargaining Games", *Games and Economic Behavior*, 7, 346-380.

Hoffman, E., McCabe, K.A., and Smith, V.L. (1996) "On Expectations and the Monetary Stakes in Ultimatum Games", *International Journal of Game Theory*, 25, 289-301.

Hoffman, E. and Spitzer, M.L. (1985) "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects` Concepts of Distributive Justice", *Journal of Legal Studies*, 14, 259- 297.

Homans, G.C. (1958) "Social Behavior as Exchange", *American Journal of Sociology*, 63, 597-606.

Homans, G.C. (1960) *Theorie der sozialen Gruppe*, Opladen.

Homans, G.C. (1968) *Elementarfaktoren sozialen Verhaltens*, Opladen.

Kraiger, K. and Ford, J.K. (1985) "A Meta-Analysis of Ratee Race Effects in Performance Ratings", *Journal of Applied Psychology, 70,* 56-65.

Kräkel, M. (2005) "Helping and Sabotaging in Tournaments", *International Game Theory Review*, 7, 211-228.

Kuiper, N.A. and Derry, P.A. (1982) "Depressed and Non-Depressed Content Self-Reference in Mild Depression", *Journal of Personality*, 50, 67-79.

Kuiper, N.A. and Mc Donald, M.R. (1982) "Self and Other Perception in Mild Depressives", *Social Cognition*, I, 233-239.

Lazear, E.P. (1989) "Pay Equality and Industrial Politics", *Journal of Political Economy*, 97, 561-580.

LePine, J.A. and Van Dyne, L. (2001) "Peer Responses to Low Performers: An Attributional Model of Helping in the Context of Groups", *Academy of Management Review*, 26, 67-84.

Levy, G. (2004) “Anti-Herding and Strategic Consultation”, *European Economic Review*, 48, 503-525.

Lichtenstein, S. and Fischhoff, B. (1977) “Do Those Who Know More also Know More About How Much They Know? The Calibration of Probability Judgements”, *Organizational Behavior and Human Performance*, 20, 159-183.

Lord, C., Ross, L., and Leppner, M. (1979) “Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence”, *Journal of Personality and Social Psychology*, 37, 2098-2109.

Ludwig, S. and Nafziger, J. (2007) “Do You Know that I Am Biased? An Experiment”, *Bonn Econ Discussion Papers No. 11/2007*.

Maccoby, E.E and Jacklin, C.N. (1974) *The Psychology of Sex Differences*, Stanford.

McLeisch, K.N. and Oxoby, R.J. (2007) “Identity, Cooperation, and Punishment”, *IZA Discussion Paper No. 2572*.

Miller, D.T. and Ross, M. (1975) “Self-Serving Biases in Attribution of Causality: Fact or Fiction?”, *Psychological Bulletin*, 82, 213-225.

Mummendey, A., Simon, B., Dietze, C., Grünert, M., Haeger, G., et al. (1992) “Categorization is not Enough: Intergroup Discrimination in Negative Outcome Allocations”, *Journal of Experimental Social Psychology*, 28, 125-144.

Ockenfels, A. and Weimann, J. (1999) “Types and Patterns: An Experimental East-West-German Comparison of Cooperation and Solidarity”, *Journal of Public Economics*, 71, 75-287.

Otten, S. and Wentura, D. (1999) “About the Impact of Automaticity in the Minimal Group Paradigm: Evidence of the Affective Priming Task”, *European Journal of Social Psychology*, 29, 1049-1071.

Oxoby, R.J. and Friedrich, C. (2006) “Trust and the Structure of Incentives”, University of Calgary Economics Working Papers, *Discussion Paper 2006-04*.

Pulford, B. and Colman, A. (1997) “Overconfidence: Feedback and Item Difficulty Effects”, *Personality and Individual Differences*, 23, 125-133.

Rabbie, J. and Horwitz, M. (1969) “Arousal of Ingroup-Outgroup Bias by a Chance Win or Loss”, *Journal of Personality and Social Psychology*, 13, 269-277.

Rabbie, J.M., Schot, J.C., and Visser, L. (1989) "Social Identity Theory: A Conceptual and Empirical Critique from the Perspective of a Behavioural Interaction Model", *European Journal of Social Psychology*, 19, 171-202.

Rabin, M. (1993) "Incorporating Fairness into Game Theory and Economics", *The American Economic Review*, 83, 1281-1302.

Rodriguez-Bailon, R., Moya M., and Yzerbyt, V. (2006) "Cuando el poder ostentado es inmerecido : sus efectos sobre la percepción y los juicios socials (English abstract title: When Power is Underserved: Its effects on Perception and Social Judgements)", *Psicothema*, 18, 194-199.

Roll, R. (1986) "The Hubris Hypothesis of Corporate Takeovers", *Journal of Business*, 59, 197-216.

Ross, L. and Leppner, M.R. (1980) "The Perseverance of Beliefs: Empirical and Normative Considerations", 117-136. In: Schweder, R.A. and Fiske, D. (Eds.): *New Directions for Methodology of Behavioral Science: Fallible Judgment in Behavioral Research*, San Francisco.

Ruffle, B. and Sosis, R. (2006) "Cooperation and the In-group-out-group Bias: A Field Test on Israeli Kibbutz Members and City Residents", *Journal of Economic Behavior and Organization*, 60, 147-63.

Ruffle, B.J. (1998) "More is Better, but Fair is Fair: Tipping in Dictator and Ultimatum Games", *Games and Economic Behavior*, 23, 247-265.

Russo, J.E. and Schoemaker, P.J.H. (1992) "Managing Overconfidence", *Sloan Management Review*, 33, 7-17.

Rutström, E.E. and Williams, M.B. (2000) "Entitlements and Fairness: An Experimental Study of Distributive Preferences", *Journal of Economic Behavior and Organization*, 43, 75-89.

Schaller, M. (1992) "In-Group Favoritism and Statistical Reasoning in Social Inference: Implications for Formation and Maintenance of Group Stereotypes", *Journal of Personality and Social Psychology*, 60, 61-74.

Segendorff, B. (2000) "A Signalling Theory of Scapegoats", Stockholm School of Economics, *SSE/EFI Working Paper Series in Economics and Finance No. 406*.

Selten, R. (1967) “Die Strategiemethode zur Erforschung eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes”, 136-168. In: Sauermann, H. (Ed.): *Beiträge zur Experimentellen Wirtschaftforschung*, Tübingen.

Selten, R. and Ockenfels, A. (1998) “An Experimental Solidarity Game”, *Journal of Economic Behavior and Organization*, 34, 517-539.

Sharp, G., Cutler, B., and Penrod, S. (1988) “Performance Feedback Improves the Resolution of Confidence Judgments”, *Organizational Behavior and Human Decision Processes*, 42, 271-283.

Sherif, M., Harvey, O.J., White, B.J., Hood, W.R., and Sherif, C.W. (1961) *Intergroup Conflict and Cooperation: The Robbers Cave Experiment,* Oklahoma.

Solow, J.L. and Kirkwood, N. (2002) “Group Identity and Gender in Public Goods Experiments”, *Journal of Economic Behavior and Organization*, 48, 403-412.

Suls, J. and Wheeler, L. (2000) *Handbook of Social Comparison. Theory and Research*, New York.

Svenson, O. (1981) “Are We All Less Risky and More Skilful than our Fellow Drivers?”, *Acta Psychologica*, 47, 143-148.

Tajfel, H., Billig, M.G., and Bundy, R.P. (1971) “Social Categorisation and Intergroup Behaviour”, *European Journal of Social Psychology*, 1, 149-178.

Tajfel, H. and Turner, J.C. (1979) “An Integrative Theory of Intergroup Conflict”, 33-47. In: Austin, W.G. and Worchel, S. (Eds.): *The Social Psychology of Intergroup Relations*, Monterey.

Taylor, S.E. and Brown J.D. (1988) “Illusion and Well-Being: A Social Psychological Perspective on Mental Health”, *Psychological Bulletin*, 103, 193-210.

Thaler, R. (1988) “The Winner's Curse”, *Journal of Economic Perspectives*, 2, 191-202.

Thral, N. and Radermacher, R. (2006) “Bad Luck versus Self-Inflicted Neediness – An Experimental Investigation of Gift Giving in a Solidarity Game” University of Cologne, *Working Paper Series in Economics No. 28*.

Weiner, B. (1986) *An Attributional Theory of Motivation and Emotion*, New York.

Weinstein, N.D. (1980) “Unrealistic Optimism About Future Life Events”, *Journal of Personality and Social Psychology*, 39, 806-820.

Wit, A.P. and Wilke, H.A.M. (1992) “The Effect of Social Categorization on Cooperation in Three Types of Social Dilemmas”, *Journal of Economic Psychology*, 13, 135-151.

Yamagishi, T., Nobuhito, J., and Kiyonari, T. (1999) “Bounded Generalized Reciprocity: Ingroup Boasting and Ingroup Favoritism”, 161–197. In: Thye, S.R., Lawler, E.J., Macy, M.W., and Walker H.A. (Eds.): *Advances in Group Processes*, Stanford.

Young, H.P. (1974) “An Axiomazation of Borda’s Rule”, *Journal of Economic Theory,* 9, 43-52.

Young, P. (1995) “Optimal Voting Rules”, *Journal of Economic Perspectives*, 9, 51-64.