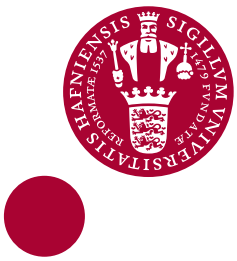

Subjective Performance Evaluations, Self-esteem, and Ego-threats in Principal-agent Relations

Sebald, Alexander Christopher; Walzl, Markus

Publication date:
2010

Document version
Publisher's PDF, also known as Version of record

Discussion Papers
Department of Economics
University of Copenhagen

No. 10-18

Subjective Performance Evaluations, Self-esteem, and Ego-threats in Principal-agent Relations

Alexander Sebald, Markus Walzl

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Subjective performance evaluations, self-esteem, and ego-threats in principal-agent relations

By Alexander Sebald and Markus Walzl*

We conduct a laboratory experiment with agents working on and principals benefiting from a real effort task in which the agents' effort/performance can only be evaluated subjectively. Principals give subjective performance feedback to agents and agents have an opportunity to sanction principals. We find that agents sanction whenever the feedback of principals is below their subjective self-evaluations even if the agents' payoff is independent of the principals' feedback. Based on our experimental analysis we propose a principal-agent model with subjective performance evaluations that accommodates this finding. We analyze the agents' (optimal) behavior, optimal contracts, and social welfare in this environment.
JEL:D01; D02; D82; D86; J41.
Keywords: Contracts, Subjective Performance Evaluations, Self-Esteem, Ego-Threats.

Providing performance feedback and creating incentives through performance pay is an integral part of numerous variants of social and economic interaction. For example, teachers regularly grade the performance of their students and give feedback, employers regularly evaluate the performance of their employees, give feedback and pay for performance. To capture performance in a purely objective way is very often hard to accomplish as a lot of valuable information about performance is captured by subjective impressions rather than objective measures. For example, spelling out tenure or hiring criteria purely on the basis of objective measures such as the number (and impact factor) of publications, the amount of research funds, or the number of supervised PhD-students hardly allows for an assessment of the applicant's qualities as a teacher, his "good citizenship", his dedication to research collaborations, or his scientific potential – issues typically assessed in the appraisal of a tenure committee or a letter of recommendation. In general, as it is costly and difficult to specify all (objective) contingencies in a contract, it is often preferred to leave (part of the) performance feedback to a more holistic subjective appraisal by e.g. a supervisor or a tenure committee who typically not only aggregate objective information but decide subjectively with considerable discretionary leeway.

Numerous contributions in economics have recognized the prevalence and importance of subjective performance evaluations and corresponding performance pay in labor market relations [see e.g. Gibbs et al. (2004), Ittner et al. (2003), Levine (2003), Milkovich and Wigdor (1991), Murphy & Oyer (2003), and Prendergast (1999)]. It has been highlighted that there is a trade-off between more holistic measures of performance [see Prendergast

* Sebald: Department of Economcs, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, Denmark. alexander.sebald@econ.ku.dk.Walzl: Department of Economics, Bamberg University, Feldkirchenstrasse 21, 96045 Bamberg, Germany. markus.walzl@uni-bamberg.de. Financial support by METEOR and the Department of Economics at the University of Copenhagen is gratefully acknowledged. Furthermore, we are grateful to Charles Bellemare, Daniel le Maire and seminar participants at the University of Copenhagen and Université Laval for very helpful comments on an earlier version of this paper. Lastly, we are grateful to Ralph Ferré Koch for programming the experimental software used in this research project. All remaining error are ours.

(1999)] and the potential for conflict due to subjective and, hence, possibly diverging performance assessments by principals and agents [see McLeod(2003)]. Furthermore, associated with the potential for conflict, a central insight in this literature is that employment relations based on subjective performance evaluations are particularly fragile in one-shot interactions. If labor contracts specify payments on the basis of the principals' subjective appraisals, principals have an incentive to claim that performance was poor in order to establish low wages. As a consequence, inefficiently low effort may be spend by agents – unless principals can credibly commit to an honest revelation of their subjective information as, for instance, in repeated interaction or with a credible payment to a third party [see e.g. Levine (2003) or McLeod (2003)].

In contrast to this, we demonstrate in our analysis that the credibility of a truthful revelation of the principals' subjective performance signal can also be established in one-shot interactions and without a third party as agents tend to create 'costly conflicts' for principals if their own subjective performance appraisals are better than the principals' feedback. That is, we first experimentally show that agents are willing to forgo own payoff to create 'costly conflicts', if their own subjective performance appraisals are better than the principals' and second, formalize and analyze a simple principal-agent model that incorporates our experimental findings.

Specifically, our laboratory experiment matches participants into pairs and randomly assigns them to one of two roles, principal or agent. The agent has to work on a real effort task. The task is such that both agent and principal only get a subjective signal about the agent's performance. The principal benefits from the effort of the agent and gives performance feedback. In reaction to the feedback, the agent has the opportunity to reduce the principal's payoff at a cost for himself. In our experiment we find that agents' reactions to principals' feedback strongly depend on their self-perceptions. Agents reduce payoffs of principals, if the principals' feedback is below their self-perception, but accept the feedback and refuse to reduce payoffs if the feedback confirms/is higher than their own evaluation.

This pattern can be observed in an *incentive* treatment where the principal's feedback determines the agent's payoff *and* in a *flat* treatment where the agent's payoff is constant and thereby unaffected by the feedback. The willingness to reduce payoffs in the *incentive* treatment can be explained by distributional concerns as e.g. inequity aversion or reciprocity. As payoffs are according to the principal's feedback, the agent may consider a feedback below his own evaluation as a decision that generates an unequal payoff distribution, or as an unkind act by the principal whose willingness to pay falls short of the agent's expectations or what the agent feels entitled to. However, the results of *incentive* and *flat* treatment taken together rather suggest a motivation for payoff reductions that is not driven by the payoff consequences of the principal's feedback but by the tension between subjective feedback and the agent's self-evaluation as such. The fact that individuals dislike a tension between feedback and self-perception, and regard it as a threat to their self-esteem, i.e. an *ego-threat*, that triggers aggressive behavior or conflict as a protection mechanism is a central finding in social psychology. As shown *e.g.* by Greenwald (1980), Bushman & Baumeister (1998) and Baumeister (2005)], people seem to care about their self-esteem and try to enhance, maintain and protect it.

Based on our experimental findings and the corresponding insights from social psychology, we present a simple principal-agent model that captures the agent's eagerness to protect his self-esteem against ego-threats. Within this model we analyze the dependence of the agent's (optimal) behavior, optimal contracts, and social welfare on the agent's sensitivity towards ego-threats, the nature of conflict created by the agent, and the quality of subjective performance evaluations. We demonstrate that an increase in the level of conflict that the agent can impose on the principal or the sensitivity of the agent towards

ego-threats enhances welfare if the effort of the agent is sufficiently valuable (i.e., if the agent works on a valuable project) even if conflict and the dis-utility from ego-threats are costly from a welfare point of view. The reason is that an enhanced level of conflict or sensitivity increases the maximum effort that can be implemented by the principal which is a binding constraint for valuable projects. Hence, in particular for these projects, high levels of potential conflict and a high sensitivity of employees towards ego-threats should be appreciated by employers as they facilitate a credible commitment towards a truthful revelation of the principals' performance signals and thereby help establishing incentives to induce high effort levels. Therefore, employers facing the problem of subjective performance evaluations in their relation to employees should not per-se try to select a work-force with a low sensitivity towards ego-threats or without independent judgment, or try to eliminate opportunities of conflict creation.

In recent years also economists have started to acknowledge the importance of self-esteem in decision making and strategic interactions [*e.g.* Elling- sen & Johannesson (2008), Compte & Postlewaite (2004), Bénabou & Tirole (2002), Köszegi (2006)]. It is argued that people strive for positive self-perceptions because they entail a consumption and motivational value. For example, Köszegi (2006) endows individuals with *ego-utility* and demonstrates its effect on choices between more or less ambitious tasks. In particular, this model explains the phenomenon of overconfidence by individuals who update believes according to Bayes' rule. Bénabou & Tirole (2002), Compte & Postlewaite (2004), and Ellingsen & Johannesson (2008) on the other hand, center on the motivational value of self-confidence. It is argued that confidence in one's ability and efficacy or a psychological payoff from being esteemed by others can help individuals to undertake more ambitious tasks. When people have imperfect knowledge about their own ability and/or when effort and ability are complements, then more self-confidence enhances peoples' motivation to act [Bénabou & Tirole (2002): 873].

As said before, however, psychologists have not only identified the implicit impact of self-esteem on information processing and motivation, but also stress people's eagerness to actively maintain and protect positive self-perceptions. First, people protect their self-esteem by systematically taking credit for success and denying blame for failure and second, people have a tendency to uncritically accept positive feedback and eagerly search for flaws/faults in other's criticism [*e.g.* Baumeister (2005), Greenwald (1980)]. Third and most importantly for our investigation, psychologists have found that conflicts and aggression tend to result from positive self-images that are challenged or threatened [*e.g.* Baird (1977), Raskin et al (1991), Bushman & Baumeister (1998)]. It is argued that hostile aggression is an expression of the self's rejection of ego-threatening evaluations received from other people [*e.g.* Baumeister et al (1996)]. People with high self-esteem usually hold confident and highly favorable ideas about themselves, *i.e.* they exhibit ego-involvement, and react belligerently to ego-threatening feedback from others [Baird (1977), Shrauger & Lund (1975) and Korman (1969)]. Our contribution demonstrates the significance of these effects in an incentivized laboratory experiment and proposes a simple model that formalizes these findings.

In the following section we present the set-up of our experiment. In section II, we discuss our experimental findings regarding conflict creation. In section III, we present a principal-agent model with subjective performance evaluations that includes preferences for conflict creation in-line with our experimental findings. Section IV concludes with some remarks on the practical implications of our analysis.

I. Experimental Set-up

In this experiment we investigate individual reactions to performance feedback in environments in which people only have subjective performance information. The experiment

took place in June and November 2009 in the laboratory of the Center for Experimental Economics at the University of Copenhagen with in total 186 participants who completed the experiment.[1] We conducted two treatments, *incentive* and *flat*, each consisting of four experimental sessions. On average participants took 45 minutes to complete the experiment and received about 110 DKK ($\sim$ 15 Euros).

At the beginning of the experiment, all participants were randomly assigned to a group and one of two different roles labeled *Person A* and *Person B*. Each group consisted of one Person A and one Person B. Participants were provided with experimental instructions (see Appendix C). After reading the instructions, participants took actions at four different stages: **i)** *control questions*, **ii)** *clicking-task*, **iii)** *evaluation and feedback* and **iv)** *reaction*.

In stage **i)** (control questions), all participants had to answer a set of control questions before being able to proceed (for the corresponding screen-shots see Appendix D).

In stage **ii)** (clicking-task), participants in the role of Person B had to work on a real-effort task (i.e. they acted as 'agents'). The real-effort task consisted of clicking away boxes on a computer screen (for a screen-shot of the clicking task see Appendix D). For a period of x seconds, 20 screens with boxes appeared for various time intervals (i.e. between 3 and 9 seconds). At the end of each time interval the screen disappeared with the remaining (i.e. un-clicked) boxes and a new screen with a new set of boxes popped up. In order to create heterogeneity in B-Persons' self evaluations, we had one session in which $x = 120$, two sessions in which $x = 90$ and one session in which $x = 50$ in each treatment.[2] Person A saw the same screen as Person B and could observe him clicking away the boxes (i.e., Person A acted as a 'principal').

In both treatments, Person A's payoff was determined by the percentage of boxes clicked away by Person B during the clicking-task.[3] If

- Person B clicked away 0-20% of the boxes: Person A received 200 points.
- Person B clicked away 20-40% of the boxes: Person A received 300 points.
- Person B clicked away 40-60% of the boxes: Person A received 400 points.
- Person B clicked away 60-80% of the boxes: Person A received 500 points.
- Person B clicked away 80-100% of the boxes: Person A received 600 points.

At stage **iii)** (evaluation and feedback), both participants were asked to evaluate B's performance by telling the percentage of boxes that B clicked away (i.e., both participants had to state one of the five categories 0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Furthermore, Person A was asked to give feedback to Person B with the same categories. In the *incentive* treatment, Person B's payoff depended on Person A's feedback as follows:

- Person A's feedback 0-20%: Person B received 100 points from A.
- Person A's feedback 20-40%: Person B received 150 points from A.
- Person A's feedback 40-60%: Person B received 200 points from A.
- Person A's feedback 60-80%: Person B received 250 points from A.
- Person A's feedback 80-100%: Person B received 300 points from A.

[1] In total 190 persons participated but 4 participants (2 groups) did not complete the experiment due to a technical problem. The analysis is based on the 186 individuals that completed the experiment.

[2] Note that the 20 screens were the same in all sessions. We only varied the number of seconds that the screen was shown.

[3] In the instructions (see Appendix C), we informed participants about the payoff scheme. Payoffs in the experiment were expressed in points and participants were informed at the beginning of the experiment that points were exchanged into Danish crowns at the end of the experiment at an exchange rate of 10 points = 3.5 DKK. For a summary statistic concerning the number of participants per treatment/session, number of appeared boxes/average number of boxes clicked away etc see Appendix E.

In the *flat* treatment, on the other hand, Person B's payoff was 200 points independent of Person A's feedback.

At stage **iv)** (reaction), Person B was able to react to Person A's feedback with a reduction of Person A's payoff by up to 100 points. To elicit reaction behavior, we used the strategy method: while Person A was giving feedback, we asked Person B to indicate for each possible feedback that he could receive by how much he would like to reduce Person A's payoff. Hence, for each possible feedback level (0-20%, 20-40%, 40-60%, 60-80%, 80-100%) Person B had to state between 0 and 100 points by which he wanted to reduce Person A's payoff in case this was Person A's actually stated feedback. For every point that Person B reduced Person A's payoff, Person B had to pay 0.25 points.

After stage **iv)** and a small questionnaire, Person B's real performance, Person A's feedback and Person B's reaction to Person A's actual feedback was used to calculate payoffs. Finally, participants were shown the actual performance of Person B, Person A's feedback, Person B's reaction and the actual payoffs on their screen. Note that at the evaluation and feedback as well as the reaction stage, Person A and B decided on the basis of their *subjective* perception of Person B's performance. Only in the end of the experiment when payoffs were listed, participants learned about B's actual performance.

II. Conflict Creation

A. Experimental Observations

Using the strategy method, we elicited the self-perception ('own evaluation') of each B-Person and the number of points that he wanted to reduce Person A's payoff ('payoff reduction'), if Person A's feedback was 0-20%, 20-40%, 40-60%, 60-80% and 80-100%. Table 1 and 2 display the median payoff reduction in the *incentive* and the *flat* treatment, respectively.

[Tables 1 and 2 here]

Tables 1 and 2 indicate that in both treatments the median payoff reduction at feedback levels below own evaluation is positive. Furthermore, for the *incentive* treatment, the median payoff reduction increases in the gap between feedback and own evaluation of B-Persons. In contrast to this, at feedback levels equal and above own evaluations the median payoff reduction is 0 in both treatments (with two exceptions: the payoff reduction of people with own evaluation 40-60 and 60-80 at feedback levels 40-60 and 60-80, respectively, in the *incentive* treatment). Hence, the median B-Persons in our experiment reduce Person A's payoff if they receive a feedback from Person A that falls short of their own evaluation but typically refuse to reduce Person A's payoff after feedback that confirms / is above their own evaluation – regardless of whether the payoff that Person B receives is dependent or independent of the feedback that Person A gives.

B. Testable Hypotheses

As a payoff reduction is costly for Person B and the interaction between Person A and B is one-shot, assuming selfishness and rationality would certainly yield the prediction of no payoff reduction in both treatments. We will refer to this case as Hypothesis 0. The above-mentioned occurrence of payoff reductions, however, suggests that individuals acting as Person B sense a non-monetary motivation to reduce payoffs.

The economic literature emphasizes payoff-related motives for costly payoff reductions such as distributional concerns [e.g., inequity aversion as in Fehr & Schmidt (1999) or Bolton & Ockenfels (2000)], and reciprocity [as in Rabin (1993), Dufwenberg & Kirchsteiger (2004), or Falk & Fischbacher (2006)]. In models of reciprocity, there exist an

Table 1—: Median payoff reduction: *incentive* treatment

Feedback	**Own Evaluation**					
	0-20	20-40	40-60	60-80	80-100	
0-20	...	20	50	80	100	
20-40	...	0	30	50	80	
40-60	...	0	10	27.5	60	
60-80	...	0	0	12	10	
80-100	...	0	0	0	0	
Total No:	0	14	15	12	1	Sum: 42

Table 2—: Median payoff reduction: *flat* treatment

Feedback	**Own Evaluation**					
	0-20	20-40	40-60	60-80	80-100	
0-20	0	30	25	20	80	
20-40	0	0	10	10	60	
40-60	0	0	0	10	40	
60-80	0	0	0	0	1	
80-100	0	0	0	0	0	
Total No:	6	7	22	13	3	Sum: 51

In Tables 1 and 2 each row (beside the last) and column correspond to a feedback level and a level of own evaluation, respectively. The last row indicates the number of B-Persons' that have levels of own-evaluation 0-20%, 20-40%, 40-60% etc. For example, there are 14 B-Persons with an own evaluation of 20-40% in the *incentive* treatment. In total we have 42 B-Persons in the *incentive* and 51 B-Persons in the *flat* treatment (the asymmetry is induced by non-show ups). Each row indicates the median payoff reduction B-Persons with a certain own-evaluation choose at this specific feedback level. For example, the median payoff reduction of B-Persons with an own evaluation of 20-40 at a feedback level of 0-20 and 20-40 in the *incentive* treatment is respectively 20 and 0 points.

(endogenous) reference point against which people judge the kindness of their own as well as other people's actions. In our setting, Person B's own evaluation and the resulting expectation concerning what he is 'entitled to' may serve as such a reference point. If Person B takes his own evaluation as a reference point for the compensation he feels entitled to, he might feel unkindly treated by Person A, if his payoff is less than according to his own evaluation [for a discussion of reference dependent preferences see e.g. Köszegi & Rabin (2006) and Abeler et al. (2009)].

For the *incentive* treatment, Person A's feedback determines the payoff for Person B (and may contain information about Person B's performance) while Person B's own evaluation is a subjective signal about his performance (and thereby also Person A's actual payoff). For the following derivation of the hypotheses, we assume that Person B considers his own signal at least as informative as Person A's feedback. For a given feedback level, the payoff distribution is therefore expected to be equal in the *incentive* treatment if Person B's own evaluation is equal to the feedback and expected to be unequal in favor of Person A if the feedback is below Person B's own evaluation. If payoff reductions are motivated by inequity aversion or (negative) reciprocity, we thus expect (weakly) larger payoff reductions if the feedback is below Person B's own evaluation than if feedback and own evaluation coincide. As a feedback above Person B's own evaluation yields an expected payoff distribution in favor of Person B, inequity aversion would not predict any payoff reduction in this case (recall that one unit of reduced payoff for Person A only costs 1/4 unit for Person B). Similarly, reciprocity would not predict larger payoff reductions than for a coincidence of feedback and own evaluations.[4] We summarize as follows:

HYPOTHESIS 1: *In the incentive treatment, for a given feedback, payoff reductions are smaller if the feedback is above/equal rather than below Person B's own evaluation.*

In the *flat* treatment, Person B's payoff is independent of Person A's feedback. Hence, Person A's feedback can neither be regarded as a kind nor an unkind act regarding monetary outcomes. Therefore, reciprocity would not predict any payoff reductions.

HYPOTHESIS 2: *In the flat treatment, no payoff reduction is observed.*

In contrast, if Person B is inequity averse, he may well have an incentive to reduce payoffs. For a given own evaluation of Person B, however, these incentives are either independent of Person A's feedback (if the feedback is considered as uninformative regarding the actual payoff distribution) or increasing in Person A's feedback (if the feedback is considered as informative regarding the actual payoff distribution and higher feedback levels imply higher expected payoff for Person A).

HYPOTHESIS 3: *In the flat treatment, for a given evaluation of Person B, payoff reductions are (weakly) larger if the feedback is above/equal rather than below Person B's own evaluation.*

In contrast to the above-mentioned payoff related motives and as discussed in the introduction, the psychological literature rather emphasizes the tension between Person

[4]Note that reciprocity could certainly explain payoff reductions if the feedback level is above Person B's own evaluation. For example, consider an individual that regards any payoff kept by Person A as unkind or feels entitled to the entire payoff generated by his effort. Such an individual would also have incentives to reduce payoffs if the feedback is above his own evaluation but still leaves some (expected) payoff to Person A. As we do not observe this kind of payoff reductions, we drop this case from our analysis.

B's own evaluations and Person A's feedback as such. If Person B regards a feedback below his own evaluation as a harmful ego-threat, while no ego-threat is perceived if feedback and self-evaluation coincide or the former exceeds the latter, and if Person B prefers to initiate costly conflict as ego-protection, then psychological costs (and the motivation to reduce payoffs) are larger if the feedback is below Person B's own evaluation rather than equal of above *in both treatments*. This resembles Hypothesis 1 for the *incentive* treatment but yields new hypothesis for the *flat* treatment.[5]

HYPOTHESIS 4: *In the flat treatment, for a given evaluation of Person B, payoff reductions are smaller if the feedback is above/equal rather than below Person B's own evaluation.*

To summarize, (costly) payoff reductions cannot be explained on the basis of rational, selfish behavior. However, several assumptions on complex preferences offer an explanation for payoff reductions. While all of these assumptions coincide regarding the hypothesis for the *incentive* treatment, different predictions are formed for the *flat* treatment. In particular, models of inequity aversion and reciprocity predict either no or a positive impact of feedback on payoff reductions in the *flat* treatment, while the protection against ego-threats suggests a negative relation (i.e., lower feedback – for a given self-evaluation – increases payoff reduction). In the next section, we will test the corresponding hypotheses.

C. Data Analysis

For the tests, we consider each feedback level separately and ask whether the behavior of B-Persons for whom this feedback level lies *Below* their own-evaluation significantly differs from B-Persons for whom this feedback level is *Equal/Above* their own evaluation.

[Tables 3 and 4 here]

Table 3—: Median payoff reduction per feedback level: *incentive* treatment

Feedback	**Feedback vs. Own Evaluation**			**WMW-Test**
	Equal/Above	*Below*	Diff	P-Value
0-20	- (0)	50 (42)	-	-
20-40	0 (14)	35 (28)	35	(0.001)
40-60	0 (29)	40 (13)	40	(0.043)
60-80	0 (41)	10 (1)	10	(0.439)
80-100	0 (42)	- (0)	-	-

Looking first at the results of the *incentive* treatment (see Table 3) shows that for feedback levels below 80-100 the median payoff reduction of B-Persons for whom the feedback level is below (i.e. *Below*) their own evaluation is higher than the median payoff reduction of B-Persons for whom the feedback level is confirming or above (i.e. *Equal/Above*) their own evaluation. Note that, as in the *incentive* treatment there is no

[5]Note that the protection against ego-threats can also be interpreted as a reciprocal mechanism. The difference to the aforementioned form of reciprocity, however, is that perceived unkindness and related psychological cost do not depend on the payoff consequence of the feedback, but rather on the tension between own evaluation and feedback as such.

Table 4—: Median payoff reduction per feedback level: *flat* treatment

Feedback	**Feedback vs. Own Evaluation**			**WMW-Test**
	Equal/Above	*Below*	Diff	P-Value
0-20	0 (6)	30 (45)	30	(0.053)
20-40	0 (13)	10 (38)	10	(0.055)
40-60	0 (35)	15 (16)	15	(0.019)
60-80	0 (48)	1 (3)	1	(0.2415)
80-100	0 (51)	- (0)	-	-

In Tables 3 and 4 we report for each feedback level the median payoff reduction of B-Persons for whom this feedback level is *Equal/Above* their own evaluation and for whom this feedback level is *Below* their own evaluation. The corresponding number of observations is given in brackets. Furthermore we report the results of the Wilcoxon-Mann-Whitney (WMW) test which analyzes whether the difference between *Equal/Above* and *Below* is significant.

B-Person with an own evaluation 0-20 we cannot report a median payoff reduction for the group *Equal/Above* at feedback level 0-20. Furthermore, as by design there are no own evaluations above 80-100 we also cannot report a median payoff reduction for the group *Below* at feedback level 80-100. As shown in Table 3, the differences in payoff reductions are significant up to the feedback level 60-80 for which the difference of the medians is still positive (10 points) but the result of the Wilcoxon-Mann-Whitney test shows an insignificant difference between *Equal/Above* and *Below*.

RESULT 1: *In the incentive treatment, for a given feedback level, (i) the median Person B reduces payoffs of Person A if feedback is below their own evaluation, but (ii) does not reduce payoffs if feedback is confirming or above own evaluation. This rejects Hypotheses 0 and confirms Hypotheses 1.*

Hence, payoff reductions at stage **iv)** in the *incentive* treatment are in-line with the assumption of distributional concerns, reciprocity, and the protection against ego-threats.

To quite some extent, the results in the *flat* treatment (see Table 4) resemble the results from the *incentive* treatment. Using the Wilcoxon-Mann-Whitney test in the *flat* treatment shows a significant difference between these two groups at all feedback levels up to 60-80. Hence, even if the payoff of Person B is independent of Person A's feedback, payoff reduction is significantly higher in situations in which the feedback falls short of the own evaluation compared to feedback that is confirming or above Person B's self perception. Summarizing,

RESULT 2: *In the flat treatment, for a given feedback level, (i) the median Person B reduces payoffs of Person A if feedback is below their own evaluation, but (ii) does not reduce payoffs if feedback is confirming or above own evaluation. This rejects Hypotheses 0 and 2.*

Hence, behavior regarding payoff reductions at stage **iv)** in the *flat* treatment can hardly be explained by reciprocity. To analyze the explanatory power of distributional concerns and the protection against ego-threats with a test of Hypotheses 3 and 4, we investigate the feedback dependence of payoff reductions for a given self-evaluation by Person B.

[Table 5 here]

Table 5—: Median actual payoff reduction per treatment

	Feedback vs. Own Evaluation		**WMW-Test**
	Equal/Above	*Below*	P-Value
Incentive	0 (21)	40 (21)	0.0002
Flat	0 (45)	30.5 (6)	0.0320

Table 5 depicts the actual median payoff reduction by B-Persons with feedback equal/above and below own evaluation for the feedback given by Person A during the experiment. As indicated in the second line (*flat* treatment), median payoff reduction is zero if the feedback is above/equal to Person B's own evaluation and 30.5 if the feedback is below Person B's own evaluation. As indicated by the corresponding Wilcoxon-Mann-Whitney test, median payoff reduction are indeed significantly different for the two cases.

RESULT 3: *In the flat treatment, for a given self-evaluation, the median Person B reduces the payoff of Person A more if the feedback is below rather than above/equal the self-evaluation. This rejects Hypothesis 3 and confirms Hypothesis 4.*

To summarize, taking together the results from the *incentive* and *flat* treatment, our data not only rejects the hypothesis that participants act as selfish maximizers of individual profit, but also demonstrates that models of distributional concerns and reciprocity only explain our findings to a limited extent. In contrast to this, assuming that Person B faces psychological costs of ego-threats (and psychological benefits from a protection of self-esteem) yields predictions regarding payoff reductions that are in line with the results of our experiment.

III. A principal agent model with subjective evaluation

In this section, we propose and analyze a simple principal agent model with subjective performance evaluation to illustrate the economic implications of conflicts generated as a protection against ego-threats. For the sake of comparability to the literature, we now switch from our neutral experimental setting (with Person A and B) to the usual terminology of principal and agent.

The model Assume there is a risk-neutral principal, P, who decides upon undertaking a project which generates a value of $\phi > 0$ if successful. The project requires effort of an agent, A. Assume that if the agent spends effort $p \in [0, 1]$, the project will be successful (create value ϕ) with probability p. The project is a complex good or service and its success is not verifiable, i.e. contracts contingent on the generation of ϕ are not feasible. Neither principal nor agent can directly observe whether the project is successful or not. Rather, both form an opinion about the agent's performance during the production process. I.e., they receive private signals about the agent's performance. The principal receives $s_P \in S_P$, where $S_P = \{L, H\}$, i.e. the principal's opinion can be such that he regards the agent's performance as either high (H) or low (L). Analogously, the agent receives $s_A \in S_A$ with $S_A = \{L, H\}$. The signals s_P and s_A are non-verifiable private pieces of information of the principal and the agent, respectively. The signals are informative with respect to the success of the project. If the project is not successful (which happens with probability $(1-p)$), principal and agent receive the signal $s_P = s_A = L$. If the project is successful, the principal receives the signal $s_P = H$ with probability g, the agent receives the same signal as the principal with probability ρ and receives $s_A = H$ as an independent signal with probability x. Hence, g measures the quality of

the principal's signal, ρ indicates the correlation between the agent's and the principal's signal - or the counter-probability of an independent judgment - and x quantifies the quality of the agent's signal if he forms an independent judgment [our specification of the information technology coincides with Mcleod (2003), p.228, for expositional ease we restrict ourselves to the case of a binary signal].

ASSUMPTION 1: ***(Information Technology)*** *We assume that the principal's and the agent's signal are imperfect, i.e., $g \in (0,1)$ and $x \in (0,1)$, and positively but imperfectly correlated, i.e., $\rho \in (0,1)$.*

We denote by γ_{kl} the conditional probability that $s_P = k$ and $s_A = l$ given that the project is a success. Then, the ex-ante probability for the signal pair $s_P = L$ and $s_A = H$, for instance, will be $p\gamma_{LH} = p(1-g)(1-\rho)x$.[6] Note that by Assumption 1, $\gamma_{HH}\gamma_{LL} > \gamma_{HL}\gamma_{LH}$.

The timing of the game is as follows:

1) The principal offers a contract to the agent and the agent decides upon acceptance. Upfront payments are arranged.

2) The agent decides upon effort p.

3) The project generates value ϕ with probability p.

4) The principal receives s_P and the agent receives s_A. The principal and the agent report (not necessarily truth-fully) on s_P and s_A. Denote the reports by t_P and t_A, respectively. t_P and t_A are verifiable.

5) The payments contingent on t_P and t_A are arranged.

6) Contingent on s_A and received payments, the agent decides upon retaliation (with effort q).

For an effort of p the agent incurs costs $v(p)$ with $v \in C^2$, $v(0) = 0$, $v'(0) = 0$, $v''(p) > 0$ and $\lim_{p\to 1} v(p) = \infty$. Had the principal access to the agent's production technology, his effort choice would solve $v'(p) = \phi$. For further reference, we will denote the first best effort level by p_{FB} and the respective surplus by Π_{FB}. Our assumptions on $v(p)$ ensure that $p_{FB} \in (0,1)$.

The agent is risk-neutral and senses a psychological payoff that depends on his opinion about his own performance, s_A, and the reported opinion of the principal, t_P. More specifically, the agent's utility function reads:

$$U = w - v(p) - Y(t_P, s_A)(1-q) - c(q) \tag{1}$$

Thereby, w denotes the wage payment, $Y(t_P, s_A)$ represents the agent's psychological payoff for a given configuration of (reported) signals, q is the level of conflict (or retaliation) created by the agent and $c(q)$ is the agent's cost for the level of conflict q with $c \in C^2$, $c(0) = 0$, $c'(0) = 0$, $c''(q) > 0$ and $\lim_{q\to 1} c(q) = \infty$.

On the background of our experimental findings and the evidence from social psychology on self-esteem, ego-threats, and conflict creation, we specify $Y(t_P, s_A)$ as follows.

[6] All γ_{kl} as functions of g, ρ, and x can be found in Appendix B.

ASSUMPTION 2: ***(Psychological Costs)*** *(i)* $Y(t_P, s_A = L) = 0$ *for all* t_P*, (ii)* $Y(t_P = H, s_A) = 0$ *for all* s_A*, and (iii)* $Y(L, H) > 0$*.*

Part (i) captures that individuals with low self-esteem (represented by $s_A = L$) do not exhibit ego-involvement and show less reaction to feedback (be it confirming or threatening) [see e.g. Baumeister, Smart & Boden (1996)]. Parts (ii) and (iii) respectively formalize the finding that individuals who hold a high opinion about themselves and are 'ego-involved' ($s_A = H$) uncritically accept positive or confirming feedback [see e.g. Baumeister (2005)] – formalized by zero psychological costs – and suffer from negative or threatening assessments [see e.g. Bushman & Baumeister (1998)] – represented by non-zero psychological cost in our model. In response to an ego-threat the agent can reduce his psychological costs that arise from the deviant (reported) opinions about his performance by creating conflict/trouble [as observed by Baird (1977), Shrauger & Lund (1975) & Korman (1969)]. For further reference, we summarize some results concerning the agent's optimal conflict level.

LEMMA 1: ***(Conflict Creation)*** *Suppose* $Y(t_P, s_A)$ *satisfies Assumption 2. Then, (i) the agent chooses* $q = argmax(Y(t_P, s_A)(1 - q) - c(q))$*., (ii) Suppose* $s_A = L$ *and/or* $t_P = H$*. Then,* $Y(t_P, s_A) = 0$ *and the agent chooses* $q = 0$*, and (iii) Suppose* $s_A = H$ *and* $t_P = L$*. Then, the agent chooses* $q \in (0, 1)$*.*

According to Lemma 1, the agent creates conflict or retaliates (i.e., chooses $q > 0$) if and only if $s_A = H$ and $t_P = L$, i.e., the agent retaliates if and only if he has a high opinion of himself and his ego / self-perception is threatened. This corresponds to our experimental finding that payoffs are only reduced if feedback is below the agent's self-evaluation. For further reference we abbreviate $Y(L, H) \equiv Y$. Moreover, $q^* > 0$ will henceforth denote the conflict level for the configuration $t_P = L$ and $s_A = H$. As the agent chooses $q = 0$ for all other configurations, no confusion should arise. Note that the higher the psychological costs created by the difference in the principal's and agent's evaluation (Y), the higher the level of conflict q^*. We assume throughout this Section that Assumptions 1 and 2 are satisfied.

The principal is risk neutral and maximizes expected profit

$$\Pi = p\phi - E\{w\} - E\{q\}\psi, \tag{2}$$

where $p\phi$ is the expected benefit generated by the agent, $E\{w\}$ are the expected wage cost of employing the agent, and $E\{q\}\psi$ are the expected costs of conflict due to retaliation. As our assumptions on $c(q)$ ensure that $q \in [0, 1]$, we can interpret q as the probability with which the agent creates costs of $\psi > 0$ for the principal. First best profits are given by $\Pi_{FB} = p_{FB}\phi - v(p_{FB})$.

A standard application of the revelation principle (for details see Lemma 3 in Appendix B) implies that we can restrict ourselves – without loss of generality – to simple bonus contracts (a fixed or up-front payment f independent on reported signals and a bonus payment b if the principal reports $t_P = H$). In particular, we can restrict ourselves to bonus payments that are independent of the agent's (opportunistic) report.

Moral Hazard The principal's objective to offer a profit maximizing contract – i.e., an optimal combination of a fixed payment and a bonus – is burdened with (i) moral hazard as the agent's effort is unobservable and (ii) a truth-telling problem as the principal has to credibly commit herself to a truthful revelation of her own signal. As long as the truth-telling constraint is non-binding (pure moral hazard case), the analysis is a straightforward application of standard incentive theory. The corresponding analysis is relegated to Appendix B. For further reference, we report the comparative statics of

optimal effort levels and profits for the pure moral hazard case.[7]

PROPOSITION 1: ***(Pure Moral Hazard)*** *There exists* $\underline{\phi} > 0$ *such that for* $\phi > \underline{\phi}$*, (i)* $\frac{d\tilde{p}}{d\phi} > 0$*,* $\frac{d\tilde{p}}{d\psi} < 0$*,* $\frac{d\tilde{p}}{dg} > 0$*,* $\frac{d\tilde{p}}{d\rho} > 0$*, and* $\frac{d\tilde{p}}{dx} < 0$ *and (ii)* $\frac{d\tilde{\Pi}}{d\phi} > 0$*,* $\frac{d\tilde{\Pi}}{d\psi} < 0$*,* $\frac{d\tilde{\Pi}}{dg} > 0$*,* $\frac{d\tilde{\Pi}}{d\rho} > 0$*, and* $\frac{d\tilde{\Pi}}{dx} < 0$*.*

Hence, if the project is sufficiently valuable to implement a positive effort level (i.e., $\tilde{p} > 0$ because ϕ is large compared to costs of effort implementation including the agent's effort costs, expected retaliation for the principal, and the expected compensation for the psychological costs of the agent), an increase in the value of the project certainly enhances marginal benefits and thereby $\tilde{p}$. Likewise, higher costs of conflict for the principal enhance marginal costs and lower the optimal effort level (recall that signals are conflicting only if the project is successful). A higher quality of the principal's signal reduces the probability of conflict which reduces marginal costs and leads to higher optimal effort levels. A higher correlation of signals or a lower quality of an independent judgment have a similar effect as they also result in lower expected conflict levels and a lower compensation of psychological costs. As indicated in Part (ii), these intuitive effects also carry over to the comparative statics of the principal's profit. The higher the value of the project and the lower expected costs associated with the retaliation of the agent, the more profit is awarded to the principal. In particular, the principal gains from a decrease in retaliation costs ψ (i.e. the size of conflict), an increase in the principal's signal quality g, an increase in the signal correlation ρ and a decrease in the probability that the agent receives an independent signal x (as all these properties of the information technology reduce the probability of conflict).

As the agent does not receive any rents in the optimal contract, the principal's profit also measures the surplus of the relationship. Hence, in the case of non-binding truth-telling constraints, conflicts (i.e. their likelihood γ_{LH} and size $q^*\Psi$ – and thereby the agent's psychological sensitivity Y) only have a welfare detrimental effect. Therefore, any property of the information technology which reduces conflict (i.e. an increase in g or ρ) is welfare-enhancing, while an increase in the quality of the agent's independent judgment x induces the adverse effect.

Truth-telling The truth-telling problem can be represented by the following table (with the principal's report depicted in the rows, the agent's report depicted in the columns, and the principal's profit as entries).[8]

	H	L
H	$p\phi - f - b$	$p\phi - f - b$
L	$p\phi - f - q^*\psi$	$p\phi - f$

Suppose $s_P = H$. Then, the principal tells the truth, whenever his payoff from doing so (which reads $p\phi - f - b$) is larger than his payoff from reporting $t_P = L$ (which reads $p\phi - f - Pr(s_A = H \mid s_P = H)q^*\psi$). This means the principal reports $t_P = H$ if

$$b \leq \frac{\gamma_{HH}}{(\gamma_{HH} + \gamma_{HL})} q^*\psi = (\rho + (1-\rho)x)q^*\psi \equiv b^{max}. \tag{3}$$

[7] We denote the optimal effort in the pure moral hazard case by $\widetilde{p}$ and the corresponding profit by $\tilde{\Pi}$.

[8] In the simple bonus contract (see also Lemma 3 in Appendix B), the agent is indifferent between all possible reports as his payment (and also his psychological payoff) will be unaffected by his own reporting decision. Hence, we can safely adopt the convention that the agent always tells the truth and omit the corresponding truth-telling problem. We therefore only depict the principal's profits in the table.

The principal can only credibly promise a bonus b below b^{max}. Note that this upper bound to credible bonuses increases in the signal correlation ρ and in the quality of an independent judgment x. An increase in each of these parameters lowers the probability of the configuration $s_P = H$ and $s_A = L$ in which case the principal could cheat without facing retaliation and therefore reduces the incentive to save the bonus payment. Moreover, b^{max} certainly increases in the level of conflict $q^*\psi$. However, the maximal credible bonus is independent of g as the principal is only tempted to lie if he received a positive signal.

If $s_P = L$, the principal tells the truth, whenever his payoff from doing so (which reads $p\phi - f - Pr(s_A = H \mid s_P = L)q^*\psi$) is larger than his payoff from reporting $t_P = H$ (which reads $p\phi - f - b$). Hence, the principal reports $t_P = L$ if

$$b \geq \frac{\gamma_{LH}}{(\gamma_{LH} + \gamma_{LL})} q^*\psi = \frac{(1-\rho)x}{(1-\rho x)} q^*\psi \equiv b^{min}. \tag{4}$$

The principal can also not promise to pay arbitrarily low bonuses as he has an incentive to evade conflict through 'unconditional bonuses'. By paying the bonus independently of his signal, the principal avoids any conflict with an agent who is prepared to protect his positive self-image. The minimal credible bonus is thereby decreasing in the signal correlation ρ and increasing in the quality of an independent judgment x because the larger ρ and the smaller x the smaller is the probability of the configuration $s_A = H$ and $s_P = L$ in which case the principal would benefit from conflict evasion. Similarly to b^{max}, b^{min} is independent of g.

Note in particular that $b^{max} > b^{min} > 0$ and that the difference between b^{max} and b^{min} gets larger and the respective interval is shifted towards larger bonuses as q^* or ψ increases. Hence, the larger the potential conflict level, the higher are the bonuses that can be implemented. In fact, for every bonus b there is a conflict level ψ such that b is credible.[9] Hence, while elevated levels of conflict were only welfare detrimental in the pure moral hazard case (see Proposition 1), they relax the upper- and tighten the lower threshold of credible bonuses which will turn out to have an ambiguous effect on welfare.

In the sequel, we call a certain effort level $p > 0$ *implementable* if for the incentive compatible bonus to implement p, $b(p)$, it holds that $b(p) \in [b^{min}, b^{max}]$. Furthermore, we define the minimum implementable effort p^{min} and the maximum implementable effort p^{max} implicitly by $b^{min} = b(p^{min})$ and $b^{max} = b(p^{max})$. Note that $b^{max} > b^{min} > 0$ implies $p^{max} > p^{min} > 0$. Let us denote the optimal effort level by p^* and corresponding profits by Π^*. If the value of the project is sufficiently large to establish a relationship[10] one can distinguish between three cases (see Proposition 5 in Appendix B): i) the case of a binding lower truth-telling constraint, ii) the case of a binding upper truth-telling constraint, and iii) the case of a non-binding truth-telling constraint. The comparative statics in the latter case have already been analyzed in Proposition 1 (as the principal simply implements $p^* = \tilde{p}$ in this case). The analysis of cases i) and ii) deserves some more attention. To this end, the following Lemma captures the comparative statics of p^{min} and p^{max} with respect to the level of conflict and the parameters of the information technology.

LEMMA 2: ***(Truth-Telling Constraints)*** *(i)* $\frac{dp^{min}}{d\Psi} > 0$ *and* $\frac{dp^{max}}{d\Psi} > 0$. *(ii)* $\frac{dp^{max}}{dg} >$

[9] A comprehensive discussion of the comparative statics of b^{max} and b^{min} can be found in Appendix B.

[10] We show in Appendix B that there exists $\overline{\phi} > 0$ with $\overline{\phi} > \underline{\phi}$ such that it is optimal for the principal to implement a positive effort level whenever $\phi > \overline{\phi}$. Hence, the truth-telling constraint aggravates agency costs and thereby requires larger project values than in the pure moral hazard case for the optimality of positive efforts.

0 *and* $\frac{dp^{min}}{dg} > 0$. *(iii)* $\frac{dp^{max}}{d\rho} > 0$ *and* $\frac{dp^{min}}{d\rho} < 0$ *if* ψ *is sufficiently large. (iv)* $\frac{dp^{max}}{dx} > 0$ *and* $\frac{dp^{min}}{dx} > 0$ *if* ψ *is sufficiently large.*

Proof. See Appendix A.

As the level of conflict ψ lifts the minimal credible bonus b^{min} and the maximal credible bonus b^{max} while leaving the incentive compatible bonus $b(p)$ unaltered, p^{min} and p^{max} increase in ψ (Part (i)). Intuitively, the more conflict, the less tempting it is to cheat on the agent (upper truth-telling constraint) and the more tempting it is to evade conflict through unconditional bonus-payments (lower truth-telling constraint).

In contrast, a higher quality of the principal's signal g lowers the incentive compatible bonus $b(p)$ (as bonuses are paid more often) but leaves b^{min} and b^{max} unaltered. Hence, the better the principal's signal, the less costly is the implementation of a certain effort level and the higher is the maximal implementable effort p^{max}. However, lower costs of effort implementation also increase the minimal effort level that can credibly be implemented (p^{min}) (Part (ii)).

In contrast to this, the impact of ρ and x on p^{min} and p^{max} is more subtle (see Parts (iii) and (iv)). These parameters of the information technology influence the minimal, the maximal *and* the incentive compatible bonus. As a result, ρ and x have a direct and an indirect effect on p^{min} and p^{max}. Both parameters modify the probability with which the principal could gain from a lie, but also change the expected psychological costs which have to be compensated by the incentive compatible bonus. Part (iii) and (iv) of Lemma 2 show that, if ψ is sufficiently large such that the gains from a lie are sufficiently pronounced, the former effect dominates the latter.

Welfare Analysis To analyze the comparative statics of the surplus (which is identical to the principal's profits in our set-up), observe that the impact of a parameter y on profits $\Pi(p)$ can be written as $\frac{d\Pi(p)}{dy} = \frac{\partial \Pi(p)}{\partial y} + \frac{\partial \Pi}{\partial p}\frac{dp}{dy}$. We will refer to the first term as the direct effect and the second as the indirect effect. The direct effect captures the impact of the parameter on profits *for a given effort level*. By the envelope theorem, this fully determines the comparative statics of equilibrium profits in the pure moral hazard case (recall that $\frac{\partial \Pi}{\partial p} = 0$ for $p = \tilde{p}$) as depicted in Proposition 1(iii). For a binding truth-telling constraint, the indirect effect can no longer be neglected and may well dominate and reverse the comparative statics of the pure moral hazard case as demonstrated by the following result.

PROPOSITION 2: ***(Comparative Statics of Welfare)*** *(i) There exists* $\tilde{\phi}$ *such that* $\frac{d\Pi^*}{d\psi} > 0$ *for all* $\phi > \tilde{\phi}$. *(ii) There exists* ϕ, Y *and* $v(p)$ *such that* $\frac{d\Pi^*}{dg} < 0$. *(iii) There exists* $\tilde{\phi}$ *and* $\tilde{\psi}$ *such that* $\frac{d\Pi^*}{dx} > 0$ *if* $\psi > \tilde{\psi}$ *and* $\phi > \tilde{\phi}$.

Proof. See Appendix A.

Proposition 2 indicates two different effects which may reverse the comparative statics of the pure moral hazard case. First, the upper truthtelling constraint may be binding. This is in particular the case for large project values ϕ which induce large marginal benefits and therefore require optimal effort levels beyond p^{max}. An increase in ψ or x is welfare detrimental for a given effort level (i.e. $(\frac{\partial \Pi}{\partial \psi, x}) < 0$) but also pushes p^{max} (as indicated in Lemma 2, p^{max} is increasing in Ψ and increasing in x if Ψ is sufficiently large) and thereby relaxes the upper truth-telling constraint. As indicated by Proposition 2(i) and (iii), the latter (indirect) effects indeed dominate the former (direct) effects if project values are sufficiently large. Hence, higher probabilities or levels of conflict are welfare enhancing in the case of valuable projects for which the upper truth-telling constraint is binding.

Second, the lower truthtelling constraint may be binding. This is in particular the case for small project values which are sufficiently attractive to sign contracts on small positive effort levels but operate with bonus payments which tempt the principal to evade conflict by paying the bonus unconditional on the signal. In this case, the principal suffers from parameter changes which tighten the lower truthtelling constraint. For instance, the higher the quality of the principal's signal g, the larger p^{min} and the more tight the lower truthtelling constraint. In contrast, an increase in g enhances the principal's profit *for a given effort level*. According to Proposition 2(ii) the latter (direct) effect may well be dominated by the former (indirect) effect. As a consequence, a better signal for the principal may be welfare detrimental in the case of small projects for which the lower truth-telling constraint is binding.

Note that similar detrimental effects cannot be derived for the correlation of signals ρ, as a higher correlation directly enhances the principal's profit and relaxes the lower and the upper truthtelling constraint as long as ψ is sufficiently large (see Lemma 2(iii)).

We conclude with a comparison of equilibrium profits with the first best solution and a discussion of the limit of a perfect signal to the principal, perfectly correlated signals, and no correct independent judgment of the agent.

PROPOSITION 3: ***(First Best Comparison)*** *(i) Suppose $g < 1$, $\rho < 1$, and $x > 0$. Then, $\Pi^* < \Pi_{FB}$. (ii) Let $\rho = 1$ and/or $x = 0$. Then, $p^* = p_{FB}$ and $\Pi^* = \Pi_{FB}$ if and only if $\frac{\phi}{g} \leq \rho q^* \psi$. (iii) Let $g = 1$. Then, $p^* = p_{FB}$ and $\Pi^* = \Pi_{FB}$ if and only if $\frac{(1-\rho)x}{1-\rho x} q^* \psi \leq \phi \leq (\rho + (1-\rho)x) q^* \psi$*

Proof. See Appendix A.

Part i) indicates that an imperfect information technology of the principal together with an imperfect correlation of the principal's and the agent's signals, and at least some correct independent judgment of the agent induces a welfare loss.

In Part ii) it can be seen that, if signals are perfectly correlated ($\rho = 1$) or the agent's independent judgement never identifies a good project ($x = 0$), a first best will be reached whenever the respective incentive compatible bonus $b(p_{FB}) = \frac{v'(p_{FB})}{g} = \frac{\phi}{g}$ is credible, i.e., $b(p_{FB}) \leq b^{max}$. As the minimal credible bonus b^{min} is zero for $\rho = 1$ or $x = 0$, only the upper truth-telling constraint matters in this case and a first best will be established, if the project value is not too large relative to the expected costs of retaliation.
This changes if we consider the limit $g = 1$. Again, a first best is reached whenever the incentive compatible bonus $b(p_{FB}) = \phi$ is credible. However, as $b^{min} = \frac{(1-\rho)x}{1-\rho x} q^* \psi$ does not vanish as long as $\rho < 1$ and $x > 0$, the first best effort can be too large or too small to be implementable. Hence, it requires a 'fine-tuning' of ϕ (relative to expected costs of conflict) to guarantee a first best solution in this case.

IV. Concluding Remarks

The objective of our paper was twofold. First, we conducted an experiment with subjective performance evaluation and feedback to investigate individual incentives to create conflict in response to a tension between self-perception and performance evaluations by others. Our experimental data indicates that individuals tend to create conflict whenever their own evaluation exceeds the feedback by another party (regardless of whether the feedback also determines the distribution of payoffs or not). This suggests that individuals regard feedback below their self-perception as an ego-threat that triggers attempts to protect one's self-esteem through the creation of conflict.

Second, we propose a simple principal agent model that captures an agent's eagerness

to protect his self-esteem and demonstrate how this facilitates principal-agent relationships even if performance signals are subjective, parties do not interact repetitively, and no third-party can enforce truth-telling. In particular, we analyzed the impact of the conflict level, the psychological sensitivity to ego-threats, and the quality of the information technology on optimal effort levels and social welfare.

Conflict Level Conflict as modeled in this paper unambiguously reduces optimal effort levels and social welfare in the absence of truth-telling constraints. In the presence of truth-telling constraints, however, we show that some conflict potential is needed to establish a positive effort by the agent and that enhanced conflict levels have a positive effect on social welfare in the case of valuable projects which require substantial bonus payments to the agent. E.g., a well-established (internal or external) system of appeals against managerial decision making is not only providing a more peaceful workforce, it may also create the conflict opportunities needed to make bonus payments credible and thereby raise firm profits. The importance of credible conflict for principal agent relations with subjective information has also been emphasized by McLeod (2003). However, while McLeod (2003) assumed that the principal could optimally *choose* conflict levels (through credible payments to a third party in case of conflicting reports), we rather investigate how psychological costs as identified in our experiment may serve the same purpose.

Sensitivity to Ego-Threats Higher levels of conflict unambiguously raise the maximum credible bonus and thereby relax the upper truth-telling constraint in a potentially welfare enhancing way. In contrast, the impact of the psychological sensitivity to ego-threats is more subtle. First of all, some sensitivity is needed to establish the prospect of conflict for the principal and thereby ensure truth-telling. The more aggressive the agent reacts to ego-threats, the higher the anticipated level of conflict and the less restrictive the upper truth-telling constraint. Hence, a more aggressive agent will induce a welfare improvement in case of valuable projects. However, the higher the sensitivity of the agent, the larger the required compensation for anticipated psychological costs. This ceteris paribus enhances necessary bonus payments for a given effort level and thereby reduces the principal's profit and social welfare. The ideal agent from the point of view of a principal who wishes to conduct a very valuable project is therefore someone who reacts very aggressively to ego-threats (i.e., who has low costs of retaliation) but does not suffer too much from an ego-threat and the corresponding retaliation (e.g. because q^* is large). This reinforces our above-made appraisal of appeal systems and suggests to ensure low costs of conflict creation for the employee (e.g. low costs of law suits etc.). Note, however, that these recommendations only hold for very valuable projects which make the upper truthtelling constraint binding. For non-binding truthtelling constraints, psychological sensitivity and the corresponding conflict remains detrimental to the principal's profits and welfare.

Information Technology We also analyzed the impact of the information technology on optimal efforts and welfare. First of all, the principal is advised to use a signal technology which sends a perfectly correlated signal to her and the agent. With perfectly correlated signals the probability of conflicting signals is zero such that the agent does not expect any psychological costs. Moreover, the lower (upper) truthtelling constraint is decreasing (increasing) in the signal correlation such that the interval of credible bonuses is maximized for a given conflict level. Whenever the first best bonus is credible for perfectly correlated signals, a first best will also be reached – in the absence of limited liability or problems of risk allocation as it is the case in our set-up. This lends support to the practice of using information for performance evaluation which is not necessarily highly correlated with actual performance but ensures a high correlation with the agent's self-assessment. Similarly, the probability of conflict will be zero if the agent does not observe good performance independent of the principal. Hence, a first best can also be

achieved with agents who lack an informative independent judgment (i.e., if $x = 0$). However, minimal and maximal implementable efforts are increasing in x (for high conflict levels), such that implementability of the first best is less straightforward for $x = 0$ than for perfectly correlated signals.

The impact of the quality of the principal's signal has shown to be subtle. A better signal reduces necessary bonus payments (due to higher expected returns and lower psychological costs for the agent) and thereby lowers agency costs which yields a welfare improvement – unless the lower truthtelling constraint binds, which may be the case for less valuable projects. Hence, the principal cannot expect higher profits from employing a better information technology for all project values. As a consequence he would not always choose a perfect information technology even if this was costless. The optimal choice of an information technology rather deals with a trade-off between agency costs (which are decreasing in the signal quality) and truthtelling constraints (which may well be tightened by a better information technology). Hence, imperfect information technologies as observed in reality may not only be optimal due to cost considerations but also due to the strategic aspects discussed in this paper.

V. References

1) Abeler, Johannes, Armin Falk, Lorenz Goette and David Huffman. Forthcoming. "Reference Points and Effort Provision." *American Economic Review*.

2) Baumeister, Roy F. 2004. "Self-Concept, Self-esteem and Identity." In textitPersonality: Contemporary Theory and Research, ed. Valerian J. Derlega, Barbara A. Winstead and Warren H. Jones, 246-280. Wadsworth Publishing Company, 3rd Edition.

3) Baumeister, Roy F., Laura Smart and Joseph M. Boden. 1996. "Relation of threatened egotism to violence and aggression: The dark side of high self-esteem.", *Psychological Review*, 103: 5-33.

4) Baird, Lloyd S. 1977. "Self and Superior Ratings of Performance: As Related to Self-Esteem and Satisfaction with Supervision." *The Academy of Management Journal*, 20(2): 291-300.

5) Bénabou, Roland and Jean Tirole. 2002. "Self-Confidence And Personal Motivation." *Quarterly Journal of Economics*, 117(3): 871-915.

6) Bushman, Brad J. and Roy F. Baumeister. 1998. "Threatened Egotism, Narcissism, Self-Esteem, and Direct and Displaced Aggression: Does Self-Love or Self-Hate Lead to Violence?" *Journal of Personality and Social Psychology*, 75(1): 219-229.

7) Compte, O. & Postlewaite, A. (2004. "Confidence-Enhanced Performance." *American Economic Review*, 94(5): 1536-1557.

8) Charness, Gary and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *Quarterly Journal of Economics*, 117(3): 817-69.

9) Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior*, 47: 268-298.

10) Ellingsen, Tore and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review*, 98(3): 990-1008.

11) Falk, Armin and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and Economic Behavior*, 54(2): 293-315.

12) Gibbs, Michael, Kenneth A. Merchant, Wim A. Van der Stede and Mark E. Vargus. 2004. "Determinants and effects of subjectivity in incentives." *The Accounting Review*, 79: 409-436.

13) Greenwald, Anthony G. 1980. "The totalitarian ego: Fabrication and revision of personal history." *American Psychologist*, 35: 603-618.

14) Ittner, Cristopher D., David F. Larcker and Marshall W. Meyer. 2003. "Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard." *The Accounting Review*, 78: 725-758.

15) Korman, Abraham K. 1969. "Toward a Hypothesis of Work Behavior." *Journal of Applied Psychology*, 54: 31-41.

16) Köszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association*, 4(4): 673-707.

17) Köszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121(4): 1133-1166.

18) Levine, Jonathan. 2003. "Relational Incentive Contracts." *American Economic Review*, 93(3): 835-857.

19) MacLeod, Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review*, 93(1): 216-240.

20) Milkovich, George and Alexandra K. Wigdor. 1991. "Pay for Performance: Evaluating Performance Appraisal and Merit Pay." eds. National Academy Press, Washington, D.C.

21) Murphy, Kevin J. and Paul Oyer. 2003 "Discretion in executive incentive contracts." Working paper, University of Southern California and Stanford University.

22) Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1): 7-63.

23) Raskin, Robert, Jill Novacek and Robert Hogan. 1991. "Narcissism, self-esteem, and defensive self-enhancement." *Journal of Personality*, 59(1): 19-38.

24) Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." *American Economic Review*, 83: 12811302.

25) Shrauger, Sidney J. and Adrian K. Lund. 1975. "Self-evaluation and reactions to evaluations from others." *Journal of Personality*. 43: 94-108.

26) Steelman, Lisa A. and Kelly A. Rutkowski. 2004. "Moderators of employee reactions to negative feedback." *Journal of Managerial Psychology*, 19(1): 6-18.

Appendix A: Proofs

Proof of Lemma 2

p^{min} is implicitly given by

$$b^{min} = \frac{(1-\rho)x}{1-\rho x} q^* \psi = \frac{1}{g}(v'(p^{min}) + (1-g)(1-\rho)x(Y(1-q^*) + c(q^*))) = b(p^{min})$$

and p^{max} is implicitly given by

$$\begin{aligned} b^{max} &= (\rho + (1-\rho)x) q^* \psi \\ &= \frac{1}{g}(v'(p^{max}) + (1-g)(1-\rho)x(Y(1-q^*) + c(q^*))) = b(p^{max}). \end{aligned}$$

We use these equations to compute the comparative statics of p^{min} and p^{max}. To be specific, let $F^{min} = b^{min} - b(p)$ and $F^{max} = b^{max} - b(p)$. Then, for a parameter y, $\frac{dp^{min/max}}{dy} = -\frac{\partial F^{min/max}/\partial y}{\partial F^{min/max}/\partial p}$. Note that $\frac{\partial F^{min/max}}{\partial p} = -\frac{v''(p)}{g} < 0$.

Part (i). Follows from $\frac{\partial b^{min} - b(p)}{\partial \psi} = \frac{(1-\rho)x}{1-\rho x} q^* > 0$ and $\frac{\partial b^{max} - b(p)}{\partial \psi} = (\rho + (1-\rho)x) q^* > 0$.

Part (ii). Observe that $\frac{\partial b^{min} - b(p)}{\partial g} = \frac{b(p^{min})}{g} + \frac{1}{g}((1-\rho)x(Y(1-q^*) + c(q^*))) > 0$ and $\frac{\partial b^{max} - b(p)}{\partial g} = \frac{b(p)}{g} + \frac{1}{g}((1-\rho)x(Y(1-q^*) + c(q^*))) > 0$.

Part (iii). $\frac{\partial b^{min} - b(p)}{\partial \rho} = \frac{-x(1-x)}{(1-\rho x)^2} q^* \psi + \frac{1}{g}((1-g)x(Y(1-q^*) + c(q^*)))$ is negative if ψ is sufficiently large (for a given $0 < x < 1$). Moreover, $\frac{\partial b^{max} - b(p)}{\partial \rho} = (1-x) q^* \psi + \frac{1}{g}((1-g)x(Y(1-q^*) + c(q^*))) > 0$.

Part (iv). $\frac{\partial b^{min} - b(p)}{\partial x} = \frac{1-\rho}{(1-\rho x)^2} q^* \psi - \frac{1}{g}((1-g)(1-\rho)(Y(1-q^*) + c(q^*)))$ is positive if ψ is sufficiently large. $\frac{\partial b^{max} - b(p)}{\partial x} = (1-\rho) q^* \psi - \frac{1}{g}((1-g)(1-\rho)(Y(1-q^*) + c(q^*)))$ is positive if ψ is sufficiently large.

Proof of Proposition 2

The impact of a parameter y on equilibrium profits $\Pi(p^*)$ can be denoted by $\frac{d\Pi(p^*)}{dy} = \frac{\partial \Pi(p^*)}{\partial y} + \frac{\partial \Pi(p^*)}{\partial p}\frac{dp^*}{dy}$. For $\frac{\partial \Pi(p)}{\partial y}$ see the proof of Proposition 1(iii). $\frac{\partial \Pi(p)}{\partial p} = \Phi - \gamma_{LH} q^* \psi - v'(p) - \gamma_{LH}(Y(1-q^*) + c(q^*))$. Note that for a fixed p, $\frac{\partial \Pi(p)}{\partial p}$ is a linear increasing function of ϕ and for a fixed ϕ it is a decreasing function of p with slope $-v''(p)$.

Part (i). Recall from Lemma 2(i) that $\frac{dp^{max}}{d\psi} > 0$. Fix any $p^{max} \in (0,1)$. Then, there exists a project value ϕ' such that $\frac{\partial \Pi(p)}{\partial p}|_{p=p^{max}} > 0$ and $p^* = p^{max}$ for all $\phi > \phi'$. In particular, $\frac{\partial \Pi(p)}{\partial p}|_{p=p^*}\frac{dp^*}{d\psi} > 0$. As $\frac{dp^{max}}{d\psi}$ and $\frac{\partial \Pi(p)}{\partial \psi}$ are independent of ϕ and $\frac{\partial \Pi(p)}{\partial p}$ is a linear increasing function of ϕ, there exists a ϕ'' such that $\frac{d\Pi(p^*)}{d\psi} > 0$ for all $\phi > \tilde{\phi} \equiv \max(\phi', \phi'')$.

Part (ii). Fix any $p^{min} \in (0,1)$ and a positive real number z. Then, there exists an effort cost function $v(p)$ such that $\frac{v'(p^{min})}{v''(p^{min})} > z$ and there exists a project value ϕ such that $0 < \tilde{p} < p^{min}$ and $\Pi(p^{min}) > 0$ (and therefore $p^* = p^{min}$). Then, by Lemma 2(ii), $\frac{dp^{min}}{dg} > 0$ and $\frac{\partial \Pi(p)}{\partial g} = -p\frac{d\gamma_{LH}}{dg}(q^* \Psi + (Y(1-q^*) + c(q^*))) > 0$ (see the

proof of Proposition 1). Now observe that $\frac{\partial \Pi(p)}{\partial g}$ is independent of $v(p)$ and its derivatives while $\frac{\partial \Pi}{\partial p}\frac{\partial p^{min}}{\partial g}$ is increasing in $\frac{v'(p)}{v''(p)}$. Hence, $\frac{d\Pi(p)}{dg} < 0$ if z is sufficiently large.

Part (iii). Recall from Lemma 2(iv) that there exists a $\tilde{\psi}$ such that $\frac{dp^{max}}{dx} > 0$ for all $\psi > \tilde{\psi}$. Fix any $p^{max} \in (0,1)$ with such a ψ. Then, there exists a project value ϕ' such that $\frac{\partial \Pi(p)}{\partial p}|_{p=p^{max}} > 0$ and $p^* = p^{max}$ for all $\phi > \phi'$. In particular, $\frac{\partial \Pi(p)}{\partial p}|_{p=p^*}\frac{dp^*}{dx} > 0$. As $\frac{dp^{max}}{dx}$ and $\frac{\partial \Pi(p)}{\partial x}$ are independent of ϕ and $\frac{\partial \Pi(p)}{\partial p}$ is a linear increasing function of ϕ, there exists a ϕ'' such that $\frac{d\Pi(p^*)}{dx} > 0$ for all $\phi > \tilde{\phi} \equiv \max(\phi', \phi'')$.

Proof of Proposition 3

Part (i). Follows from non-zero psychological costs and costs of conflict in this case.

Part (ii) and (iii). $g = 1$, $\rho = 1$, or $x = 0$ implies that $\gamma_{LH} = 0$ and therefore $\Pi(p) = p\phi - v(p)$ such that $\tilde{p} = p_{FB}$. However, $b(p_{FB})$ has to be in the interval $[b^{min}, b^{max}]$ which results in the condition displayed in the proposition (recall that for $x = 0$ or $\rho = 1$, $b^{min} = 0$).

Supplementary Material - Not For Publication

Appendix B

Information Technology

Conditional probabilities $\gamma_{k,l}$ for signal configuration ($s_P = k$,$s_A = l$) are

$$\gamma_{HH} = g\left(\rho + (1-\rho)\, x\right) \text{ and } \gamma_{HL} = g(1-\rho)(1-x),$$

$$\gamma_{LL} = (1-g)\left(\rho + (1-\rho)\,(1-x)\right) \text{ and } \gamma_{LH} = (1-g)\,(1-\rho)\, x.$$

Reduced Form Contracts

In our setting with unobservable effort and subjective measures of performance, a contract Γ can only be contingent on the reported subjective opinions of the principal and the agent. Hence, a contract fixes payments for all configurations of reports t_P and t_A and reads $\Gamma = \{w_{kl} \mid k \in S_P, l \in S_A\}$. The agent accepts a contract if he expects a (weakly) positive utility from it (individual rationality) and chooses p as to maximize his utility (incentive compatibility). If a contract Γ is individually rational and the agent chooses effort p, we say that Γ *implements* p. Principal and agent report their opinions, i.e. signals, truthfully if and only if they weakly benefit from doing so.

LEMMA 3: ***Reduced Form Contracts***

Suppose there exists a contract Γ which implements $p > 0$. Then, there always exists a contract $\hat{\Gamma}$ which implements p at weakly lower costs and (i) principal and agent tell the truth, (ii) $w_{kl} = w_{km} \equiv w_k$ for all $k \in S_P$ and $l, m \in S_A$, and (iii) $w_H > w_L$.

PROOF:

To save on notation, we denote $Y(t_P = l, s_A = k)(1-q^*) - c(q^*) \equiv Y_{kl}$ throughout this proof.

Part(i). For a given contract Γ and signals s_P and s_A, the principal and the agent decide upon their report. Let $\sigma_P : S_P \to \Delta(S_P)$ and $\sigma_A : S_A \to \Delta(S_A)$ be the principal's and agent's reporting strategies (*i.e.,* mappings from the set of signals S_P and S_A to the set of probability distributions over S_P and S_A, respectively). Suppose that (σ_P^*, σ_A^*) is the pair of optimal reporting strategies for contract Γ. Then, the revelation principle implies that there exists a contract $\hat{\Gamma}$ which implements the same effort at the same costs and induces truthful reports by principal and agent. We will, henceforth, restrict our analysis to this type of (revelation) contracts.

Part (ii). Suppose that $\Gamma = \{w_{kl}\}$ is a revelation contract, *i.e.,* the principal and the agent tell the truth under contract Γ. As Γ implements $p > 0$, the incentive compatibility constraint

$$\Sigma_{k \in S_P, l \in S_A}(w_{kl} - Y_{kl}) \frac{dPr\{s_P = k, s_A = l\}}{dp} = v'(p)$$

is satisfied. Consider a contract $\hat{\Gamma}$ which fixes payments of

$$\hat{w}_k = \sum_{l \in S_A} w_{kl} Pr\{s_P = k, s_A = l\}$$

if the principal receives signal $s_P = k$, i.e., payments are independent of s_A. These payments also satisfy the incentive compatibility constraint (see above).[11] Moreover, the agent weakly benefits from telling the truth. Finally, the principal's truth-telling constraint is also satisfied under $\hat{\Gamma}$. To see this observe that the principal reports k given that he has received k under contract Γ if

$$\begin{aligned} & Pr\{s_A = H \mid s_P = k\}(w_{oH} - w_{kH}) + Pr\{s_A = L|s_P = k\}(w_{oL} - w_{kL}) \\ & \geq Pr\{s_A = H|s_P = k\}((q^*\psi)_{kH} - (q^*\psi)_{oH}) \\ & +Pr\{s_A = L|s_P = k\}((q^*\psi)_{kL} - (q^*\psi)_{oL}) \end{aligned} \tag{5}$$

for all $o \in S_P$ (where $(q^*\psi)_{t_A,t_P}$ denotes the anticipated conflict costs for a reported configuration (t_A, t_P)). This set of inequalities holds because Γ implements truth-telling by assumption. $\hat{\Gamma}$ implements truth-telling if

$$\begin{aligned} \hat{w}_o - \hat{w}_k \quad \geq \quad & Pr\{s_A = H|s_P = k\}((q^*\psi)_{kH} - (q^*\psi)_{oH}) \\ & +Pr\{s_A = L|s_P = k\}((q^*\psi)_{kL} - (q^*\psi)_{oL}). \end{aligned}$$

holds for all $o, k \in S_P$. Inserting $\hat{w}_k$ and $\hat{w}_o$ yields

$$\begin{aligned} & Pr\{s_A = H|s_P = k\}(w_{oH} - c_{kH}) + Pr\{s_A = L|s_P = k\}(w_{oL} - w_{kL}) \\ & \geq Pr\{s_A = H|s_P = k\}((q^*\psi)_{kH} - (q^*\psi)_{oH}) \\ & +Pr\{s_A = L|s_P = k\}((q^*\psi)_{kL} - (q^*\psi)_{oL}). \end{aligned}$$

which coincides with Eqs. 5 and therefore shows that for $\hat{\Gamma}$ the principal's truthtelling constraint is satisfied as well. Hence, any revelation contract Γ can be substituted by a revelation contract $\hat{\Gamma}$ with w_{kl} independent of l which also implements $p > 0$ and leaves the principal weakly better off.

Part (iii). Suppose by contradiction that Γ implements $p > 0$ with $w_H = g$ and $w_L = g + \epsilon$ with $\epsilon \geq 0$. Then, the incentive compatibility constraint of the agent can be written as

$$\epsilon = \frac{v'(p) + \gamma_{LH} Y_{LH}}{(\gamma_{LH} + \gamma_{LL} - 1)}.$$

Observe that the numerator of the *rhs* is strictly positive and the denominator is strictly negative. Hence, the *rhs* is strictly negative and the incentive compatibility constraint is not satisfied for any $\epsilon \geq 0$. A contradiction.

Pure Moral Hazard Problem

In this section we abstract from the truthtelling problem inherent to the principal-agent relationship in order to analyze the isolated impact of moral hazard on the optimal effort level chosen by the principal and social welfare. Hence, we assume throughout this section that the contract $\Gamma = (f, b)$ guarantees truth-telling (i.e., truth-telling constraints are non-binding).

For a given contract $\Gamma = (f, b)$, the agent chooses effort p as to maximize his utility (see Eqn. 1) while anticipating the generation of ex-post conflict at level q^* as depicted

[11] Individual rationality is trivially fulfilled as expected payments for the agent are the same under Γ and $\hat{\Gamma}$, and Γ is individually rational by assumption.

in Lemma 1. This means, he maximizes

$$U(p) = p(\gamma_{HH} + \gamma_{HL})b + f - v(p) - p\gamma_{LH}(Y(1-q^*) + c(q^*))$$

which induces the first order condition[12]

$$\begin{aligned} b(p) &= \frac{v'(p) + \gamma_{LH}(Y(1-q^*) + c(q^*))}{\gamma_{HH} + \gamma_{HL}} \\ &= \frac{1}{g}(v'(p) + (1-g)(1-\rho)x(Y(1-q^*) + c(q^*))). \end{aligned} \tag{6}$$

Note that $\frac{d^2U(p)}{dp^2} = v''(p) > 0$ such that the agent's optimization problem is well-behaved. Eqn. (6) shows that the incentive compatible bonus that the principal pays to the agent in case he beliefs that the agent did a good job has to overcome marginal effort costs and marginal psychological costs. If the principal wants to induce a positive effort level, he has to offer a positive bonus. Note, however, that the required bonus does not vanish in the limit of small efforts, because marginal psychological costs do not vanish for $p = 0$. Finally, observe that the incentive compatible bonus increases in target effort p, psychological costs Y, and the conditional probability of conflict (γ_{LH}). In particular, a higher quality of the principal's signal g reduces the incentive compatible bonus because the agent expects higher returns to effort and the probability of conflict decreases. Likewise, a lower correlation of the signals or a higher probability of a positive independent evaluation by the agent enhances the compensation requested by the agent for a given effort level.

The agent accepts a contract $\Gamma = (f, b)$ whenever his expected utility from it is weakly positive, i.e.

$$p(\gamma_{HH} + \gamma_{HL})b + f - v(p) - p\gamma_{LH}(Y(1-q^*) + c(q^*)) \geq 0.$$

To maximize her profits, the principal sets the upfront payment for a given bonus b to

$$f(b) = -p(\gamma_{HH} + \gamma_{HL})b + v(p) + p\gamma_{LH}(Y(1-q^*) + c(q^*)).$$

Observe that the upfront-payment can well be negative (*i.e.,* a franchise fee) as the agent is not protected by limited liability. Note in particular that $f(b)$ can always be fixed such that the agent does not receive any rents from the relationship.

To implement effort $p > 0$ the principal's costs are $C(p) = f + p(\gamma_{HH} + \gamma_{HL})b(p) = v(p) + p\gamma_{LH}((1-q^*)Y + c(q^*))$. Note that $C(p)$ is convex and that $C(0) = 0$. We adopt the convention that an effort $p > 0$ which is not implementable requires infinite costs. The principal's profit now reads

$$\Pi(p) = p\phi - p\gamma_{LH}q^*\psi - C(p)$$

which is zero for $p = 0$ and concave for $p > 0$. We denote the maximum of $\Pi(p)$ on $[0, 1]$ by $\widetilde{p}$ and the corresponding profit for the principal by $\tilde{\Pi}$ and derive the following set of results.[13]

PROPOSITION 4: ***Pure Moral Hazard***

[12] We denote a bonus which implements an effort level of p by $b(p)$.

[13] $\widetilde{p}$ and $\tilde{\Pi}$ are equilibrium effort and profit whenever the truth-telling constraints are non-binding.

(i) $\widetilde{p} > 0$ *if and only if* $\phi > \underline{\phi} \equiv \gamma_{LH}(q^*\Psi + ((1-q^*)Y + c(q^*)))$.

(ii) Suppose $\phi > \underline{\phi}$. *Then,* $\frac{d\tilde{p}}{d\phi} > 0$, $\frac{d\tilde{p}}{d\psi} < 0$, $\frac{d\tilde{p}}{dg} > 0$, $\frac{d\tilde{p}}{d\rho} > 0$, *and* $\frac{d\tilde{p}}{dx} < 0$.

(iii) Suppose $\phi > \underline{\phi}$. *Then,* $\frac{d\tilde{\Pi}}{d\phi} > 0$, $\frac{d\tilde{\Pi}}{d\psi} < 0$, $\frac{d\tilde{\Pi}}{dg} > 0$, $\frac{d\tilde{\Pi}}{d\rho} > 0$, *and* $\frac{d\tilde{\Pi}}{dx} < 0$.

PROOF:

Part(i). Consider

$$\Pi(p) = p\phi - p\gamma_{LH}q^*\psi - C(p)$$

with $C(p) = v(p) + p\gamma_{LH}((1-q^*)Y + c(q^*))$. Observe that $\Pi = ap - v(p)$ with $a = \phi - \gamma_{LH}(q^*\Psi + ((1-q^*)Y + c(q^*)))$. Recall that $v(0) = 0$, $v'(0) = 0$, and $v''(p) > 0$. Then, $\widetilde{p} > 0$ if and only if $a > 0$.

Part (ii). We use the first order condition

$$\frac{d\Pi}{dp} = \phi - \gamma_{LH}q^*\psi - v'(p) - \gamma_{LH}(Y(1-q^*) + c(q^*)) = 0. \tag{7}$$

as an implicit function of $\tilde{p}$. With we get

$$\frac{d\tilde{p}}{d\phi} = -\frac{1}{-v''(\tilde{p})} > 0,$$

$$\frac{d\tilde{p}}{d\psi} = -\frac{-\gamma_{LH}q^*}{-v''(\tilde{p})} < 0,$$

$$\frac{d\tilde{p}}{d\gamma_{LH}} = -\frac{-q^*\psi - (Y(1-q^*) + c(q^*))}{-v''(\tilde{p})} < 0,$$

$$\frac{d\tilde{p}}{dY} = -\frac{-\gamma_{LH}\psi\frac{dq^*}{dY} - \gamma_{LH}(1-q^*)}{-v''(\tilde{p})} < 0.$$

which implies Part (ii) (recall that $\frac{d\gamma_{LH}}{dg} < 0$, $\frac{d\gamma_{LH}}{d\rho} < 0$, and $\frac{d\gamma_{LH}}{dx} > 0$).

Part (iii). Follows directly from

$$\frac{\partial\Pi(p)}{\partial\phi} = p > 0,$$

$$\frac{\partial\Pi(p)}{\partial\Psi} = -p\gamma_{LH}q^* < 0,$$

$$\frac{\partial\Pi(p)}{\partial g} = -p\frac{d\gamma_{LH}}{dg}(q^*\Psi + (Y(1-q^*) - c(q^*))) > 0,$$

$$\frac{\partial\Pi(p)}{\partial\rho} = -p\frac{d\gamma_{LH}}{d\rho}(q^*\Psi + (Y(1-q^*) - c(q^*))) > 0,$$

$$\frac{\partial\Pi(p)}{\partial x} = -p\frac{d\gamma_{LH}}{dx}(q^*\Psi + (Y(1-q^*) - c(q^*))) < 0$$

for any $p > 0$ and the envelope theorem $\frac{d\tilde{\Pi}}{dy} = \frac{\partial\Pi}{\partial y}|_{p=\tilde{p}}$ for a parameter y.

Comparative Statics of Bonuses

Recall that $b(p) = \frac{v'(p)+\gamma_{LH}(Y(1-q^*)+c(q^*))}{\gamma_{HH}+\gamma_{HL}} = \frac{1}{g}(v'(p)+(1-g)(1-\rho)x(Y(1-q^*)+c(q^*)))$ which implies.

LEMMA 4: ***Comparative Statics of*** $b(p)$
(i) Suppose $p > 0$. *Then,* $b(p) > 0$. *(ii)* $\lim_{p\to 0} b(p) > 0$. *(iii)* $\frac{db(p)}{dp} > 0$. *(iv)* $\frac{db(p)}{dg} < 0$. *(v)* $\frac{db(p)}{d\rho} < 0$. *(vi)* $\frac{db(p)}{dx} > 0$.

The definition of b^{min} and b^{max}

$$b^{max} = \frac{\gamma_{HH}}{(\gamma_{HH}+\gamma_{HL})}q^*\psi = (\rho + (1-\rho)x)q^*\psi$$

$$b^{min} = \frac{\gamma_{LH}}{(\gamma_{LH}+\gamma_{LL})}q^*\psi = \frac{(1-\rho)x}{(1-\rho x)}q^*\psi$$

implies the following results.

LEMMA 5: ***Comparative Statics of*** b^{max} ***and*** b^{min}
(i) $b^{min} > 0$. *(ii)* $b^{max} > b^{min}$. *(iii)* $\Delta b \equiv b^{max} - b^{min}$ *is monotone increasing in* q^* *and* ψ. *(iv)* b^{min} *is monotone increasing in* q^*, ψ, *and* x *and monotone decreasing in* ρ. *(v)* b^{max} *is monotone increasing in* q^*, ψ, ρ, *and* x. *(vi) Both* b^{max} *and* b^{min} *are independent on* g.

PROOF:
(i) and (ii) follow from the positive correlation of signals, *i.e.,* $\rho > 0$ or $\gamma_{HH}\gamma_{LL} > \gamma_{HL}\gamma_{LH}$.
(iii) Follows from $\Delta b = \frac{\gamma_{HH}\gamma_{LL}-\gamma_{HL}\gamma_{LH}}{(\gamma_{HH}+\gamma_{HL})+(\gamma_{LH}+\gamma_{LL})}q^*\psi$.
(iv), (v), and (vi) follow directly from Eqs 3 and 4.

Truth-telling problem

We denote the maximum of $\Pi(p) = p\phi - p\gamma_{LH}q^*\psi - C(p)$ on $\{0\} \cup [p^{min}, p^{max}]$ by p^*. p^* will be referred to as the optimal effort level (p^* is the optimal effort level for the principal given that only effort levels between p^{min} and p^{max} are feasible) and $\Pi^* = \Pi(p^*)$ will be the corresponding profit for the principal.

PROPOSITION 5: ***Optimal Effort Level***

$p^* > 0$ *if and only if* $\phi > \overline{\phi} > \underline{\phi}$ *with* $\Pi(p^{min})|_{\phi=\overline{\phi}} = 0$.
Now suppose that $\phi > \overline{\phi}$.

(i) Binding Lower Truth-Telling Constraint: If $0 < \tilde{p} < p^{min}$, *then the principal implements* $p^* = p^{min}$ *with bonus* b^{min} *[Figure 1].*

(ii) Binding Upper Truth-Telling Constraint: If $\tilde{p} > p^{max}$, *then the principal implements* $p^* = p^{max}$ *with bonus* b^{max} *[Figure 2].*

(iii) Non-Binding Truth-Telling Constraint: If $\tilde{p} \in [p^{min}, p^{max}]$, *then the principal implements* $p^* = \tilde{p}$ *by paying* $b(\tilde{p})$ *[Figure 3].*

PROOF:

$''\Leftarrow''$ Suppose $\phi > \overline{\phi}$. As $\frac{\partial \Pi}{\partial \phi} > 0$, $\Pi(p^{min})|_{\phi} > 0 = \Pi(p = 0)$ and therefore $p^* > 0$. By Proposition 1, this implies that $\underline{\phi} \leq \overline{\phi}$.

$''\Rightarrow''$ Suppose $p^* > 0$. Then, $\overline{\phi} > \underline{\phi}$ (see Proposition 1). Hence, $\Pi(p)$ is continuous in $p \geq 0$ and concave with a unique maximum at $\tilde{p} > 0$. Now suppose that $\phi < \overline{\phi}$ such that $\Pi(p^{min})|_{\phi} < 0$. Then, $\tilde{p} < p^{min}$ and $\Pi(p) < 0$ for all $p \in [p^{min}, p^{max}]$. A contradiction.

To see that $\overline{\phi} \neq \underline{\phi}$, recall from Lemma 5 that $b^{min} > 0$ which implies $p^{min} > 0$. Now suppose that $\phi = \overline{\phi} = \underline{\phi}$. Then, $\Pi(p^{min}) = 0$ by definition of $\overline{\Phi}$. Then, continuity and concavity of $\Pi(p)$ imply $0 < \widetilde{p} < p^{min}$ where the first inequality contradicts Proposition 1(i).

Items (i) to (iii) are a direct implication of the fact that $\Pi(p)$ is continuous in $p \geq 0$ and concave with a unique maximum at $\tilde{p} > 0$ whenever $\phi > \overline{\phi}$, and the observation that $p^{max} > p^{min} > 0$.

According to Proposition 5, there will be no principal-agent relationship (i.e. $p^* = 0$) whenever the returns to the project are below a certain threshold. Note in particular, that the presence of a truth-telling problem increases the corresponding threshold value compared to the pure moral hazard case ($\overline{\phi} > \underline{\phi}$) which already indicates potential welfare losses due to truth-telling constraints. Finally, observe that in the absence of conflict (i.e., $q^*\Psi = 0$) it holds that ($p^{min} = 0$) such that profits for the principal are zero at p^{min} for any ϕ. This establishes the familiar result that no positive effort can be implemented in the absence of conflict if performance evaluations are subjective.

[Figures A.1-A.3 here]

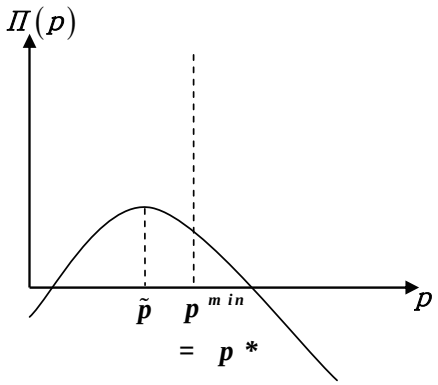

Figure A.1. : Binding Lower Truth-Telling Constraint

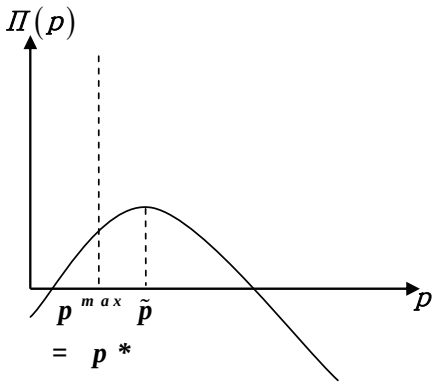

Figure A.2. : Binding Upper Truth-Telling Constraint

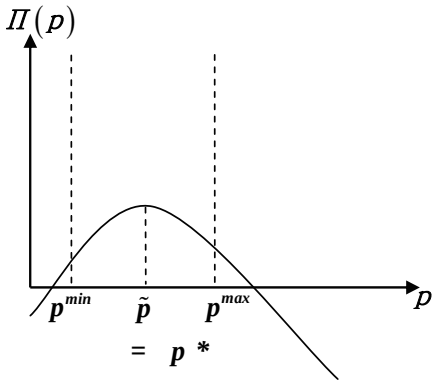

Figure A.3. : Non-Binding Truth-Telling Constraint

Supplementary Material - Not For Publication

Appendix C

Experimental instructions

Dear Participant,

Welcome to the experiment.

Important: Please do not communicate with other participants during the experiment and switch off your mobile phones. Read the instructions carefully. If something is not well explained or any question turns up now or at any time later in the experiment, then ask one of the experimenters. Do, however, not ask out loud, but raise your hand! We will clarify questions privately. You can use the instructions throughout the experiment whenever you want to re-clarify certain things and you may take notes on them, if you wish.

This experiment is a project from researchers from the University of Copenhagen and Bamberg University (Germany). It studies people's behavior in work situations.

You can earn money in this experiment. The amount of money that you will receive depends on your decisions as well as another person's decisions. All earnings will be paid out at the end of the experiment.

During the experiment, your income will be calculated in points. These points are converted into Danish kroner (DKK) according to the following exchange rate:

10 points = 3.5 DKK

In this experiment you will be randomly grouped into pairs and assigned to one of two different roles. We name these roles Person A and Person B. This means, during the experiment you will be paired with one other person in this room and you will be either Person A or Person B. If you are Person A, you will be paired with Person B and vice versa.

Note, both of you start with an endowment of 200 points in the beginning of the experiment that will be part of your final payoff.

On the following page we will reveal your role, i.e. Person A or Person B, and explain to you what the experiment is about.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

You have randomly been assigned to the role of **Person B**.

During this experiment you are paired with another person in this room: Person A.

The experiment has 4 stages:

Stage 1 (Questions): After reading the instructions, please answer the questions that you find on the screen. These questions are related to the instructions and they should check in how far the information in the instructions is clear. Please answer all the questions. When Person A has answered the questions a "Next" button will appear at the bottom of your screen. Please click it. When the answers to all questions are correct, clicking the button "Next" will start stage 2 of the experiment.

Stage 2 (Clicking-Task): You will be given a task. The task that you will be given is "clicking away boxes". This means, for a period of 90 seconds screens with boxes will appear for various time lengths and your task is to click the boxes away.

Note, Person A will be able to observe on his screen how you work on your task. This means, he / she will see the same screen as you and observe you clicking away the boxes.

Important: your performance will generate Person A's payoff.

If you click away:

- 0-20% of the boxes that appear during the 90 sec., then Person A will receive 200 points,
- 20-40% of the boxes that appear during the 90 sec., then Person A will receive 300 points,
- 40-60% of the boxes that appear during the 90 sec., then Person A will receive 400 points,
- 60-80% of the boxes that appear during the 90 sec., then Person A will receive 500 points,
- 80-100% of the boxes that appear during the 90 sec., then Person A will receive 600 points.

Stage 3 (Evaluation and Feedback): After the clicking-task, both of you will be asked to evaluate your performance. Note, these evaluations will NOT be communicated to the person that your

are paired with. In addition, you will be given feedback by Person A which is communicated to you at the end of the experiment.

Important: by giving feedback to you, Person A decides how much he / she wants to give you from his / her payoff that was generated through your performance during the clicking-task.

If Person A's feedback is:

- 0-20%, then you receive 100 points from Person A,
- 20-40%, then you receive 150 points from Person A,
- 40-60%, then you receive 200 points from Person A,
- 60-80%, then you receive 250 points from Person A,
- 80-100%, then you receive 300 points from Person A.

Stage 4 (Reaction): Following the feedback stage, you will be able to use the 200 points initial endowment and the points that you receive because of Person A's feedback to react to his / her evaluation of your performance. This means, you will be asked by how much you would like to reduce Person A's payoff given his / her feedback.

Your answer to this question can reduce Person A's payoff by up to 100 points. However, note that for every point that you reduce Person A's payoff, you have to pay 0.25 points. This means, for example, a reduction of 40 points of Person A's payoff, costs you 10 points etc.

Important: Note, neither you nor Person A will be informed about your actual performance in the clicking-task before the end of the experiment. So all decisions that you and Person A take during the experiment are based on your own subjective opinions. Note, however, that whatever decision you take your final payoff will NOT be negative.

On the next page you find a simple payoff-example:

Consider the following payoff-example:

Example:

- If you click away 20-40% of the boxes Person A receives 300 points.
- This means, Person A has a total of 300 + 200 = 500 points including his / her initial endowment of 200 points.
- If his / her feedback to you after the clicking-task is 20-40%, then you receive 150 points from Person A's 500 points, i.e. Person A is left with 500 - 150 = 350 points.
- Person A's feedback also implies that you have a total of 150 + 200 = 350 points including your initial endowment of 200 points.
- If you than reduce Person A's payoff by 40 points in reaction to his / her feedback, this costs you 10 points from your 350 points.
- Given this, Person A's final payoff (in points) is 300 + 200 – 150 – 40 = 310 points including the initial endowment of 200 points.
- Your final payoff (in points) is 150 + 200 – 10 = 340 points including the initial endowment of 200 points.

At the end of the experiment:

At the end of the experiment there will be a small questionnaire to fill out. Furthermore, payoffs will be calculated - on the basis of your performance, your feedback and Person B's reaction to it – a summary of all this will be displayed on your screen. Please remain seated until your client number (which you will find on your summary screen) is announced. Upon announcement please come forward so that you can be paid.

Please raise your hand now, if you have any questions. Otherwise, please answer the questions on the screen and press "Next" to start stage 2 of the experiment.

Supplementary Material - Not For Publication

Appendix D

Screen-shots

In this Appendix you can find a selection of the screen-shots. The full set of screen-shots can be obtained from the authors upon request.

[Figures A.4 - A.8]

Supplementary Material - Not For Publication

Appendix E

Data

[Tables A.1 and A.2 here]

Table A.1—: Descriptive Statistics A: *incentive* treatment

Sessions	Number of A-Persons	Number of B-Persons	Number of Appeared Boxes	Average Number of Clicked Boxes
50 seconds:	17	17	400	101
90 seconds:	17	17	400	193.35
120 seconds:	9	9	400	202

Table A.2—: Descriptive Statistics B: *incentive* treatment

Sessions	Average Ratio of A's evaluation / Feedback	Average Ratio of B's evaluation / Feedback
50 seconds:	1.48	1.85
90 seconds:	1.20	1.58
120 seconds:	1.05	1.26

Looking at Table A.1 one can see that in the *incentive* treatment there were 17 A- and B-Persons in the 50 second session, 17 A- and B-Persons in the 90 second sessions and 9 A- and B-Persons in the 120 second session. In all sessions 400 boxes appeared in the effort task and the average number of boxes increased the more time B-Persons had.
Looking at Table A.2 one can see that in the *incentive* treatment the average ratio of As' evaluations to feedback is for all sessions above 1. This means, the evaluations of A-Persons were on average better than their feedback in the *incentive* treatment. As can easily be seen, the same is true for the average ratio of B-Persons' evaluations and feedback. Interestingly, the ratio is higher in case of B-Persons. This indicates that on average B-Persons had a better evaluation of their own work as A-Persons.

[Tables A.3 and A.4]

Looking at Table A.3 one can see that in the *flat* treatment there were 12 A- and B-Persons in the 50 second session, 27 A- and B-Persons in the 90 second sessions and 13 A- and B-Persons in the 120 second session. In all sessions 400 boxes appeared in the effort task and the average number of boxes increased the more time B-Persons had. The increase is actually a bit sharper than in the *incentive* treatment.

Looking at Table A.4 one can see that in the *flat* treatment the average ratio of As' evaluations to feedback is above 1 only for the 90 second sessions. Generally, the ratio is

Table A.3—: Descriptive Statistics A: *flat* treatment

Sessions	Number of A-Persons	Number of B-Persons	Number of Appeared Boxes	Average Number of Clicked Boxes
50 seconds:	12	12	400	87
90 seconds:	27	27	400	196
120 seconds:	13	13	400	247.5

Table A.4—: Descriptive Statistics B: *flat* treatment

Sessions	Average Ratio of A's evaluation / Feedback	Average Ratio of B's evaluation / Feedback
50 seconds:	0.841	0.552
90 seconds:	1.082	1.084
120 seconds:	0.989	1.014

lower than the comparable ratio for the *incentive* treatment. With regard to the average ratio of B-Persons' evaluations and feedback, one can see that in the fast 50 second sessions the ratio is even lower than the average ratio of As' evaluations to feedback. In the other two sessions it is higher, but it is generally lower compared to the *incentive* treatment.

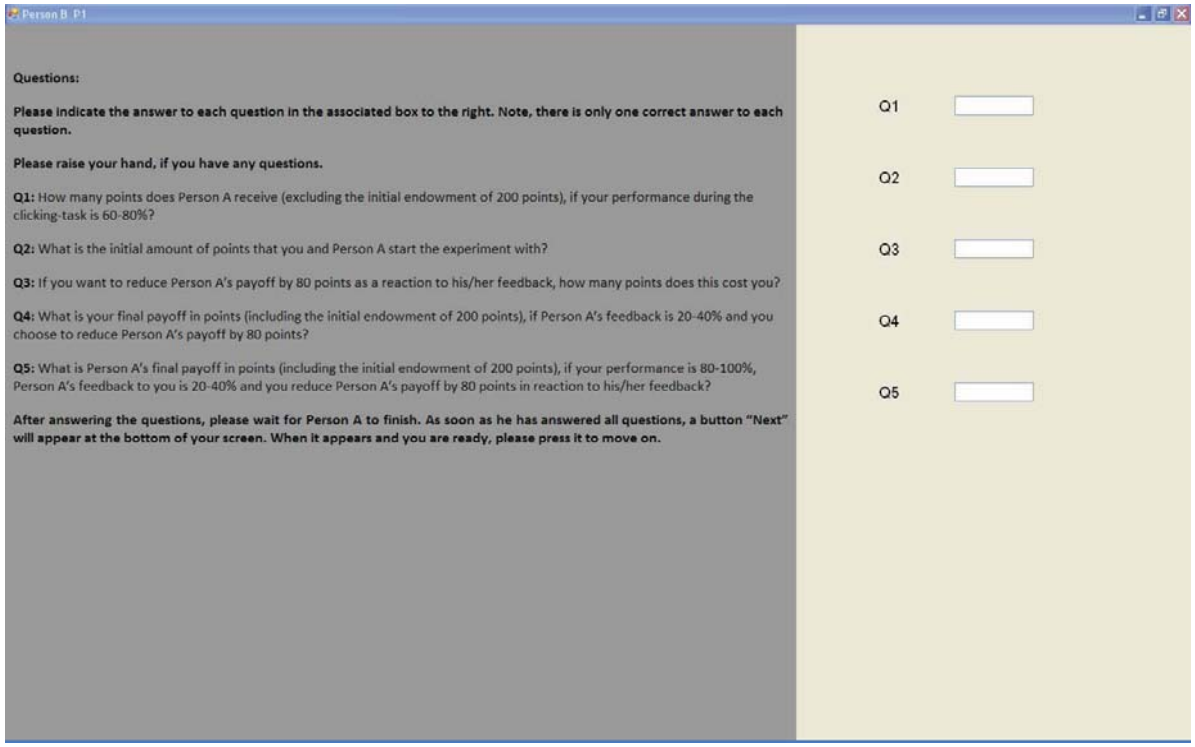

Figure A.4. : Control Questions: Person B

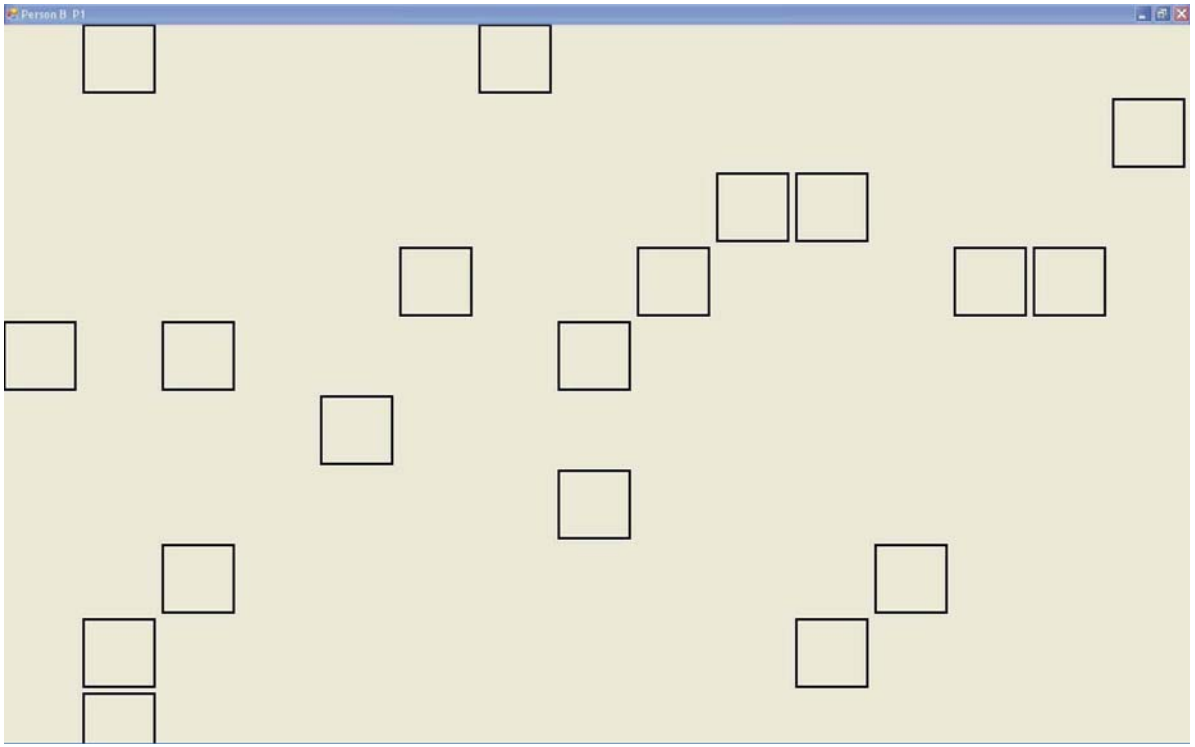

Figure A.5. : Clicking Task: Person B

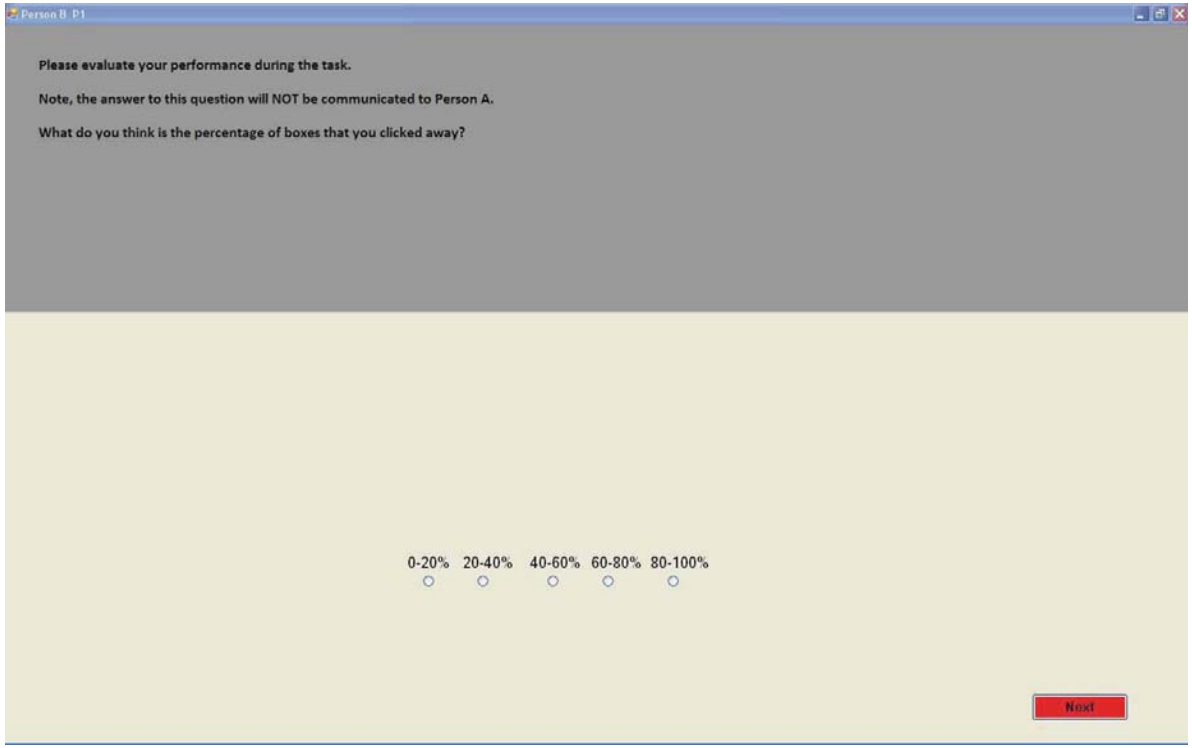

Figure A.6. : Evaluation: Person B

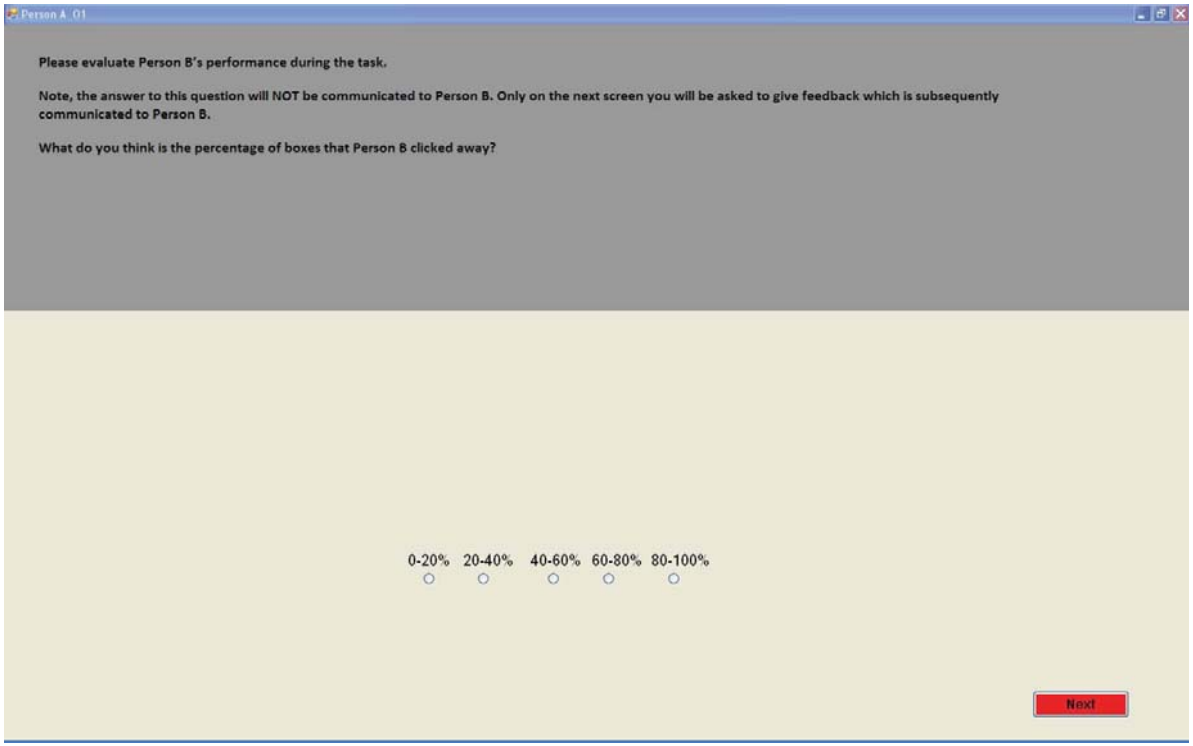

Figure A.7. : Evaluation: Person A

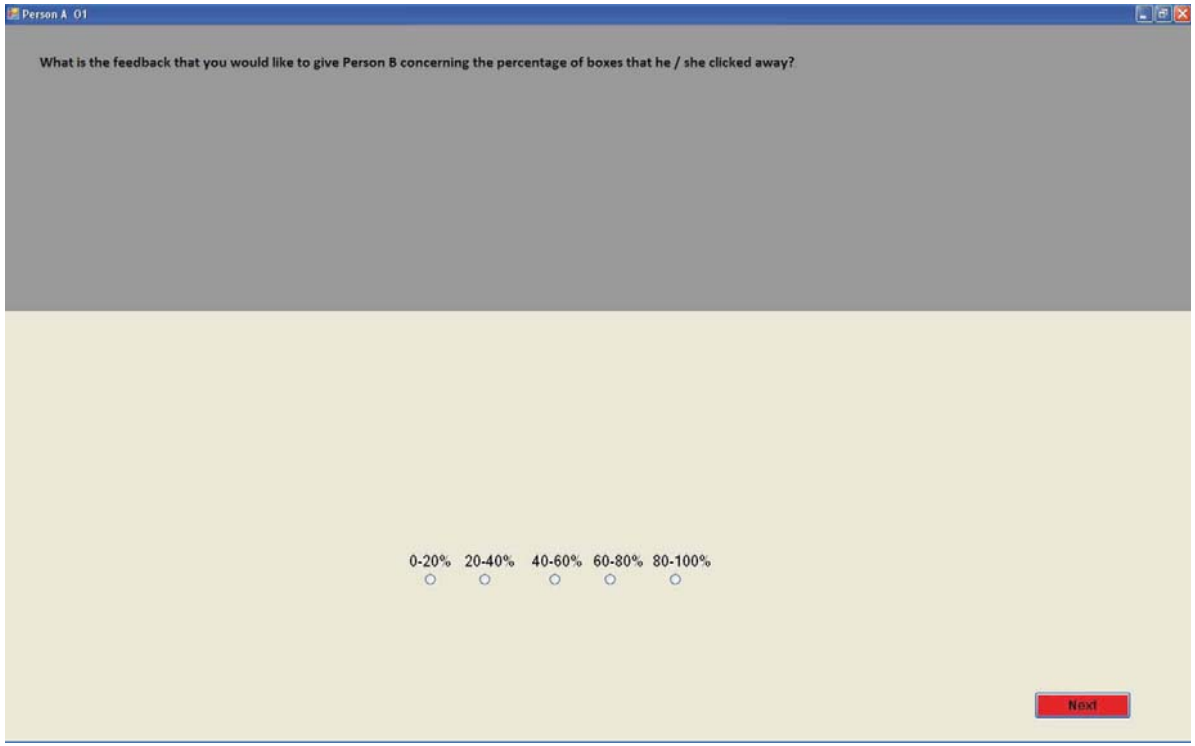

Figure A.8. : Feedback: Person A