Ederer, Florian; Fehr, Ernst

**Working Paper**

# Deception and incentives: how dishonesty undermines effort provision

IZA Discussion Papers, No. 3200

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

*Suggested Citation:* Ederer, Florian; Fehr, Ernst (2007) : Deception and incentives: how dishonesty undermines effort provision, IZA Discussion Papers, No. 3200, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/34795

DISCUSSION PAPER SERIES

IZA DP No. 3200

# Deception and Incentives: How Dishonesty Undermines Effort Provision

Florian Ederer
Ernst Fehr

December 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

# Deception and Incentives: How Dishonesty Undermines Effort Provision

**Florian Ederer**
*MIT*

**Ernst Fehr**
*University of Zurich*
*and IZA*

Discussion Paper No. 3200
December 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3200
December 2007

# ABSTRACT

# Deception and Incentives: How Dishonesty Undermines Effort Provision*

In this paper we show that subtle forms of deceit undermine the effectiveness of incentives. We design an experiment in which the principal has an interest in underreporting the true performance difference between the agents in a dynamic tournament. According to the standard approach, rational agents should completely disregard the performance feedback of self-interested principals and choose their effort level as if they had not been given any information. However, despite substantial underreporting many principals seem to exhibit lying aversion which renders their feedback informative. Therefore, the agents respond to the feedback but discount it strongly by reducing their effort relative to fully truthful performance feedback. Moreover, previous experiences of being deceived exacerbate the problem and eventually reduce average effort even below the level that prevails in the absence of any feedback. Thus, both no feedback and truthful feedback are better for incentives than biased feedback.

JEL Classification: D83, C92, M12

Keywords: deception, dishonesty, communication, cheap talk, dynamic tournaments

Corresponding author:

Ernst Fehr
Institut für Empirische Wirtschaftsforschung
Universität Zürich
Blümlisalpstrasse 10
CH-8006 Zürich
Switzerland
E-mail: efehr@iew.uzh.ch

* We would like to thank David Abrams, Daron Acemoglu, David Autor, Joshua Angrist, Arthur Campbell, Mathias Dewatripont, Daniel Gottlieb, Glenn Ellison, Urs Fischbacher, Xavier Gabaix, Robert Gibbons, Lorenz Götte, Bengt Holmstrom, Margaret Meyer, Whitney Newey, Catarina Reis, Johannes Spinnewijn, Jean Tirole, Christian Zehnder and seminar participants at the EEA-ESEM 2006 meetings for helpful comments as well as Thomas Epper and Franziska Heusi for outstanding research assistance.

"Distortion or inaccuracy in performance appraisal is nothing more than good management. Manager who shamelessly manipulate the performance appraisal system to achieve their primary goals (i.e., get the most possible out of the human and physical resources at their disposal) do more for their organizations than managers who follow all of the rules and who turn in accurate performance appraisals, regardless of the consequences."

Murphy & Cleveland (1995, p. 348)

# 1 Introduction

Incentives are pervasive in economic life. Without proper incentives economic exchange and production will be suboptimal. Deliberate incentive design intends to shape and enhance agents' performance and rests, therefore, always on some sort of performance appraisal. However, as the above quote from the famous performance appraisal textbook by Murphy and Cleveland (1995) indicates, principals often face the temptation to manipulate and bias performance appraisals. In fact, these authors strongly endorse biased performance appraisals and there is evidence that managers indeed provide untruthful performance feedback.[1]

In this paper we document that biased performance appraisals may thoroughly undermine incentives. We show that even quite subtle forms of deceit can be detrimental for incentives. It is obvious that incentives break down if promised payments are not delivered although the agent performed well. However, here we show that deceitful principals undermine incentives even if they cannot manipulate the payments to the agents, that is, even if the agents are correctly paid according to their output performance.

In order to study the impact of deception on incentives we examine behavior in a dynamic tournament. Tournaments are ubiquitious in economic organizations because performance is often rewarded by promoting high-performing employees. Such tournaments are inherently dynamic as during the contest the principal often observes some interim performance measures that she can use to provide feedback to the agents. In these settings the principals often face strong incentives to provide wrong or misleading feedback. Based on a simple two-stage model of a dynamic tournament with two competing agents we show that the agents provide higher effort in the second stage of the tournament if the absolute output difference between the agents after the first stage is smaller. This result is intuitive as agents whose interim output performance is

[1] Longecker, Sims and Gioia (1987), for example, interviewed 60 upper-level executive and found evidence of deliberate manipulation of formal performance appraisal ratings by executives. Executives adjusted or manipulated employee ratings in order to accomplish their own goals or agendas, most often in an attempt to increase performance by employees.

closer to each other have a stronger incentive for subsequent performance because the likelihood of winning the overall tournament is still high for both of them. Therefore, if the performance difference is not directly observable to the agents a self-interested principal has an incentive to report smaller than actual output differences to the agents.[2]

We examined the prediction that agents' effort after feedback is decreasing in the absolute output difference by conducting a laboratory experiment with truthful information feedback (the Truthful Feedback condition, TF), that is, the agents knew that they received correct feedback information after the first stage of the tournament. The actual effort behavior in this treatment is remarkably close to the prediction of the above model. Thus, based on this result, we know that if principals are free to provide false feedback they face an incentive to underreport the actual output differences because this increases the effort of gullible agents. However, rational agents will anticipate the principals' incentive to underreport and, therefore, treat the principals' feedback information as completely uninformative.[3]

Thus, if agents are rational and the principals' messages are indeed completely uninformative we have a stark prediction: in a condition in which the principals provide feedback, the agents' effort is identical to a condition in which the agents receive no feedback at all. In other words, the principals' messages are without any consequences on effort. In order to test this prediction we conducted two further treatments: a treatment in which the principals were free to choose the feedback information (the Principal Feedback treatment, PF) and a treatment without any information feedback after the first stage (the No Feedback treatment, NF). We deliberately chose the parameters of the experiment in such a way that the equilibrium effort in the second stage of the NF treatment is higher than the average effort after feedback in the equilibrium of the TF treatment. This means that if the cheap talk hypothesis holds, the agents in the PF condition should provide a higher average effort after feedback than in the TF condition.

In sharp contrast to the cheap talk hypothesis, the feedback by the principals was informative and had a sizeable negative effect on the agents' effort choices. However, as the principals also frequently sent false feedback the agents strongly discounted the principals' messages: when a principal reported an absolute output difference of $\Delta$ the agents behaved as if the true output difference is not $\Delta$ but $k\Delta$ with $k > 1$. Therefore, for any given output difference reported by a

[2] As employees' performance is often difficult to assess, representing a combination of many "soft" factors such as communication and team leadership performance, innovativeness, initiative and originality, the principal's assessment of these factors will finally decide the outcome of the tournament. It is therefore often the case that the principals' are better informed about performance and, in particular, performance differences among employees than the employees themselves. In addition, employees may suffer from self-serving biases in their performance assessment which renders balanced assessments of employee performance by human resource departments even more important.

[3] We call this in the following the "cheap talk hypothesis" because in our setting the fact that "talk is cheap" implies that all communication among self-interested players is uninformative.

principal in the PF treatment, the agents provided much less effort compared to the effort level they chose in response to truthfully reported output differences in the TF treatment. This effort response to principals' messages also had implications for average effort levels after feedback. Despite the fact that the agents in the PF treatment face a much higher frequency of small reported output differences their average effort after feedback is eventually even lower than in the TF and the NF treatment because the agents discount the principals' information quite strongly.

One driving force behind these effort patterns is the fact that principals' messages were not merely cheap talk, that is, they were partially informative. Some principals provided feedback that was close to the truth while others reported consistently much smaller output differences than the actual ones. On average, an increase in the reported output difference by one unit was associated with a true increase in the output difference by two units. Therefore, it was rational for the agents to respond to variations in the reported output differences in the PF condition even more strongly than to variations in truthful output differences in the TF condition. Interestingly, this is exactly what we observe in these conditions.

The negative effect of deceitful messages on effort was not limited to the post-feedback stage. Over time the agents' average effort in the PF condition even declines in the pre-feedback stage, indicating a pervasive negative effect of principals' deceitful messages on the overall effort level. This effect seems to be driven by the frequency of deceitful messages that an agent has received: agents who faced a higher frequency of small reported output differences in past interactions were significantly more likely to choose a low effort level even in the pre-feedback period of a tournament. This finding indicates a more general undermining of incentives through deception because, as we show in the paper, a rational agent's first stage effort should not be affected by the expectation that a principal sends false feedback about the first stage output difference.

In the final part of the paper we show that a model of lying aversion provides a plausible explanation for many key features in our data. This model assumes a heterogeneous population of lying averse subjects: some subjects are assumed to have higher psychic costs of lying than others. The assumption of heterogeneous lying aversion is supported by previous work of Gneezy (2005), Mazar and Ariely (2006) and Sanchez-Pages & Vorsatz (2007) which provides evidence for heterogeneous preferences for truth-telling. Heterogeneity is also supported by our finding that strongly Machiavellian subjects, that is, subjects with a high degree of selfishness and opportunism, lie significantly more compared to subjects that score low on the Machiavelli scale.

Principals who exhibit some degree of lying aversion are willing to report the truth if the costs of truth-telling are not too large. In our setting the costs of truth-telling arise because principals who report a lower output difference to the agents will on average induce a higher second stage performance in the tournament. Empirically it turns out that if the true output

difference is already "small" the benefits of reporting an even smaller output difference are small. Thus, for small true output differences a model based on heterogeneous principals predicts a higher frequency of truth-telling, which is indeed the case. Furthermore, the existence of lying averse principals makes it rational for the agents to attribute some informative content to the principals' messages while the existence of principals with little or no lying aversion renders it rational to discount the message of a principal partially because the agent in the experiment does not know what principal he faces in a given period. Thus, the partial informativeness of the principals' messages and the subsequent response of the agents follows directly from heterogeneity in lying aversion and the assumption that agents behave rationally.

Finally, heterogeneous lying aversion can also explain why agents who experienced deceitful messages more often, provide lower second stage effort than agents who experience fewer deceitful messages. A plausible reason for this fact is that agents who experience deceitful messages more often have less trust in the principals, that is, they update the probability that they face a deceitful principal. As a consequence, they discount the principals feedback information more strongly and provide a lower effort for a given reported output difference. We believe that this feature in our data captures an important effect that is also likely to be relevant in real organizations.

It is ironic that the negative impact of the PF treatment on effort is based on the fact that some principals also send messages that are correct or close to the truth. If the principals' messages had been completely unreliable it would have made no sense for the agents to take the principals' messages seriously – not even in a discounted way. In this case it would have been rational to completely disregard principals' messages and choose the same high second stage effort as in the NF treatment. Thus, we observe here an interesting non-monotonous effect of feedback information on tournament incentives: no information or completely uninformative feedback and fully truthful information is better than partially informative feedback. This finding may explain a stylized fact from the organizational behavior literature, namely, that performance feedback is often quite uninformative (Meyer, Kay & French (1965), Beer (1987), Gibbs (1991)). In view of the difficulty of forcing managers to provide truthful performance feedback to the employees it might be better to provide no or merely uninformative feedback.

Our paper is related to the theoretical and experimental literature on the efficiency and effectiveness of tournament incentives (Lazear & Rosen (1981), Lizzeri, Meyer & Persico (1999), Aoyagi (2004), Ederer (2005), Bull, Schotter & Weigelt (1987), Schotter & Weigelt (1992), Harbring, Irlenbusch, Kräkel & Selten (forthcoming)). However, none of these papers provides an empirical examination of the impact of principals' lying incentives on effort provision in a dynamic tournament. Our paper is also related to the seminal theoretical work of Crawford & Sobel (1982) and Crawford (2003) on strategic information transmission and experimental work on sender-receiver games where the interests of the sender and the receiver of a message

are in conflict (e.g., Dickhaut, McCabe & Mukherji (1995), Blume, DeJong, Kim & Sprinkle (1998)). Central aspects of Crawford & Sobel's theory of information transmission have found strong support in papers such as Cai & Wang (2006), Wang, Spezio & Camerer (2006) and Sanchez-Pages & Vorsatz (2007). One intriguing empirical finding in these papers concerns overcommunication by the senders, i.e., the fact that the senders' messages are too informative about the true state of the world relative to the equilibrium with rational and self-interested players. Moreover, the work of Sanchez-Pages & Vorsatz (2007) suggests that this overcommunication is at least partly driven by preferences for truth-telling.

None of the above papers examines the impact of strategic information transmission on the effectiveness of performance incentives. We are also not aware of any other work that compares agents' effort behavior in situations where they receive no information about the "state of the world", fully truthful information and information that can be biased because it comes from an interested principal. It is the comparison between these three situations which enables us to identify the extent to which deceitful messages undermine incentives for effort provision. The impact of deception on effort provision is so pervasive that it does not only change the agents' direct effort response to the performance feedback at the second stage of the tournament but it also causes a large decline in first stage effort levels, supporting the conclusion that no performance feedback as well as truthful performance feedback is superior to an arrangement where the principal is not constrained to tell the truth.

In view of the profound effects of heterogenous preferences for honesty in our setting we believe that contract theory and, more generally, economics could benefit from taking these preferences into account. There is a large class of economic problems that are characterized by lying incentives and as our results suggest heterogenous preferences for honesty may affect behavior in these situations in important ways.[4]

The remainder of the paper is organized as follows. Section 2 proposes a model of the effect of interim information and communication in dynamic tournaments. Section 3 describes the experiment and Section 4 presents the experimental results which we interpret in Section 5.

[4]For example, a used car salesman has an incentive to lie about the quality of the car he is attempting to sell (Akerlof 1970) or a regulated firm has an incentive to claim high costs in order to receive higher transfer payments (Laffont & Tirole 1993). In the absence of reputation incentives financial brokers will be tempted to recommend stocks or funds that further their own interests but not necessarily the clients' interests (Davis 2004) and tax payers have an incentive to underreport their income. (Allingham & Sandmo 1972). Similarly, in the insurance market the insured parties have incentives to overstate the value of claims to insurance companies. In all these cases, economic models typically assumed that a player lies whenever it is in his or her economic interest to do so. However, the existence of heterogeneous lying aversion may, for example, change the optimal insurance contract because insurance companies may want to induce self-selection of liers and honest people into different contracts. Likewise, employers may want to screen people for special jobs on the basis of the strength of an employee's lying aversion.

Section 6 concludes. Appendix A contains additional proofs and Appendix B proposes a simple model of lying aversion. Appendix C shows that alternative social preference theories cannot explain the experimental results.

# 2 A Simple Model of Dynamic Tournaments

In this section we present a model of dynamic tournaments when there is no feedback (NF), when feedback is truthful (TF) and when principals have the opportunity to provide wrong feedback (PF). For additional technical details and less restrictive modeling assumptions the reader is referred to Ederer (2005).

Consider a tournament for a fixed prize between two risk-neutral agents $i = A, B$ which takes place over 2 stages, $t = 1, 2$. The utility of winning the contest is $P$ and the utility of losing is $p$. Agent $i$'s output in stage $t$ is given by $x_t^i = e_t^i + \varepsilon_t^i$ where $e_t^i$ is the privately chosen effort level and $\varepsilon_t^i$ is an error term. At the end of stage 2, the principal aggregates the scores from both stages to determine the winner of the contest. Agent $A$ wins the contest if his accumulated output is greater than that of agent $B$, i.e. if $x_1^A + x_2^A > x_1^B + x_2^B$ agent $A$ wins and agent $B$ wins if the reverse inequality holds.

Each agent's effort $e_t^i$ is his private information and is not observed by the other contestant. We assume that the noise difference $\varepsilon_t^A - \varepsilon_t^B$ is independently normally distributed for each $t$ with cumulative density function $F(\cdot)$ and density $f(\cdot)$. Let $G(\cdot)$ and $g(\cdot)$ denote the distribution and density functions of the sum of the noise differences $\varepsilon_1^A - \varepsilon_1^B + \varepsilon_2^A - \varepsilon_2^B$. Note that the noise difference in each stage and thus the sum of noise differences in both stages are normally distributed, i.e.,

$$\begin{aligned} \varepsilon_t^A - \varepsilon_t^B &\sim N(0, \sigma^2) \\ \varepsilon_1^A - \varepsilon_1^B + \varepsilon_2^A - \varepsilon_2^B &\sim N(0, 2\sigma^2). \end{aligned}$$

When exerting effort in stage $t$ agent $i$ incurs a cost $c(e_t^i)$. The payoff to agent $i$ is given by

$$U^i = p + (P - p)\Pr(x_1^i + x_2^i > x_1^j + x_2^j) - c(e_1^i) - c(e_2^i). \tag{1}$$

We intend to contrast the following three settings: no feedback, truthful feedback and feedback by self-interested principals. In the no feedback scenario, neither of the agents knows the first stage output difference $\Delta$ when choosing second stage effort. The first stage output difference is defined in the following way:

$$\Delta \equiv x_1^A - x_1^B = e_1^A + \varepsilon_1^A - e_1^A - \varepsilon_1^B.$$

In the truthful feedback scenario, the first stage output difference $\Delta$ is truthfully revealed to the two contestants. Hence, each agent learns the first stage outcome $\Delta$ before choosing second

stage efforts. Finally, in the last scenario feedback is given to the agents by self-interested principals who do not face any truth-telling restrictions. This third case falls between the two polar opposites of no feedback and truthful feedback. When the feedback the agents receive from principals is completely uninformative then rational agents will behave as in the no feedback setting, whereas when it is fully informative they will choose efforts as in the truthful feedback scenario. As we shall see, the predictions for effort choice when agents do not receive any feedback are identical to those when feedback is given by self-interested principals since feedback is completely uninformative in the latter setting.

We denote equilibrium effort by $e^*$, $\tilde{e}$ and $\hat{e}$ when agents receive no feedback, truthful feedback and feedback from principals, respectively.

## 2.1 Dynamic Tournaments without Feedback

The probability of winning the contest for a given level of effort choices is $G(\cdot)$. Therefore the maximization of equation (1) is tantamount to the maximization of

$$p + (P - p)G(e_1^i + e_2^i - e_1^j - e_2^j) - c(e_1^i) - c(e_2^i).$$

The resulting first order condition with respect to $e_t^i$ for an interior solution for both stages is given by

$$c'(e_t^i) = (P - p)g(e_1^i + e_2^i - e_1^j - e_2^j).$$

In the unique symmetric Nash equilibrium of the game this condition simplifies to

$$c'(e^*) = (P - p)g(0). \tag{2}$$

Since the sum of the noise differences is normally distributed according to $N(0, 2\sigma^2)$, the effort in both stages of the NF condition is given by

$$c'(e^*) = \frac{P - p}{2\sigma\sqrt{\pi}}. \tag{3}$$

## 2.2 Dynamic Tournaments with Truthful Feedback

### 2.2.1 Second Stage

At the beginning of the second stage agent $i$ knows the first stage output difference $\Delta$ so that his overall utility can be written as

$$U^i = p + (P - p)F(e_2^i - e_2^j + \Delta) - c(e_2^i)$$

where $F(e_2^i - e_2^j + \Delta)$ is the probability of winning the contest for given second stage effort levels and a given output difference $\Delta$. In the unique symmetric equilibrium of the second stage each agent $i$ chooses $e_2^i$ such that $\frac{\partial U^i}{\partial e_2^i} = 0$ which yields the following first order condition

$$c'(\tilde{e}_2) = (P - p)f(\Delta). \tag{4}$$

Since the second stage error difference is normally distributed the effort in the second stage of the TF condition is given by

$$c'(\tilde{e}_2) = \frac{P - p}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\Delta}{\sigma}\right)^2\right]. \tag{5}$$

Equation (5) implies that the second stage equilibrium effort is decreasing in the absolute magnitude of the first stage output difference $\Delta$. When the difference between the two agents is large, the intensity of competition is low and thus the effort of both agents is low while effort is high when agents' outputs in the first stage are close to each other.

### 2.2.2 First Stage

Having determined the equilibrium effort choice of the second stage we can solve for the agents' effort choice in the first stage. Agent $i$ chooses $e_1^i$ to maximize

$$U^i = p + (P - p)G(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j) - c(e_1^i) - E\left[c(\tilde{e}_2^i)\right]$$

where the second stage effort levels $\tilde{e}_2^i$ and $\tilde{e}_2^j$ depend on $\Delta$. Note that the first stage effort $e_1^i$ affects the agent's utility $U^i$ in three ways. First, there is a direct effect of $e_1^i$ on $U^i$ because a higher $e_1^i$ increases the probability of winning and increases the first stage cost of effort. Second, a higher $e_1^i$ changes $\Delta$ which in turn changes the second stage equilibrium effort $\tilde{e}_2^i$ according to equation (5). Finally, a higher $e_1^i$ also affects the opponent's effort choice $\tilde{e}_2^j$ because $j$ also responds to the change in $\Delta$ with a change in $\tilde{e}_2^j$. Thus, agent $i$'s first order condition can be written as

$$\frac{dU^i}{de_1^i} = \frac{\partial U^i}{\partial e_1^i} + \frac{\partial U^i}{\partial \tilde{e}_2^i}\frac{d\tilde{e}_2^i}{de_1^i} + \frac{\partial U^i}{\partial \tilde{e}_2^j}\frac{d\tilde{e}_2^j}{de_1^i}. \tag{6}$$

The second term in equation (6) is zero because for any given output difference $\Delta$ agent $i$ chooses the utility maximizing level of $\tilde{e}_2^i$ so that $\frac{\partial U^i}{\partial \tilde{e}_2^i} = 0$. It turns out that because of the symmetry of the normal distribution and the symmetry of the equilibrium the third term in equation (6) is also zero. Since the proof of this assertion involves some cumbersome manipulations we relegate it to Appendix A (see also Ederer (2005)). Thus, as the second and the third term in equation (6) are zero we are left with

$$\frac{\partial U^i}{\partial e_1^i} = (P - p)g(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j) - c'(e_1^i)$$

which simplifies in the symmetric equilibrium to

$$c'(\tilde{e}_1) = (P - p)g(0). \tag{7}$$

### 2.2.3 Comparison

From the first order conditions (2) and (7) it follows that first stage efforts under the NF and the TF condition are equal, i.e., $e^* = \tilde{e}_1$. Due to the symmetric effort choice of the two agents in the first stage of the TF condition, equation (5) for second stage effort in the TF condition reduces to

$$c'(\tilde{e}_2) = \frac{P - p}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_1^A - \varepsilon_1^B}{\sigma}\right)^2\right]. \tag{8}$$

Equation (8) shows that second stage effort under truthful feedback, contrary to all other previous effort choices, is random. When $\Delta = \varepsilon_1^A - \varepsilon_1^B = 0$, the exponent of the right-hand side of (8) is zero and thus second stage effort $\tilde{e}_2$ is highest at this point. Furthermore, at $\Delta = 0$ second stage effort $\tilde{e}_2$ is higher than $e^*$ as can be seen from comparing the right-hand sides of equations (8) and (3).

In Appendix A we further show that expected second stage effort under the TF condition is equal to effort in the NF condition, $E[\tilde{e}_2] = e^*$, if the cost of effort is quadratic (and marginal cost is linear), $E[\tilde{e}_2] > e^*$ if marginal cost is concave and $E[\tilde{e}_2] < e^*$ if marginal cost is convex. Intuitively, the convexity of marginal costs implies that due to the variability of output differences the expected marginal costs for second stage effort would be higher in the TF condition compared to the NF condition if the agent chose first and second stage effort levels that are equal in expectation. However, in equilibrium marginal costs of first and second stage effort are equalized and therefore $E[\tilde{e}_2]$ is below $e^*$. This result is important for our purposes because in our experiments we implemented convex marginal cost of effort, which implies that the average second stage effort under truthful feedback, $E[\tilde{e}_2]$ is predicted to be lower than the second stage effort under no feedback $e^*$.

Our analysis above shows that the first stage effort under no feedback (NF) and under truthful feedback (TF) are identical, i.e., the fact that agents receive information about the first stage output difference does not affect first stage efforts. Aoyagi (2004) shows that this result holds more generally, in particular, for the case where feedback information about $\Delta$ is only partially informative. That is, the equality of first stage efforts does not only hold for the two extreme cases of full information and no information feedback, but also for all intermediate cases of information release.

## 2.3 Dynamic Tournaments with Feedback by Self-interested Principals

Consider the introduction of a risk-neutral self-interested principal. The principal's role is to transmit information about the output difference between the two contestants after the first stage. Her period payoffs are a linear function of the two contestants' outputs in both stages minus the wage costs for the two agents, that is, her payoff is given by

$$U^P = \theta \left( x_1^A + x_1^B + x_2^A + x_2^B \right) - P - p,$$

where $\theta$ is a positive constant and it is commonly known that $\theta > 0$.

The principal observes the first stage output difference of the two contestants and must then send a private message $\hat{\Delta}^i$ to each agent $i$ about the output difference. We assume that the principal's message is non-verifiable and the principal is not obliged to report truthfully to either of the agents. Since the principal's message does not affect her utility function the message is *cheap talk* (Crawford & Sobel 1982).

For the principal feedback scenario we denote equilibrium effort by $\hat{e}$ and reports to agent $i$ of the output difference by $\hat{\Delta}^i$.

### 2.3.1 Communication and Effort Equilibrium

In the above setting for all $\theta > 0$ the unique Perfect Bayesian Equilibrium is a *babbling equilibrium* in which all messages sent to the agents by the principal are uninformative. To see this, consider the case where agents (naively) believe the principal's message. In this case, the optimal message is $\hat{\Delta}^i = 0$ because according to equation (5) this maximizes the agent's effort. More generally, for any given belief the agents hold about the principal's message the principal will want to send the message that leads the agent to believe that the absolute value of the output difference is as low as possible as this will maximize the agent's effort. As a result, for any given first stage output difference $\Delta$ the reported output difference is independent of the actual output difference. Hence, all communication is uninformative and no information is transmitted in equilibrium.

Note that the present model does not make clear predictions as to what messages will actually be sent in equilibrium. As communication is completely uninformative all messages will be ignored by agents. That is, regardless which messages principals send they will not influence agents' effort choices and so the messages sent by principals are indeterminate.

Furthermore, since the communication by principals is completely uninformative, agents ignore the messages. As a result, with regards to effort choice the model reduces to the no feedback setting analyzed in Section 2.1.

# 3 An Experimental Approach for the Study of Dishonesty and Incentives

The ideal data set for studying the effects of dishonesty on incentives and effort choices is based on a truly exogenous ceteris paribus variation in the degree of credibility of information. The exogenous variation allows the researcher to make causal inferences on the impact of different degrees of information credibility. Such a data set permits, for instance, the examination of how the possibility of dishonest reporting behavior affects communication behavior of principals and the effort choices of contestants. The problem is, however, that it seems almost impossible to find or generate field data to approximate this ideal data set. In real life situations, there are few completely truthful or untruthful signals and one rarely knows how credible a given piece of information is. Thus, it is particularly difficult to measure and analyze the effects of deception on incentives and effort provision. Experiments designed suitably allow for causal inferences, however – because one can implement exogenous ceteris paribus variations in the extent of information manipulation – thus overcoming some of the measurement and endogeneity problems present in the field data. In particular, the following experiment allows us to study behavior in the three settings where agents do not receive any feedback, where they receive completely truthful feedback and where the feedback they receive can be manipulated by principals.

## 3.1 Experiment Design

We implemented the following three treatment conditions in order to examine how dishonesty affects communication and effort choices. In the first condition, which we call the no feedback condition (NF condition), the contestants were given no information about first stage outcomes before the second stage. In the second condition, which we call the truthful feedback condition (TF condition), the experimenter exogenously enforced truthful feedback about the first stage to the contestants and the subjects were informed about this. In the third condition, which we call the principal feedback condition (PF condition), feedback about the first stage was given by principals to the contestants; principals observed outcomes in the first stage and had to send non-verifiable messages about the first stage output levels and the output difference separately and privately to each agent. The principals were free to report any output level and output difference irrespective of the actual outcomes and the agents knew this. Thus the design of the experiment was chosen in such a way as to analyze differences in behavior resulting from differences in information transmission.

By comparing effort choices across the NF and TF conditions the effect of credible information on effort choices can be examined. Furthermore, these two conditions provide the benchmarks against which the outcomes of PF condition can be measured. Our major re-

search questions are addressed by analyzing communication behavior in the PF condition and by comparing effort choices across the PF, TF and NF conditions.

There were 15 periods in each of the three conditions. In all conditions the subjects were anonymously and randomly matched at the beginning of each period. Each period had two stages. In stage 1, workers chose their first stage efforts $e_1^i$. After choosing effort $e_1^i$, a normally distributed random variable $\varepsilon_1^i$ was added to each agent's effort choice. An agent's first stage output $x_1^i$ was the sum of his effort choice $e_1^i$ and the realization of his random variable $\varepsilon_1^i$. After the end of the first stage, in the PF condition principals received information about the first stage output levels and the output difference between the agents. The principals privately sent information about the first stage output levels and the output difference to each of the two agents they were paired with. Before the beginning of the second stage of the TF and the PF condition, each contestant received information about his own first stage output $x_1^i$, the first stage output of his opponent $x_1^j$ and the associated output difference $\Delta$. In the TF condition this information was relayed to each agent directly and truthfully through the computer, whereas it was sent by a principal in the PF condition. The agents in the PF condition knew that the principal could only determine the feedback they received about first stage outcomes and that she could neither determine the overall outcome of the tournament nor influence the allocation of prizes. This allocation was objectively determined by a comparison of the agents' actual total output in the two stages in all three conditions. Contestants in the NF condition did not receive any feedback about the first stage.

In stage 2, contestants in all conditions had to choose their second stage effort levels $e_2^i$ to which another random variable $\varepsilon_2^i$ was added to produce the agent's second stage output $x_2^i$. At the end of the second stage each agent received truthful feedback from the computer about the *sum* of his own first and second stage output level and the output sum of the other contestant. In addition, each agent was informed about who received the higher prize $P$. This allowed the agents to verify the validity and fairness of the prize allocation process. Since the agents received information about the sum of outputs rather than about the separate outputs in each stage they could only imperfectly infer the truthfulness of the principal's previous report about stage 1 outputs in the PF condition. Each principal received information about the agents' output levels in both stages and her own profit at the end of the second stage.

## 3.2 Parameters, Procedures, and Subject Pool

The material payoff of a contestant $i$ was given by

$$U^i = \begin{cases} P - c(e_1^i) - c(e_2^i) & \text{if } x_1^i + x_2^i > x_1^j + x_2^j \\ p - c(e_1^i) - c(e_2^i) & \text{if } x_1^i + x_2^i < x_1^j + x_2^j. \end{cases}$$

The material payoff of a principal was given by

$$U^P = \theta(x_1^i + x_2^i + x_1^j + x_2^j) - P - p.$$

The set of feasible effort levels was given by $e \in \{1, 2, \ldots, 100\}$. The exact parameter values and theoretical predictions are shown in Table 1. The parameter values were chosen so that the equilibrium effort choice did not constitute an immediate focal point such as 50 or the socially efficient effort level 52.38. According to the design parameters the equilibrium effort in the NF condition, $e^*$, equals 37 in both stages. The same equilibrium effort prevails in the first stage of the TF condition. Furthermore, as we implemented convex marginal costs, the average equilibrium effort in the second stage of the TF condition, $\tilde{e}_2$, is lower than the equilibrium effort in stage 2 of the NF and PF condition.

**Insert Table 1 here.**

The agents' payoff function, the number of contestants and principals, the exact parameterization and the fact that there were 15 periods were common knowledge. However, in order to avoid vertical fairness comparisons between the principals and the agents, the agents only had qualitative information about the principals' payoff function, i.e., they did not know the value of $\theta$. Agents only knew that the principals' payoff was increasing in the effort exerted by them, but they did not know how much the principal earned. We implemented this procedure because we were not interested in the potential impact of vertical fairness (payoff) comparisons; instead we wanted to study the impact of different information transmission mechanisms on the agents' behavior.

The experiments were programmed and conducted with the software z-Tree (Fischbacher 2007). A total of 192 subjects participated in our experiments. We conducted four sessions in each condition giving us a total of 16 matching groups which constituted the independent units of observation. A session lasted, on average, 75 minutes. Subjects were science students from the University of Zurich and the Federal Institute of Technology in Zurich. During the experiment experimental currency units (ECU) were used to keep track of monetary earnings. The exchange rate was set at 100 ECU = 1 CHF. On average, a subject earned CHF 43.40 (≈$36) in an experimental session.

## 3.3 Predictions

Our experiment allows us to test the empirical predictions of the model presented in Section 2; we can test, in particular, the effort predictions implied by the model. These predictions follow from the optimizing behavior of self-interested agents and principals and are discussed below.

The first hypothesis concerns second stage effort which should respond to feedback about the first stage output difference.

**Feedback Hypothesis**: *In the second stage of the TF condition effort is decreasing in the absolute magnitude of the first stage output difference* $\Delta$.

The predictions of the feedback hypothesis are similar to the predictions of a tournament in which one contestant has an absolute output advantage over his opponent. Since the density of the noise difference is highest at 0 (see the right-hand side of equation (5)), the greater is the advantage of one contestant the weaker are the incentives for effort, and hence the lower should be the equilibrium effort choices of both contestants. Note that if agents behave similar to this prediction in the PF condition, principals should have a strong incentive to misreport, as higher effort can be elicited by reporting an output difference that is small in absolute value.

Based on the lying incentive for principals, we can now predict and analyze behavior in the PF condition where interim information is given by principals.

**Cheap Talk Hypothesis**: *In the PF condition communication by the principals does not influence the agents' effort choice, i.e., the agents behave in the same way as in the NF condition.*

It is crucial for the cheap talk hypothesis that agents know that principals are rational and selfish. When this is the case, the contestants in the PF condition realize that the first stage output differences that principals report, only serve to elicit higher effort levels from contestants regardless of the actual first stage output differences. As a result, rational agents understand that messages are completely uninformative and therefore they should behave as in the NF condition.

Finally, based on the model in Section 2 and for the given experimental parameters (see Table 1) we can stipulate the following hypothesis.

**Average Effort Hypothesis**: *The average efforts in both stages of the NF condition and in the first stage of the TF condition are equal. Furthermore, the average effort in the second stage of the TF condition is lower than in the second stage of the NF condition.*

The equality of effort in both stages of the NF and the first stage of the TF condition follows from the identical first order conditions given in equations (2) and (7). Expected second stage effort in the TF condition, however, is lower than effort in both stages of the NF condition since we implemented convex marginal costs.

# 4 Results

We now present the key results obtained in our experiments. For all the tests and regressions that follow, we use clustering on matching groups because the matching groups constitute our independent units of observation.

## 4.1 Feedback Hypothesis

We first focus on the behavior of contestants in the second stage of the TF condition. As shown in Section 2, the theoretical model predicts that equilibrium effort is decreasing in the absolute value of the output difference arising at the first stage. Our first results show that this prediction is confirmed by the data.

**Result 1** (information feedback): *If the information about the agents' output differences after stage 1 is truthful, the agents' average response to the feedback information closely tracks the theoretical prediction. The second stage effort is decreasing in the absolute value of the output difference.*

Figure 1 shows the results of a bivariate Epanechnikov kernel regression together with the model's theoretical prediction. Average effort is remarkably close to the theoretical predictions of the model. The figure neatly indicates the strong effort response to the information feedback for larger absolute values of the output difference. Note that the experimental results are particularly close to the theoretical predictions when an agent is at a disadvantage at the beginning of the second stage. The fit is slightly worse for the case when the contestant is ahead.

Table 2 summarizes the regression results. In accordance with the theoretical predictions column (1) reports a highly significant negative coefficient on the absolute value of the first stage output difference. In column (2) we further decompose this effect by estimating separate coefficients for first stage winners and losers, both of which are highly significant and of the correct sign. Column (2) indicates that the discouraging effect on effort of the first stage output difference seems to be weaker for winners. Columns (3) and (4) show that these results are robust to additional controls for round and an interaction effect between round and the output differences.

**Insert Table 2 here.**

Taken together, Result 1 shows that agents' response to truthful feedback is remarkably close to the theoretical predictions of our model. It also powerfully demonstrates that there

are strong incentives to manipulate feedback: when the output difference $|\Delta|$ falls by one unit, the average second stage effort increases by 0.30 to 0.53 units. Hence, self-interested principals have strong incentives for lying if they have the opportunity to provide and manipulate the feedback information.

## 4.2 Cheap Talk Hypothesis

So far, we have focused on a setting in which feedback is always truthful. We are interested in what happens when information is relayed to the contestants through a self-interested principal who maximizes total effort and therefore has objectives that conflict with truth-telling behavior. We therefore turn to the analysis of the PF condition. The previous results highlight the experimental support for the standard theoretical model when there is completely credible feedback. In contrast, we will see that the theory does not perform similarly well in the present setting where there are incentives for dishonesty and information manipulation.

The theory makes a particularly stark prediction with regards to the efforts chosen by the agents when self-interested principals relay information. It predicts that all communication should be uninformative and so the effort choice of the contestants should be the same as in the case where no feedback is given to them.

**Result 2** (cheap talk effort response): (a) *The cheap talk hypothesis is unambiguously refuted because the agents' second stage effort choices respond strongly to the information provided by interested principals.* (b) *However, for a given reported output difference the agents provide considerably less second stage effort in the PF condition compared to the TF condition suggesting that agents partially discount the information provided by the principals.*

Empirical support for Result 2a comes from Figure 2 which plots the predicted effort response in stage 2 in the PF condition (which is completely flat), the actual average effort response to output differences in the TF condition (using the same Epanechnikov Kernel regression as in Figure 1) and the actual average effort response in the PF condition.[5]

The figure transparently shows the refutation of the cheap talk hypothesis since there is a very strong effort response to the reported output difference. Average effort choices in the PF condition vary considerably with the reports sent by the principals. Further support for Result 2a comes from Table 3 which reports the effects of principals' messages on the second stage effort choices of agents in the PF condition. In column (1) in Table 3 we see that the coefficient on the reported absolute output difference is negative and of similar magnitude as in

[5]Note that the range for which the Kernel regression for the PF condition can be drawn is much smaller than the range for the TF condition since in the former more than 95% of all the reported first stage output differences fall in the range between -25 and 25.

the setting where feedback was fully credible. The discouraging effect of the reported first stage output difference is even larger when we winsorize the sample to eliminate extreme outliers (24 reported output differences that fall outside the range between -30 and 30) as shown in column (2) of Table 3. In column (3) we separately estimate the behavioral effects on second stage efforts for interim winners and losers. The effect is particularly pronounced for interim losers and in particular for the winsorized sample as shown in column (4).

Since agents respond to the information given by principals in the way described above, it is perhaps not surprising that principals who lie more also earn more. Reporting a lower than actual output difference elicits higher effort and thus a higher payoff for the principal. On average, an incremental lie about the first stage output difference of one unit gives the principal an additional payoff of 2.5 ECU's.

**Insert Table 3 here.**

Figure 2 also provides first evidence for Result 2b. For a given reported output difference average effort choices in the PF condition are lower than in the TF condition. This suggests that when contestants receive information from a principal they realize, on average, that principals have an incentive to report output differences that are lower in absolute value than the actual output difference. For this reason agents may therefore adjust their effort response accordingly.

We therefore investigate in detail whether the second stage effort response to information given in the PF condition is significantly different from that observed in the TF condition. Table 4 reports second stage effort regressions with pooled data from both of these conditions. Once again, we eliminate the same extreme outliers in the specifications reported in columns (2) and (4). Here, *PRINCIPAL* is a binary variable which is equal to one for all observations in the PF condition and zero otherwise. On average, contestants exert significantly less effort for any given level of the first stage output difference when the information is given to them by a self-interested principal as shown by the large negative coefficient on the *PRINCIPAL* variable which is statistically significant in all the specifications reported in Table 4.

In addition, we are also interested in differences in the slope of the average effort response function. The coefficient on the interaction term $|\hat{\Delta}|$**PRINCIPAL* is insignificant in columns (1) and (2), but larger in magnitude in the winsorized sample shown in column (2). The point estimate in column (2) shows that for a one point increase in the absolute value of the reported output difference in the PF condition contestants reduce their effort by 0.66 units whereas they only reduce it by 0.41 in the TF condition. This stronger effort response by agents in the PF condition further compensates for the "underreporting" behavior by principals which we document in later results. As the effort response is potentially asymmetric for interim winners

and losers, we estimate separate slopes and interaction terms in columns (3) and (4). Again, the interactions with *PRINCIPAL* are not significant as shown in column (3). Only in the winsorized sample the slope of the effort response to information about the first stage output difference for interim losers is significantly larger in the PF condition than in the TF condition whereas the effect is insignificant for the winners as shown in column (4).

**Insert Table 4 here.**

Faced with this evidence we conclude that information feedback does indeed affect contestants' behavior even when the information can be manipulated by a self-interested principal. In particular, contestants shift their effort response downward suggesting that when choosing their effort levels they anticipate the misreporting behavior of principals. When compared to the TF condition, the feedback given by the principals clearly elicits lower effort at all reported output differences.

## 4.3 Average Effort Hypothesis

Although the agents' effort response to reported output differences is significantly lower in the PF condition it need not be the case that average effort is also lower. This is because in the PF condition the principals control the information feedback. If the principals compress the reported output differences relative to the true output differences, the agents in the PF condition may face a very different distribution of information feedback compared to the TF condition and, therefore, it is not clear whether average effort rises or falls in the PF condition relative to the other two conditions. Our next result addresses this question

**Result 3** (average effort across treatments): a) *In the PF condition the average effort in the first and the second stage decreases over time. Therefore, towards the final period average effort in the PF condition is significantly lower than in the NF and the TF condition.* (b) *In contrast, in the NF and the TF condition average effort is constant over time and, as predicted by theory, effort levels in both of the NF and the first stage of the TF condition are statistically indistinguishable. Moreover, in line with theory, the average second stage effort in the TF condition is smaller than the second stage effort in the NF condition.*

A first indication for Result 3 is provided by Figures 3 and 4. In Figure 3 we show how average effort in the first stage evolves over time across conditions. The figure shows that in the NF and the TF condition average effort is fairly stable over time while in the PF condition it strongly decreases over time. In fact, towards the final periods, average effort in the PF

condition is almost 25% lower than in the other two conditions. The impression conveyed by Figure 3 is also supported by more formal statistical tests. In the NF and the TF condition the average effort in the first three rounds is not significantly different from the respective effort levels in the final three rounds ($t$-test; $p$-value 0.259 for the NF condition; $p$-value 0.622 for the TF condition). In contrast, the same test confirms a significant difference between the first three and the last three rounds in the PF condition ($t$-test, $p$-value 0.042), so that the average effort in the PF condition is also significantly below the level in the other two conditions ($t$-test, $p$-value 0.005 for the NF condition; $p$-value 0.013 for the TF condition). All these results are also robust if we compare just the first and the last round or the first 5 and the last five rounds within each treatment or across the treatments. Finally, as predicted by the model in Section 2, the first stage average effort in the NF and the TF condition are not significantly different from each other ($t$-test, $p$-value 0.109).

A similar picture emerges when we examine how second stage effort evolves over time although the downwards trend in the PF condition starts a bit later than in the case of first stage effort. Figure 4 shows that the second stage effort in the NF and the TF condition is rather stable over time. If we compare average effort in the first three and the last three periods we find no difference ($t$-test; $p$-value 0.736 for the NF condition; $p$-value 0.642 for the TF condition) while if we perform the same comparison in the PF treatment we observe a significantly lower average effort during the final three periods ($t$-test, $p$-value 0.023). The decline in average effort in the PF condition also implies that towards the end (in the final three periods) average effort in the PF condition is significantly lower than in the NF and the TF condition ($t$-test; $p$-value 0.001 for the comparison with the NF condition; $p$-value 0.007 for the comparison with the TF condition). Figure 4 also neatly shows that the average second stage effort in the NF condition is higher than in the TF condition, a result that is also significant ($t$-test, $p$-value 0.005).

Taken together, Result 3 indicates that the model presented in Section 2 organizes the data very well in those conditions in which there is no feedback or fully truthful feedback. However, if the agents receive feedback from the principals their effort is not only smaller for any given reported output difference (Result 2) but average effort eventually declines and is, finally, significantly below the effort in the other two treatment conditions (Result 3). Apparently, the opportunity for the principals to deceive the agents is detrimental for effort provision and seems to undermine the agents' willingness to compete in the tournament. It is also interesting that the effort decline is already present in the first stage of the tournament. In fact, as Figure 3 and 4 reveal, the decline is even more pronounced in the first stage. Recall, that according to the theory presented in Section 2, the first stage effort should not be affected by the informativeness of the feedback that is given to the agents. Thus, the fact that the principals can lie in the PF treatment should not affect the agents' effort choices regardless of whether the principals' feedback is informative or not.

## 4.4 Principals' Communication Behavior

Result 2 and 3 unambiguously show that the agents' behavior in the PF condition differs substantially from the theoretical predictions. We next ask what drives these departures from the theory. One potential explanation for agents' second stage effort response documented in Result 2 is that feedback is informative. We therefore investigate the reporting behavior of principals in the PF condition. In this setting, the theory makes the particularly stark prediction that communication by the principals should be completely uninformative since the preferred message of the principal does not vary with the output difference between the two agents. A selfish principal unconstrained by honesty norms only intends to maximize her payoff and therefore, regardless of the actual output difference, she will report the output difference that for a given belief of the agent maximizes the agent's second stage effort response. For example, if the beliefs held by the agent are such that he exerts higher effort when the reported output difference is smaller in absolute value, then a profit-maximizing principal should report an output difference close to zero.

As shown in Result 4 the experimental evidence is at odds with the theoretical prediction that the information provided by the principals is completely uninformative.

**Result 4** (feedback informativeness): *The principals' messages are partly informative for the agents because higher reported output differences are associated with higher actual output differences.*

Figure 5 shows a scatter plot of actual versus reported first stage output differences in the interval $[-30, 30]$ for both variables as well as an OLS regression line within this range of actual output differences. The shallow slope of the regression line is a result of the compression of reported output differences relative to actual output differences. Although there is a large amount of untruthful reporting there is also a substantial number of observations where principals report truthfully. In particular, for small output differences in the interval $[-10, 10]$ many observations are on the truthtelling line.[6]

Since agents do not know what kind of principal they face the appropriate level for the analysis of the informativeness of the principals' feedback information is the aggregate data. When agents observe a message of $\hat{\Delta}^i$ they do not know whether the message comes from a truthful or a lying principal. We therefore performed OLS regressions (not reported) of $\Delta$ on $\hat{\Delta}^i$, $\left(\hat{\Delta}^i\right)^2$ and $\left(\hat{\Delta}^i\right)^3$. The results of these regressions provide qualitative insight into the extent to which the principals' messages reveal information about actual output differences.

[6] There are also a few observations where principals "overreport" the absolute value of the actual output difference. However, these observations almost exclusively occur in the first three rounds until principals realize that "underreporting" is more profitable.

In particular, they address the question what prediction an agent would make about $\Delta$ after observing $\hat{\Delta}^i$ if he had all the data of the experiment at his disposal.

In these regressions the coefficient on the reported output difference $\hat{\Delta}^i$ is positive and significant at the 1% level throughout, indicating that there is clear link between actual realization and the content of messages sent. Moreover, the $R^2$ of the regressions shows that almost a quarter (23%) of the variance in the actual output difference is explained by the reported output difference suggesting that communication is at least partially informative for the agents. In particular, if a contestant would have all the information that the experimenter has at his disposal he would estimate that a reported output difference of $\hat{\Delta}^i = y$ would on average correspond to an actual output difference of about $\Delta \approx 2y$. In other words, on average the actual output difference is about twice as large as the reported output difference.

If agents believe that there is a positive relationship between reported and actual output differences, which indeed there is as documented in Result 4, then to effect an increase in their own payoff principals should report an output difference that is generally lower in absolute value than the actual output difference. This aggregate underreporting behavior documented in Result 4 suggests that principals clearly seem to understand that if their message is believed and acted upon accordingly by the agents, reporting a lower output difference will lead to higher second stage effort choices and hence higher payoffs.

The informativeness of feedback also goes some way of explaining the effort response of agents in the PF condition. Having established that the messages sent by principals are at least partially informative in the aggregate data, we now turn to a closer analysis of the reporting behavior of principals at the individual level. In particular, we ask what factors drive the principals' reporting behavior.

**Result 5** (principals' communication behavior): *Reporting behavior is driven by the potential gains from lying. Principals are more likely to report the truth when the potential gains from lying are small. Furthermore, the partial informativeness of messages is due to the heterogeneity among the principals: some principals always lie maximally about the first stage output difference, but a significant share of the principals reports close to the truth if the potential gains from lying are relatively small. This heterogeneity is further supported by the fact that Machiavellian personality traits predict lying behavior.*

Figure 5 provides a first indication that principals do not lie maximally by always reporting an output difference that is equal to 0, but sometimes choose to report larger output differences. To empirically test the prediction that principals are more likely to lie when the potential gains to lying are larger we ran several regressions of the absolute divergence of the report from the truth $|\hat{\Delta}^i - \Delta|$ on the maximum potential gain $G$ from untruthful reporting. Denote the maximum possible theoretical gain $G$ of an untruthful report, if followed naively by an agent,

by

$$G \equiv \theta \left[ e_2(\hat{\Delta}^i = 0) - e_2(\hat{\Delta}^i = \Delta) \right].$$

Table 5 reports the results of OLS regressions which yield highly significant positive coefficient for the gain $G$ regardless of whether we control for round effects and interaction effects between rounds and $G$. On average an increase in $G$ by one unit increases the extent of untruthful reporting by 0.4 units. Similar results hold for probit estimates (not reported) with the probability of reporting untruthfully as the dependent variable.

**Insert Table 5 here.**

We also investigated the principals' reports about agents' *absolute* output in addition to their reporting behavior with regard to the output difference discussed above. In this case, principals almost always report the truth (results not reported), i.e., the agent typically receives truthful information about his own output while the information about the output difference, and, hence, the opponent's output is biased. This provides additional evidence consistent with our conjecture that lying behavior is essentially driven by the potential gains from untruthful reporting because agents' behavior is driven by reported output differences and not by their absolute output levels.

In order to investigate heterogeneity in principals' communication behavior in more detail, we performed the analysis displayed in Figure 5 for each of the 32 principals in our experimental data. These individual level regressions reveal heterogeneity with respect to the principals' willingness to report the truth. Figure 6 shows the results for two principals (subjects #24 and #43). Whereas the first principal (subject #24) consistently reports output differences that are close to zero regardless of the actual output difference, the second principal (subject #43) truthfully reports for small absolute values of the first stage output difference.

The data contained in Figure 5 and, in particular, Figure 6 hint at the differences in reporting behavior across principals. Given that such differences exist, we are interested in analyzing reporting behavior using subject-specific characteristics. Table 6 reports the results of OLS regressions with the reported output difference $\hat{\Delta}^i$ as the dependent variable. If all principals lie maximally all the time, then the coefficient on $\Delta$ should be equal to 0. As can be seen from Table 6, throughout all specifications this coefficient is significantly different from 0. This means that the actual output difference influences reporting behavior. However, aggregate data is not ideally suited to our analysis since there is substantial heterogeneity with respect to reporting behavior. To control for heterogeneity in reporting behavior we use the Machiavelli score. We measured subjects' Machiavellism – a combination of selfishness and opportunism – with the Machiavelli questionnaire (Christie and Geis 1970). In this questionnaire the subjects

indicate their degree of agreement with statements such as "It's hard to get ahead without cutting corners here and there" and "The best way to deal with people is to tell them what they want to hear". Based on their scores on this questionnaire we classified principals in 3 different groups: low Machiavelli (lowest 10%), medium Machiavelli (middle 80%) and high Machiavelli (highest 10%). For subjects with low and high Machiavelli scores we create the binary variables *LOWMACH* and *HIGHMACH*. As can be seen from Table 6 in specifications (3) and (4), the interaction effect $\Delta$**LOWMACH* is positive and statistically significant, showing that principals with lower Machiavelli scores on the questionnaire also lied significantly less during the actual experiment. We can observe the opposite effect for highly Machiavellian individuals since the interaction effect $\Delta$**HIGHMACH* is negative. However, even for these highly Machiavellian individuals the combined coefficient on $\Delta$ and $\Delta$**HIGHMACH* is significantly different from 0 ($p$-value 0.0005) so that we can reject the null-hypothesis that for subjects with strong Machiavellian personality traits the actual output difference does not influence the output difference they report. Specifications (2) and (4) show that our analysis is robust to round effects.[7]

**Insert Table 6 here.**

## 4.5 What Drives the Effort Decline in the PF Condition?

While Result 4 provides a rationale for the agents' second stage effort response to feedback in the PF condition, it does not explain the decline of average first and second stage effort in the PF condition. Result 6 shows that this decline can be attributed to the experience of lying behavior.

**Result 6** (deception undermines incentives): *The decline in average effort over time is strongly affected by the agents' deception experiences. The more often an agent has been deceived in the past the more the agent reduces first and second stage effort.*

As documented previously in Results 4 and 5, principals engage in untruthful "underreporting" of the absolute value of the output difference and hence a sophisticated agent who is or has become aware of the principals' reporting strategy should respond to the messages accordingly by reducing second stage effort more than he would have done if the feedback given to him

[7] Note also that we included the *ROUND* and Machiavelli score variables separately in the regressions reported in Table 6. These non-interaction terms are not significant and therefore omitted in the presentation of Table 6.

was entirely truthful. Result 6 indeed shows that on average agents respond correctly to the manipulated information of principals.

How did the agents find out whether they have been deceived by a principal? When an agent has been told repeatedly in the past rounds that the first stage output difference is very low in absolute value he might come to suspect that principals reported untruthfully. In fact, according to responses of subjects in the post-experiment questionnaire such a very low first stage output difference alerts agents that some principals may report untruthfully. When asked how they detected whether principals engaged in lying behavior agents spontaneously cited low absolute values of reported first stage output differences as the most important measure (64% of agents). This intuition of agents is indeed correct since output differences that are small in absolute value are more likely to come from a dishonest than an honest principal.

We therefore construct a measure that adds up the occurrences of very low reported output differences for each agent over *previous* rounds. This measure of *past* lying behavior is the lagged cumulative sum of the occurrence of low difference reports (*LOW DIFFERENCE*). We classified any report of the first stage output difference between $-3$ and 3 as a low difference thus creating a binary variable.[8] About 38% of all reports over all the rounds of the experiment fall into this range which also includes the modal report (2). The lagged cumulative sum of the number of low reported output differences captures the idea that agents who have more reason to believe that they have been deceived in the past are more likely to discount feedback information in the current round. The number of agents who experience such a low output difference increases over time. Whereas at the start of round 2 only 18 out of 64 contestants have received very low feedback about the output difference, in round 15 there are 61 contestants who have previously received a very low output difference report.

We investigate whether this measure of past lying behavior influences second stage effort choice by running the regressions reported in Table 7. The experience of past lying behavior leads contestants to exert less effort in stage 2 as specifications (2), (3) and (4) reveal. In all these specifications the coefficient on the lagged cumulative sum of low output difference report occurrences is negative and statistically significant and the effort response to information feedback is now even more pronounced than in the models for the PF condition presented in Table 3. On average, for any low reported output difference that a contestant has received from a principal in the past he reduces effort by 0.99 as shown in column (2). Note that the coefficient on the *ROUND* variable which is significant in column (1), is no longer significant whenever we control for past lying experiences. In fact, when we control for past lying behavior the coefficient for the round variable becomes even positive (though insignificant). Thus the decline in second stage effort over time is not simply a time trend but it seems to be driven by

[8] We also conducted robustness checks with different cut-off levels and different measures of deception occurrences. The results were qualitatively similar.

the experience of past lying behavior.

**Insert Table 7 here.**

Next, we analyze whether similar effects can be found for first stage effort. According to the model in Section 2 the information agents receive at the end of the first stage is only relevant for second stage effort. First-stage effort should be the same regardless of how much information is transmitted; therefore it should be the same across all three conditions (NF, TF and PF) and constant over time. As Result 3b showed this is in fact the case for the first stage effort in the NF and TF condition, where average first stage efforts were statistically indistinguishable and constant over time despite the differences in information release before the beginning of the second stage.

As before, we investigate whether the effort decline in the PF condition is a reaction to the increasing awareness of agents that principals are lying. We run OLS regressions of first stage effort on the past lying measure (*LOW DIFFERENCE*) and the number of rounds shown in Table 8. On average, for any low reported output difference that a contestant has received from a principal in the past he reduces effort by 1.07 as shown in column (2). As in our analysis of second stage effort, the coefficient on the *ROUND* variable which is significant in column (1) is no longer significant. Finally, we note that our results are robust to different cut-off levels for the classification of low difference reports. Model (3) in Table 8 also reports regression results for the symmetric cut-off level of $[-1, 1]$.

**Insert Table 8 here.**

We also investigated whether there is a similar effect on first and second stage effort in the TF condition where feedback is truthful. As expected, there is no significant evidence of such an effect regardless of the cut-off we choose for classifying a low difference. Thus the behavior documented in Tables 7 and 8 is due to agents anticipating that principals are lying rather than the consequence of a simple learning effect for second stage effort.

# 5 Interpretation

The previous section shows that when there is either no feedback (NF condition) or completely truthful feedback (TF condition) the theoretical model proposed in Section 2 works remarkably well. The experimental evidence largely confirms the predictions of the model; the agents effort

is decreasing in $|\Delta|$ and the average effort in both stages of the NF condition and the first stage of the TF condition are very similar. In addition, the average effort in the second stage of the TF condition is smaller than effort in the NF condition. Note that this prediction relies on a subtle aspect of the model (the convexity of the marginal cost function) and we find it quite remarkable that the data support it.

In contrast, when information can be manipulated and feedback is given by self-interested principals, the theoretical model is no longer supported by the experimental data. Contrary to the theoretical predictions, communication by principals is at least partially informative and agents strongly respond to the messages sent by principals. Furthermore, agents significantly reduce their average effort in the first and the second stage if they have experienced lies in the past. It seems that the principals' deceitful behavior thoroughly undermines effort provision.

These experimental results naturally lead to us to ask a new set of questions: Can we explain the failure of the cheap talk hypothesis with a simple and parsimonious theoretical model? Are other theories equally successful at explaining the theoretical results about principals' reporting behavior and agents' effort choice?

In this section we interpret our experimental findings in the light of the failure of the cheap talk hypothesis, the evidence gathered from the previously discussed results and a simple and parsimonious model of lying aversion, which explains many of the puzzling features documented previously. Furthermore, we show that social preference theories based on motives of altruism, reciprocity and inequity aversion cannot explain key observations such as the informativeness of the principals' messages. The lying aversion model is developed in greater detail in Appendix B, and Appendix C contains a more detailed analysis of the predictions of social preference theories.

Before we address the question of how the existence of lying averse subjects can account for many observations in the PF condition, it is worthwhile to stress that lying aversion cannot play a role in an environment in which lying is impossible. Thus, lying aversion obviously cannot affect behavior in the NF and TF treatment, implying that the predictions of the model described in Section 2 apply. Therefore, the fact that behavior in the NF and TF condition largely meets the predictions of the model in Section 2 is perfectly consistent with the assumptions of lying aversion.

## 5.1 Principals' Reporting Behavior

We begin our analysis with the reporting behavior of principals. In our baseline theoretical model of Section 2 principals would report whichever output difference maximized an agent's effort for a given belief of the agent; principals should therefore deviate from reporting the truth whenever it is in their interest to do so. Consider the following situation in which the principal

holds fixed beliefs such that an agent will exert higher effort for a reported output difference that is lower in absolute value. Clearly, this is a reasonable assumption given the observed average effort response in the PF condition shown in Figure 2.

Now, consider the case of heterogenous principals who have different preferences for truth-telling because they suffer different disutility when they report untruthfully. When deciding what message $\hat{\Delta}^i$ to send to agent $i$ they have to weigh the benefits against the disutility of lying. When agents exert higher effort for a reported output difference that is lower in absolute value then the benefits of lying come from "underreporting" the output difference. However, since such "underreporting" also involves costs when the principal is averse to lying, she may often choose to keep lying disutility low by not lying maximally, i.e. she will not report an output difference very close or equal to zero, or even avoid lying costs altogether by reporting truthfully. When principals have a preference for reporting truthfully then the actual output difference matters for the principal's reporting behavior. This is one of the key predictions (Theoretical Result 1) of the lying aversion model presented in Appendix B which fleshes out the above argument more formally and in more detail. The rationale for lying as a cost-benefit decision is further supported by the evidence summarized in Result 5 where we show that reporting behavior is driven by the gains from lying. Clearly, the fact that there is some truthful reporting and the reports in general depend on the actual realization of the output difference is what makes communication at least partially informative.

Theoretical Result 1 also shows that each reported first stage output difference should be either equal to or smaller in absolute value than the actual output difference. This effect is due to the effort-increasing effects for "underreporting" described above. Again, the theoretical result is strongly supported in our data where more than 95% of all reports are either truthful (i.e., equal to the actual output difference) or smaller in absolute value than the actual output difference.

While our model of lying aversion can explain the principals' behavior, theories of social preference based on motives of altruism, reciprocity or inequity aversion fail to do so. We show this in more detail in Appendix C and provide some intuition here. When the principal has altruistic motives and cares equally about the two agents, she is indifferent which agent wins the tournament. The principal's private reporting decision to an agent is therefore only affected by second stage effort (multiplied by $\theta$) chosen by the agent and the associated cost of effort. Since second stage effort only depends on the reported output difference, but not the actual output difference, which is not observed by the agent, the optimal report of an altruistic principal is independent of the actual first stage output difference. As a result, we should not observe any systematic relationship between actual and reported output differences whereas in the data we clearly do. The same argument applies to an inequity-averse principal or a principal who responds reciprocally to the first stage output levels by putting more weight on the welfare of

the agent with a higher output.

It is also worthwhile to point out that the concept of quantal response equilibrium does not seem to be able to explain the dependence of the principals' messages on the true output differences. The reason for this is that the quantal response approach assumes that players play noisy best replies, i.e., they tend to make mistakes and these mistakes are more likely if the costs of a mistake are low. In our PF condition, the agents' effort response to a reported output difference is given by the graph displayed in Figure 2. A self-interested principal who plays a noisy best reply to this effort response should always report the output difference that maximizes her payoff with the highest probability but – due to the noise in the best reply – she will also report neighboring output differences with positive probability. Thus, this approach can explain why principals will not always tell the truth but it cannot explain why the principals' feedback information depends on the true output difference.

As documented in Result 5 there is heterogeneity in reporting behavior across principals; some principals lie more and more often than others. In our model of lying aversion this difference in preferences is captured by a lying aversion parameter which determines the strength of lying aversion and varies across principals. Consider two principals with different levels of lying aversion who observe the same first stage output difference and decide how to report to the agents. Reporting untruthfully generates the same benefits for the two principals, however for the same deviation from the truth the more lying-averse principal suffers greater disutility than her less scrupulous counterpart. As a result, more lying-averse principals will lie less in equilibrium. This prediction is also apparent in our experimental data where principals with weak Machiavellian personality traits lie considerably less often than principals with intermediate or strong Machiavellian traits as shown in Result 5 and Table 6.

## 5.2 Agent Effort Response

Turning our focus to the effort choice of agents it seems that the main reason why the baseline model fails in the PF condition is the informative content of principals' messages. In fact, once we recognize that communication is informative key aspects of agents' behavior in the PF condition make sense.

The previous section showed that all principals will either "underreport" or report truthfully, but will never report untruthfully when they do not gain from doing so. As a result, any agent should expect that the actual first stage output difference $\Delta$ relevant for his second stage effort decision is larger in absolute value than the first stage output difference $\hat{\Delta}^i$ reported to him. As we showed in our discussion of Result 4, on average the actual first stage output difference is about twice as large as the reported one and more than 95% of all reports are either truthful (i.e., equal to the actual output difference) or smaller in absolute value than the actual output

difference. Hence, for any given reported output difference a rational agent should exert less effort than in the case where feedback is completely truthful. This is exactly the behavior we derive in Theoretical Result 2 in Appendix B: for a given report agents are predicted to exert less effort in the PF condition than in the TF condition.

While the experimental data strongly support the qualitative predictions of our model, we are also interested in how close actual second stage effort choices are to optimal effort choices. An optimal effort choice requires agents to perform the correct inference from a message sent by the principal as well as making an appropriate effort choice. In our discussion of the TF condition we already showed that agents effort choices are close to optimal when all communication is truthful and there is no complicated inference process. Consider now the case of a rational agent in the PF condition. Since there is heterogeneity in lying behavior among principals, when receiving a report about the first stage output difference $\hat{\Delta}^i$ the agent does not know exactly what the actual output difference $\Delta$ is. However, he knows the distribution of the actual output difference. In other words, the agent knows the posterior distribution of $\Delta$ following a message $\hat{\Delta}^i$ which is given by $\Pr(\Delta \mid \hat{\Delta}^i)$. Consequently, he also knows his conditional probability of winning the tournament which is given by

$$\Pr(e_2^i - e_2^j + \Delta + \Delta\varepsilon_2 > 0 \mid \hat{\Delta}^i).$$

Given this posterior distribution the agent chooses his second stage effort level appropriately.

We can back out the posterior distribution of $\Delta$ given $\hat{\Delta}^i$ from our experimental data using Kernel estimation and then compute the optimal response for a given $\hat{\Delta}^i$.[9] This optimal response function is shown in Figure 7 alongside the actual average second stage effort response for the range of $\hat{\Delta}^i \in [-25, 25]$ into which more than 95% of all messages fall. Figure 7 shows that the actual effort responses in the PF condition are relatively close to the optimal effort responses although the actual responses do not match the optimal response in every detail. In fact, the deviation of actual from optimal behavior is not statistically significant ($t$-test, $p$-value 0.16).

## 5.3 Dynamics

We are also interested in understanding the dynamics of effort choices in the PF condition. As shown by Result 6 contestants who receive a higher number of low output difference reports in previous periods, adjust their second stage effort response downwards. Our model of lying aversion (Theoretical Result 3) explains this behavior by a simple updating process about the

[9] In the specification we report in Figure 7, we assumed that following a message $\hat{\Delta}^i$ player $i$ knows the exact equilibrium effort of his opponent. However, we also performed the same analysis for the case where following a message $\hat{\Delta}^i$ player $i$ only knows the distribution of efforts of his opponent. The results are essentially the same.

relative proportions of high and low lying-averse principals on behalf of the agents. If agents are uncertain about the exact proportions of honest and dishonest principals, they will update their beliefs about the relative proportions of honest (more lying-averse) and dishonest (less lying-averse) principals according to the messages they have received in the past. The interaction with more dishonest principals which entails repeatedly observing an output difference that is particularly low in absolute value, will then lead agents to rationally expect that the actual proportion of dishonest principals is high. In turn, when agents believe that the proportion of lying principals is high they will adjust their effort response downward by treating any message they receive more skeptically. By reporting output differences that are low in absolute value a principal achieves a higher payoff for himself, but also undermines the collective credibility of principals which leads to a changed second stage effort response in later periods.

There is one feature in our data that our model, which departs from the standard model only by introducing rational lying aversion among the principals, cannot readily explain. In Result 3 we show that first stage average effort strongly declines over time. However, rational and self-interested agents should not change their first stage efforts regardless of how much the principals underreport the true output differences (see Section 2). The rational response of a self-interested agent to underreporting is to discount the principals' messages and to choose the second stage effort accordingly. Yet, in fact the principals' underreporting of the truth seems to more thoroughly undermine the agents' willingness to provide effort by reducing first stage performance. It is possible to account for this phenomenon if one extends the notion of lying aversion such that lying aversion not only incorporates the psychic costs of lying but also the resistance of the victim of the lie. The evidence in Sanchez-Pages & Vorsatz (2007) and Brandts & Charness (2003) indicates the existence of subjects that are willing to punish lying per se. These authors show that the same payoff consequences trigger quite different punishment behaviors depending on whether the payoffs have been generated by a lie. Subjects are much more willing to punish an opponent for generating a payoff allocation if the opponents' behavior is based on a lie. In our setting, the anticipation of being the victim of a lie could have induced the agents to decrease also their first stage effort. Moreover, agents who have more reason to believe that the principal will lie, i.e., those agents who experienced a higher number of small output reports in the past, also have more reason to respond with a lower first stage effort.

# 6 Conclusion

In this paper we have shown that deceptive information by the principals eventually undermines the effectiveness of performance incentives. The principals initially benefit from deceiving the agents because the underreporting of output differences induces agents to work harder com-

pared to a truthful output report. However, despite massive underreporting the average effort in the second stage of the tournament eventually falls below the effort in the No Feedback treatment and the Truthful Feedback treatment because the agents strongly discount the principals' feedback information. Moreover, the agents not only reduce their effort in direct response to the principals' feedback but they also show a strong decline in first stage effort levels indicating a pervasively negative effect of deception on tournament incentives. This decline in average effort levels seems to be driven by agents' previous experiences of deception.

In order to assess these results it is important to keep in mind that the agents knew that the winner of the tournament was determined objectively by the total output difference after stage two. Therefore, the principals could not engage in favoritism by falsely proclaiming one player the winner of the tournament, nor did the principals have incentives to do so in our setting. In our view, this renders the fact that the agents average first and second stage effort in the PF condition falls below the effort in the No Feedback and the Truthful Feedback condition all the more remarkable. This finding suggests that for an organization it may be better to provide no feedback to the agents in promotion tournaments unless the principal can commit to provide fully truthful feedback. In view of our results it is interesting that the organizational behavior literature (Meyer, & French (1965), Beer (1987), Gibbs (1991)) documents the widespread absence of performance feedback in firms.

Our data suggest that the detrimental impact of the principals' feedback on average effort is due to the deceptive nature of this feedback. However, the data also suggest that the principals' imperfect lying behavior, i.e., the fact that the principals' messages are informative about the true output differences, also played a role. If the principals' feedback had been completely unreliable the agents would have had an incentive to behave as if they had been given no information. Recall that according to the chosen experimental parameters effort in the NF condition is predicted to be (and actually turned out to be) even higher than in the TF condition, that is, if the agents had completely disregarded the principals' feedback we should not have observed a detrimental effect of feedback information on incentives.

We regard it, however, as an open question whether agents would indeed not respond at all to completely deceptive (i.e., completely uninformative) feedback by the principals. It is hard to believe that extreme forms of intentional lying do not provoke some effect on agents' effort. In fact, the strong decline in first stage effort levels in our setting is consistent with the idea that people dislike being the victim of a lie. In view of our results we thus believe that it is worthwhile to study the psychological forces associated with lying and the consequences of these forces for contracts and incentives. A large class of contracting problems is characterized by lying incentives, and a better understanding of the motivational forces associated with lying or with being the victim of a lie may enhance our understanding of these problems.

# A Omitted Proofs

## A.1 First Stage Effort in the TF Condition

When feedback is truthful at the beginning of the first stage agent $i$ chooses $e_1^i$ to maximize his utility

$$U^i = p + (P-p)G(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j) - c(e_1^i) - E\left[c(\tilde{e}_2^i)\right].$$

Define the noise difference $\delta_t$ as

$$\delta_t \equiv \varepsilon_t^i - \varepsilon_t^j$$

then

$$\Delta = e_1^i - e_1^j + \delta_1.$$

Since the noise differences are independently (normally) distributed we can rewrite the principal's utility in the following way

$$U^i = p + (P-p)E_{\delta_1}\left[F(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j + \delta_1)\right] - c(e_1^i) - E_{\delta_1}\left[c(\tilde{e}_2^i)\right].$$

To obtain the first order conditions we differentiate this expression with respect to $e_1^i$. The first order conditions are given by

$$\begin{aligned} c'(e_1^i) &= (P-p)E_{\delta_1}\left[f(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j + \delta_1)\right] \\ &\quad +(P-p)E_{\delta_1}\left[f\left(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j + \delta_1\right)\frac{d\tilde{e}_2^i}{de_1^i}\right] - E_{\delta_1}\left[c'\left(\tilde{e}_2^i\right)\frac{d\tilde{e}_2^i}{de_1^i}\right] \\ &\quad -(P-p)E_{\delta_1}\left[f(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j + \delta_1)\frac{d\tilde{e}_2^j}{de_1^i}\right]. \end{aligned} \tag{A1}$$

The first line of equation (A1) captures the direct effect of $e_1^i$, the second line represents the effect of $e_1^i$ on $\tilde{e}_2^i$, and the third line is the strategic effect of $e_1^i$ on $\tilde{e}_2^j$. The three lines correspond to the three different effects discussed in the main part of the paper. We now show that the second and the third line are zero.

Denote the second line of (A1) by $K$. We note the symmetry of the second-period equilibrium effort levels, $\tilde{e}_2^i = \tilde{e}_2^j$ and use the law of iterated expectations to rewrite the second line in the following way

$$\begin{aligned} K &= E_\Delta E_{\delta_1}\left[\left[(P-p)f(\Delta) - c'(\tilde{e}_2^i(\Delta))\right]\frac{d\tilde{e}_2^i(\Delta)}{de_1^i} \mid \Delta\right] \\ &= E_\Delta \frac{d\tilde{e}_2^i(\Delta)}{de_1^i} E_{\delta_1}\left[(P-p)f(\Delta) - c'(\tilde{e}_2^i(\Delta)) \mid \Delta\right] \\ &= 0 \end{aligned}$$

since

$$(P-p)f(\Delta) = c'(\tilde{e}_2^i(\Delta))$$

which is the first order condition of the second stage effort choice given in equation (4).

The strategic effect given by the third line of (A1) is zero as well in a symmetric equilibrium. Denote the term in the third line of (A1) by $S$. First, there is a first stage equilibrium that is symmetric in effort levels, i.e.,

$$\tilde{e}_1^i = \tilde{e}_1^j \equiv \tilde{e}_1$$

and therefore $\Delta = \delta_1$. Second, the second stage effort equilibrium is symmetric

$$\tilde{e}_2^i(\Delta) = \tilde{e}_2^j(\Delta) = \tilde{e}_2(\Delta)$$

and $\tilde{e}_2(\Delta)$ is symmetric in $\Delta$ around 0, i.e.,

$$\tilde{e}_2(\Delta) = \tilde{e}_2(-\Delta).$$

Substituting these expressions into the third line we obtain

$$\begin{aligned} S &= (P-p)E_{\delta_1}\left[f(e_1^i + \tilde{e}_2^i - e_1^j - \tilde{e}_2^j + \delta_1)\frac{d\tilde{e}_2^j(\Delta)}{de_1^i}\right] \\ &= (P-p)E_{\delta_1}\left[f(\delta_1)\frac{d\tilde{e}_2(\delta_1)}{d\delta_1}\right] \\ &= (P-p)E_{\delta_1}\left[f(\delta_1)\tilde{e}_2'(\delta_1)\right] \end{aligned}$$

where

$$\tilde{e}_2'(\delta_1) \equiv \frac{d\tilde{e}_2(\delta_1)}{d\delta_1}.$$

From the properties of the density function we know that $f$ is symmetric in $\delta_1$ around 0. Furthermore, from equation (5) characterizing equilibrium second stage stage effort we know that $\tilde{e}_2(\delta_1)$ is symmetric in $\delta_1$ around 0 and that it achieves a maximum at $\delta_1 = 0$, so

$$\begin{aligned} f(\delta_1) &= f(-\delta_1) \\ \tilde{e}_2(\delta_1) &= \tilde{e}_2(-\delta_1) \\ \tilde{e}_2'(\delta_1) &= -\tilde{e}_2'(-\delta_1) \\ \tilde{e}_2'(0) &= 0. \end{aligned}$$

Rewriting the expression for $S$ we have

$$\begin{aligned} S &= (P-p)\int_{-\infty}^{\infty} f(\delta_1)\tilde{e}_2'(\delta_1)f(\delta_1)d\delta_1 \\ &= (P-p)\left[\int_{-\infty}^{0}\left[f(\delta_1)\right]^2\tilde{e}_2'(\delta_1)d\delta_1 + \int_{0}^{\infty}\left[f(\delta_1)\right]^2\tilde{e}_2'(\delta_1)d\delta_1\right] \\ &= (P-p)\left[-\int_{-\infty}^{0}\left[f(-\delta_1)\right]^2\tilde{e}_2'(-\delta_1)d\delta_1 + \int_{0}^{\infty}\left[f(\delta_1)\right]^2\tilde{e}_2'(\delta_1)d\delta_1\right] \\ &= (P-p)\left[-\int_{0}^{\infty}\left[f(\delta_1)\right]^2\tilde{e}_2'(\delta_1)d\delta_1 + \int_{0}^{\infty}\left[f(\delta_1)\right]^2\tilde{e}_2'(\delta_1)d\delta_1\right] \\ &= 0. \end{aligned}$$

The first order condition therefore simplifies to

$$c'(\tilde{e}_1) = (P-p)g(0).$$

## A.2 Expected Second Stage Effort in the TF Condition

In equilibrium, the expected effort in the second stage in the TF condition $E[\tilde{e}_2]$ where feedback is truthful, is lower (higher) than effort in the NF condition $e^*$ where no feedback is given, if the marginal cost function $c'$ is convex (concave).

Since the noise differences are independently (normally) distributed equation (2) can be rewritten in the following way

$$\begin{aligned} c'(e^*) &= (P-p)g(0) \\ &= (P-p)E_{\delta_1}[f(\delta_1)] \\ &= E_{\delta_1}[(P-p)f(\Delta)] \\ &= E_{\delta_1}[c'(\tilde{e}_2(\Delta))] \end{aligned}$$

where we used the fact that $\tilde{e}_1^i = \tilde{e}_1^j \equiv \tilde{e}_1$ (and therefore $\Delta = \delta_1$) as well as equation (4). Using Jensen's inequality we can provide a ranking of $\tilde{e}_2(\Delta)$ which is a random variable and $e^*$ which is a constant. If $c'$ is convex, then Jensen's inequality implies

$$c'(e^*) = E_{\delta_1}[c'(\tilde{e}_2(\Delta))] \geq c'(E_{\delta_1}[\tilde{e}_2(\Delta)])$$

and hence $e^* \geq E[\tilde{e}_2]$ since $c'$ is increasing. Clearly, if $c'$ is concave the reverse inequalities hold. If the cost function is quadratic, then marginal cost is linear and $e^* = E[\tilde{e}_2]$.

# B A Model of Communication with Lying Aversion

In this appendix we present a model that provides the theoretical underpinnings for our discussion and interpretation of results in Section 5.

There is evidence from psychological and economic research that individuals are averse to lying even when it is in their interest to do so (Mazar & Ariely (2006), Gneezy (2005), Sanchez-Pages & Vorsatz (2007)).

Consider the following change in the principal's payoff function

$$U^P = \theta\left(x_1^A + x_1^B + x_2^A + x_2^B\right) - P - p - l\sum_{i=A,B} h\left(|\hat{\Delta}^i - \Delta|\right) \tag{B1}$$

where $h\left(|\hat{\Delta}^i - \Delta|\right)$ is a lying term that is increasing ($h' \geq 0$) and convex ($h'' \geq 0$) in the absolute value of the difference between the reported output difference $\hat{\Delta}^i$ and the actual output difference $\Delta$. The principal suffers disutility when she reports an output difference that differs from the truth and this disutility is increasing in the difference between the reported value and the truth. Note that $l$ is a positive constant that may vary across principals. The higher is $l$ the stronger is the principal's aversion to lying.

## B.1 Communication Equilibrium

We now characterize a communication equilibrium with naive senders. When considering what message $\hat{\Delta}^i$ to send to agent $i$ the principal has to weigh the benefits against the disutility of lying. Naive senders believe that agents do not anticipate that the sender will not report truthfully. In other words, they believe that the agents will not adjust their effort behavior in response to the principals' lies.

For a given $\Delta$ taking first order conditions for a report to agent $i$ we find that principal's utility varies with $\hat{\Delta}^i$ in the following way

$$\frac{\partial U^P}{\partial \hat{\Delta}^i} = \theta\frac{\partial \hat{e}_2^i}{\partial \hat{\Delta}^i} - lh'\left(|\hat{\Delta}^i - \Delta|\right). \tag{B2}$$

The first term on the left-hand side of the equation is the marginal benefit of lying whereas the second term gives the marginal cost of lying.

In contrast to our model without lying disutility the optimal message now depends on $\Delta$. Furthermore, note that $\hat{e}_2^i(\Delta)$ is a function that has the shape of a normal density. For values of $\Delta$ close to zero second stage effort $\hat{e}_2^i$ is not very responsive to changes in $\Delta$, that is, the marginal benefit of lying is very small in that range. Hence, if the marginal cost of lying is strictly increasing ($h' > 0$), for $l$ large enough and $\Delta$ close enough to zero the principal's optimal report $\hat{\Delta}^{i*}$ is a corner solution, that is the principal reports the truth. For a principal with lying aversion $l$ denote this cutoff for $|\Delta|$ below which truthful reporting occurs by $\underline{\Delta}(l)$, in other words, if $|\Delta| < \underline{\Delta}(l)$ then

$$\hat{\Delta}^{i*}(\Delta) = \Delta.$$

Note that this cutoff $\underline{\Delta}(l)$ is increasing in $l$. Clearly, as the principal's lying aversion goes to zero, this cutoff will go to zero, i.e., $\underline{\Delta}(0) = 0$, so a principal who is not lying-averse will lie maximally (report 0) all of the time.

The marginal benefit of lying increases as $|\Delta|$ increases and may eventually overwhelm the disutility of lying. This is because as $|\Delta|$ increases the slope of $\hat{e}_2^i(\Delta)$ may become steeper than the slope of the lying term. Lying, however, is not maximal since the optimal message $\hat{\Delta}^{i*}$ is not equal to 0, but is given by the report that balances the marginal benefit and marginal cost of lying,

$$\theta \frac{\partial \hat{e}_2^i}{\partial \hat{\Delta}^i} = lh'\left(|\hat{\Delta}^i - \Delta|\right). \tag{B3}$$

One should also note that since the second stage effort has the same shape as a normal density, which is flat around the mean of 0 and also flat in the tails, the marginal benefit to lying decreases again as $|\Delta|$ increases above 40. Thus, at some point given by $\overline{\Delta}(l)$ it will no longer be profitable at all for the principal to lie. This feature of our model is unobservable in our data since even very honest principals (with high $l$) would report truthfully only for very large values of $|\Delta|$. For example, according to our model a very honest principal (such as the principal shown in right panel of Figure 6) who reports truthfully for $|\Delta| < 20$ (so $\underline{\Delta}(l) = 20$) should also report truthfully for $|\Delta| > 64$ (so $\overline{\Delta}(l) = 64$). In our data, this principal is never put in such a situation. In fact, when calculating the principal-specific cutoffs $\underline{\Delta}(l)$ from our data and the implied cutoffs $\overline{\Delta}(l)$ we find that extreme realizations of $|\Delta| > \overline{\Delta}(l)$ constitute less than 0.5% of our data. We therefore ignore this slightly peculiar feature of our model.

In summary, this gives our first theoretical result.

**Theoretical Result 1**: *If the principals believe that agents will naively follow their report, for small values of* $|\Delta| < \underline{\Delta}(l)$ *the principal's optimal reporting strategy* $\hat{\Delta}^{i*}(\Delta)$ *for a given* $\Delta$ *is to report truthfully, i.e.*

$$\hat{\Delta}^{i*}(\Delta) = \Delta \text{ if } |\Delta| < \underline{\Delta}(l)$$

*and it is characterized by the first order condition in (B3) for intermediate values,* $\underline{\Delta}(l) < |\Delta| < \overline{\Delta}(l)$. *Furthermore, the reported output difference is always lower than the true output difference,*

$$|\hat{\Delta}^i| \leq |\Delta|.$$

*Principals with lower* $l$ *lie more. If lying aversion is sufficiently large, principals always report the truth* $\hat{\Delta}^{i*}(\Delta) = \Delta$ *and principals always lie maximally* $\hat{\Delta}^{i*}(\Delta) = 0$ *if lying aversion is sufficiently small.*

## B.2 Effort Choice

Consider now the agents' effort choice. As mentioned before, there is heterogeneity with respect to lying aversion among the principals and agents do not know with which principal they are matched. The agents therefore do not know each principal's aversion to lying given by the parameter $l$ and the resulting cutoff $\underline{\Delta}(l)$ above which lying occurs. A report $\hat{\Delta}^i$ may be truthful if it is coming from a truthful principal with high $l$ or it may be a lie sent by a principal with low $l$. However, from Theoretical Result 1 we know that

$$|\hat{\Delta}^i| \leq |\Delta|$$

since principals would never want to tell a lie if it decreases their pecuniary payoff. Realizing that principals have an incentive to underreport agents should adjust their response accordingly. This gives us our second theoretical result.

**Theoretical Result 2**: *For a given reported first stage output difference $\hat{\Delta}^i$ second stage effort in the PF condition is lower than in the TF condition, that is we have*

$$\hat{e}_2^i\left(\hat{\Delta}^i\right) \leq \tilde{e}_2^i\left(\hat{\Delta}^i\right) \qquad \forall \hat{\Delta}^i$$

*where $\hat{e}_2^i\left(\hat{\Delta}^i\right)$ and $\tilde{e}_2^i\left(\hat{\Delta}^i\right)$ represents the second stage efforts in the PF condition and in the TF condition, respectively.*

## B.3 Dynamics

We now turn to a theoretical analysis of the dynamics observed in the data. Consider the following simplification. Assume that the principals are independently drawn from a population of two types of senders. With probability $\mu$ the sender has low lying aversion $\underline{l} > 0$, otherwise (probability $1 - \mu$) the sender has high lying aversion with $\bar{l} > \underline{l}$. Contestants are initially uncertain about the exact proportion $\mu$ of sender types, but hold a correct prior $m$ where

$$E[\mu] = m.$$

The agents update their beliefs about $\mu$ to form a posterior $\hat{\mu}^i$ according to information obtained while playing in the experiment. As shown before, for a given output difference $\Delta$ a principal with low lying aversion will lie more than a principal with high lying aversion. Thus, a reported output difference that is particularly low in absolute value, i.e. $|\hat{\Delta}^i| < k$ where $k$ is positive but small, increases the probability that it was sent by a principal with low lying aversion.

More importantly though, in this setting where the exact probability $\mu$ is unknown (in contrast to where it is known with certainty) receiving a low output difference message does

not only provide information about the type of the sender but also provides information about the actual proportions of low- and high-lying-averse principals. *Ceteris paribus*, a contestant who has received a low output difference message $|\hat{\Delta}^i| < k$ in previous rounds will now hold a higher posterior $\hat{\mu}^i$ than a contestant who has not received such a message. Furthermore, a contestant with a higher posterior $\hat{\mu}^i$ anticipates there to be more lying (due to a higher proportion of lying principals) and will therefore adapt his effort response accordingly. In particular, we have

$$\hat{e}_2^i\left(\hat{\Delta}^i;\hat{\mu}^i\right) \leq \hat{e}_2^i\left(\hat{\Delta}^i;\hat{\mu}^{i\prime}\right) \qquad \forall \hat{\Delta}^i \text{ and } \hat{\mu}^i \geq \hat{\mu}^{i\prime}.$$

This yields our third theoretical result which captures the dynamic effort response effects documented in Result 6.

**Theoretical Result 3**: *For a given reported output difference contestants who in previous rounds have received low output difference reports such that $|\hat{\Delta}^i| < k$ where $k > 0$ but small should exert lower second stage effort levels than contestants who have not received such low output differences.*

# C Social Preference Theories

In Appendix B we outlined a simple model of lying aversion that is able to explain the failure of the cheap talk hypothesis. In this appendix we consider other theories that depart from selfish maximization of monetary gains. We show, in particular, that social preference theories based on motives of altruism, reciprocity and inequity aversion cannot explain key features of our experimental results such as the dependence of $\hat{\Delta}^i$ on $\Delta$.

## C.1 Altruism

Consider the case of an altruistic principal with the following objective function

$$U^P = \theta\left(x_1^A + x_1^B + x_2^A + x_2^B\right) - P - p + \alpha \sum_{i=A,B} U^i$$

where $\alpha$ reflects how much the principal cares about the payoffs of the agents. We can rewrite the objective function in the following way

$$U^P = \theta\left(x_1^A + x_1^B + x_2^A + x_2^B\right) - (1-\alpha)(P+p) - \alpha \sum_{t=1}^{2} \sum_{i=A,B} c(e_t^i).$$

In contrast to the first stage effort choices $e_1^i$ which do not depend on the principal's messages $\hat{\Delta}^i$, the second stage effort $\hat{e}_2^i(\hat{\Delta}^i)$ is influenced by the principal's message. However, since

the agent never observes the actual output difference $\Delta$, the second stage effort $\hat{e}_2^i(\hat{\Delta}^i)$ does not depend on $\Delta$. The interior solution for the message to agent $i$ is given by the first order condition

$$\frac{\partial U^P}{\partial \hat{\Delta}^i} = \frac{\partial \hat{e}_2^i}{\partial \hat{\Delta}^i}\left[\theta - \alpha c'(e_2^i)\right] = 0 \text{ for } i = A, B. \tag{C1}$$

Note, that this expression does not depend on $\Delta$ and hence for every value of $\Delta$ the principal should report the same value of $\hat{\Delta}^i$. This is clearly refuted by the data since for almost all principals there is a significantly positive relationship between $\Delta$ and $\hat{\Delta}^i$.

## C.2 Reciprocal Altruism

Consider now the case of a reciprocally altruistic principal. Since the interim winner of a tournament has produced more output for the principal the principal might care more about the interim winner's payoff than the interim loser's. When agent A is ahead, that is $\Delta \geq 0$, then the principal attaches a weight $\alpha$ to agent A's and a weight $\beta$ to agent B's payoff and vice versa when $\Delta < 0$, where $\alpha \geq \beta$. This captures the effect of reciprocity.

Since the setting is symmetric without loss of generality, let $\Delta \geq 0$ so that the principal feels more altruistic to agent A than to agent B. The principal's objective function is therefore given by

$$U^P = \theta\left(x_1^A + x_1^B + x_2^A + x_2^B\right) - P - p + \alpha U^A + \beta U^B.$$

The first order conditions for $\hat{\Delta}^i$ are

$$\begin{aligned}
\frac{\partial U^P}{\partial \hat{\Delta}^A} &= \frac{\partial \hat{e}_2^A}{\partial \hat{\Delta}^A}\left[\theta + (\alpha - \beta)(P - p)f(\hat{e}_2^A - \hat{e}_2^B + \Delta) - \alpha c'(e_2^A)\right] = 0 \\
\frac{\partial U^P}{\partial \hat{\Delta}^B} &= \frac{\partial \hat{e}_2^B}{\partial \hat{\Delta}^B}\left[\theta - (\alpha - \beta)(P - p)f(\hat{e}_2^A - \hat{e}_2^B + \Delta) - \beta c'(e_2^B)\right] = 0.
\end{aligned}$$

Multiplying the first condition by $\frac{\partial \hat{e}_2^B}{\partial \hat{\Delta}^B}$ and the second condition by $\frac{\partial \hat{e}_2^A}{\partial \hat{\Delta}^A}$ and adding these two equations we obtain after simplification

$$2\theta - \alpha c'(\hat{e}_2^A) - \beta c'(\hat{e}_2^B) = 0 \tag{C2}$$

which has to hold for interior solutions of $\hat{\Delta}^i$ regardless of the value of $\Delta$. As in our previous analysis this expression does not depend on $\Delta$ and hence for every value of $\Delta$ the principal should report the same values of $\hat{\Delta}^i$ which is refuted by the data.

## C.3 Inequity Aversion

Consider the case of an inequity-averse principal with preferences given by

$$U^P = \Pi^P - \frac{\alpha}{2}\sum_{i=A,B}\max\left\{U^i - \Pi^P, 0\right\} - \frac{\beta}{2}\sum_{i=A,B}\max\left\{\Pi^P - U^i, 0\right\}$$

where

$$\Pi^P = \theta\left(x_1^A + x_1^B + x_2^A + x_2^B\right) - P - p$$

and $\alpha \geq \beta > 0$.

We analyze three cases here where the principal's profit is either greater or smaller than the payoff of both agents, or between the two. As we will see, inequity aversion cannot explain the reporting behavior of principals in any of the three cases.

First, if $\Pi^P \geq \max\{U^A, U^B\}$ then the principal's utility is given by

$$U^P = \Pi^P - \frac{\beta}{2}\sum_{i=A,B}\max\left\{\Pi^P - U^i, 0\right\}$$

and hence the same conclusions hold as for the case of simple altruism where the state of nature $\Delta$ does not influence the principal's choice of message.

Second, if $\Pi^P \leq \min\{U^A, U^B\}$ then preferences are given by

$$U^P = \Pi^P - \frac{\alpha}{2}\sum_{i=A,B}\max\left\{U^i - \Pi^P, 0\right\}.$$

In this case the principal behaves spitefully but otherwise the same argument as in the first case applies, i.e., the message should not depend on $\Delta$.

Third, consider the case where $\Pi^P$ lies between the two agents, e.g. $U^A \leq U^P \leq U^B$. In this case, the principal's first order conditions are given by

$$\begin{aligned}
\frac{\partial U^P}{\partial \hat{\Delta}^A} &= \frac{\partial \hat{e}_2^A}{\partial \hat{\Delta}^A}\left[\left(1 + \frac{\alpha-\beta}{2}\right)\theta + \frac{\alpha+\beta}{2}(P-p)f(\hat{e}_2^A - \hat{e}_2^B + \Delta) - \frac{\beta}{2}c'(\hat{e}_2^A)\right] = 0 \\
\frac{\partial U^P}{\partial \hat{\Delta}^B} &= \frac{\partial \hat{e}_2^B}{\partial \hat{\Delta}^B}\left[\left(1 + \frac{\alpha-\beta}{2}\right)\theta - \frac{\alpha+\beta}{2}(P-p)f(\hat{e}_2^A - \hat{e}_2^B + \Delta) + \frac{\alpha}{2}c'(\hat{e}_2^B)\right] = 0
\end{aligned}$$

which we can add and simplify to obtain

$$2\left(1 + \frac{\alpha-\beta}{2}\right)\theta + \frac{\alpha}{2}c'(\hat{e}_2^A) - \frac{\beta}{2}c'(\hat{e}_2^B) = 0. \tag{C3}$$

As in the previous cases this condition does not depend on $\Delta$ and hence the optimal message should not depend on the first stage output difference $\Delta$.

# References

**Akerlof, George A.** 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics*, 84(3): 488-500

**Allingham, Michael, and Angmar Sandmo.** 1972. "Income Tax Evasion: A Theoretical Analysis." *Journal of Public Economics*, 1(3-4): 323-338.

**Aoyagi, Masaki.** 2004. "Information Feedback in a Dynamic Tournament." Discussion Paper No. 580, The Institute of Social and Economic Research, Osaka University.

**Beer, Michael.** (1987), "Performance Appraisals." In *Handbook of Organizational Behavior*, ed. Jay Lorsch, 286–301. Englewood Cliffs, NJ: Prentice Hall.

**Blume, Andreas, Douglas V. DeJong, Yong-Gwan Kim, and Geoffrey B. Sprinkle.** 1998. "Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games." *American Economic Review*, 88(5): 1323–1340.

**Bull, Clive, Andrew Schotter, and Keith Weigelt.** 1987. "Tournaments and Piece Rates: An Experimental Study." *Journal of Political Economy*, 95(1): 1–33.

**Brandts, Jordi, and Gary Charness.** 2003. "Truth or Consequences: An Experiment." *Management Science*, 49(1): 116–130.

**Cai, Hongbin, and Joseph Tao-Yi Wang.** 2006. "Overcommunication in Strategic Information Transmission." *Games and Economic Behavior*, 56(1): 7–36.

**Crawford, Vincent P.** 2003. "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review*, 93(1): 133–149.

**Crawford, Vincent P., and Joel Sobel.** 1982. "Strategic Information Transmission." *Econometrica*, 50(6): 1431–51.

**Davis, Ann.** 2004. "Client Comes First? On Wall Street, It Isn't Always So." *The Wall Street Journal*, December 16, A1. http://online.wsj.com/article/SB110315096744601376.html.

**Dickhaut, John W., Kevin A. McCabe, and Arijit Mukherji.** 1995. "An Experimental Study of Strategic Information Transmission." *Economic Theory*, 6(3): 389–403.

**Ederer, Florian.** 2005. "Feedback and Motivation in Dynamic Tournaments." Working Paper, MIT.

**Fischbacher, Urs.** 2007. "z-Tree - Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10(2): 171–178.

**Gibbs, Michael. 1991.** "An Economic Approach to Process in Pay and Performance Appraisals." Working Paper, University of Chicago, Graduate School of Business.

**Gneezy, Uri. 2005.** "Deception: The Role of Consequences." *American Economic Review*, 95(1): 384–394.

**Harbring, Christine, Bernd Irlenbusch, Matthias Kräkel, and Reinhard Selten.** Forthcoming. "Sabotage in Asymmetric Contests – An Experimental Analysis." *International Journal of the Economics of Business.*

**Laffont, Jean-Jacques, and Jean Tirole.** 1993. *A Theory of Incentives in Procurement and Regulation.* Cambridge: MIT Press.

**Lazear, Edward P., and Sherwin Rosen.** 1981. "Rank-order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, 89(5): 841–864.

**Lizzeri, Alessandro, Margaret A. Meyer, and Nicola Persico.** 1999. "Interim Evaluations in Dynamic Tournaments: the Effects of Midterm Exams." Working Paper, University of Pennsylvania.

**Longnecker, Clive O., Henry P. Sims, and Dennis A. Gioia.** 1987. "Behind the Mask: The Politics of Performance Appraisal." *The Academy of Management Executive*, 1: 183-193.

**Mazar, Nina, and Dan Ariely.** 2006. "Dishonesty in Everyday Life and its Policy Implications." *Journal of Public Policy and Marketing*, 25(1): 117–126.

**Murphy, Kevin R., and Jeanette N. Cleveland.** 1995. *Performance Appraisal: An Organizational Perspective*, Boston: Allyn and Bacon.

**Sanchez-Pages, Santiago, and Marc Vorsatz.** 2007. "An Experimental Study of Truth-Telling in a Sender-Receiver Game." *Games and Economic Behavior*, 61(1): 86–112.

**Schotter, Andrew, and Keith Weigelt.** 1992. "Asymmetric Tournaments, Equal Opportunity Laws, and Affirmative Action: Some Experimental Results," *Quarterly Journal of Economics*, 107(2): 511–539.

**Wang, Joseph Tao-Yi , Spezio Michael, and Colin Camerer.** 2007. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-telling and Deception in Games." CalTech Working Paper.

# Table 1
# Parameterization and Predictions

Table 1 outlines the parameterization chosen for the laboratory experiments and presents the theoretical predictions derived from the theoretical model using the chosen parameter values. The parameters $\theta$, $c(e)$, $P$, $p$ and $\sigma$ denote the marginal benefit of agents' effort for the principal, the agent's cost of effort, the prizes for winners and losers of the tournament and the standard deviation of the noise. Starred effort levels are predictions for the NF condition, whereas effort levels with a tilde are predictions for the TF condition. We also give predictions for the cost of effort, the expected utility of an agent and the utility of the principal in the TF condition.

| **Parameterization** | | **Predictions** | |
|---|---|---|---|
| $\theta$ | 4 | $e^* = \tilde{e}_1$ | 37.00 |
| $c(e)$ | $(r/3)e^3$ | $\tilde{e}_2(\Delta)$ | $37\sqrt[4]{2}\exp\left[-\frac{(\Delta)^2}{1600}\right]$ |
| $r$ | 1/686 | $E[\tilde{e}_2]$ | 35.93 |
| $P$ | 300 | $c(e^*)$ | 24.60 |
| $p$ | 100 | $E[U^i(e^*)]$ | 150.80 |
| $\sigma$ | $20\sqrt{2}$ | $U^P(e^*)$ | 192.00 |

Note: We chose parameters in such a way that the equilibrium effort choices did not constitute an obvious focal point (e.g. 50) and were below the socially efficient level (52.83).

# Table 2
# OLS Regressions Examining the Feedback Hypothesis

Table 2 presents OLS regressions for second stage effort choices in the TF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\tilde{e}_2$ | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| *CONSTANT* | *** 48.82 (1.97) | *** 48.82 (1.97) | *** 49.64 (1.49) | *** 48.82 (1.97) |
| $\lvert\Delta\rvert$ | *** -0.42 (0.05) | | | |
| $\Delta > 0$ | | *** -0.30 (0.05) | *** -0.30 (0.05) | *** -0.30 (0.05) |
| $\Delta < 0$ | | *** 0.53 (0.05) | *** 0.53 (0.05) | *** 0.53 (0.05) |
| *ROUND* | | | -0.10 (0.15) | -0.10 (0.15) |
| $\Delta > 0$ * *ROUND* | | | | -0.002 (0.01) |
| $\Delta < 0$ * *ROUND* | | | | 0.007 (0.01) |
| Clusters | 8 | 8 | 8 | 8 |
| *N* | 1440 (full) | 1440 (full) | 1440 (full) | 1440 (full) |
| $R^2$ | 19.11% | 23.26% | 23.32% | 23.39% |

## Table 3
## OLS Regressions Effort PF Condition Stage 2

Table 3 presents OLS regressions for second stage effort choices in the PF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\widehat{e}_2$ | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| *CONSTANT* | *** 43.06 (1.27) | *** 44.30 (1.64) | *** 42.07 (1.48) | *** 43.52 (1.75) |
| $\lvert\hat{\Delta}^i\rvert$ | *** -0.45 (0.09) | ** -0.67 (0.22) | | |
| $\hat{\Delta}^i > 0$ | | | -0.03 (0.22) | -0.08 (0.24) |
| $\hat{\Delta}^i < 0$ | | | *** 0.53 (0.14) | *** 0.97 (0.23) |
| clusters | 8 | 8 | 8 | 8 |
| *n* | 960 (full) | 936 (wins.) | 960 (full) | 936 (wins.) |
| $R^2$ | 5.52% | 3.44% | 7.39% | 7.15% |

## Table 4
## OLS Regressions Effort TF Condition Stage 2 and PF Condition Stage 2

Table 4 presents OLS regressions for second stage effort choices in the TF and PF conditions. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\widehat{e}_2, \widetilde{e}_2$ | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| *CONSTANT* | *** 48.82 (1.90) | *** 48.82 (1.90) | *** 48.82 (1.90) | *** 48.82 (1.90) |
| $\lvert\hat{\Delta}^i\rvert$ | *** 0.41 (0.05) | *** 0.41 (0.05) | | |
| $\hat{\Delta}^i > 0$ | | | *** -0.30 (0.05) | *** -0.30 (0.05) |
| $\hat{\Delta}^i < 0$ | | | *** 0.53 (0.05) | *** 0.53 (0.05) |
| $\lvert\hat{\Delta}^i\rvert$ * *PRINCIPAL* | -0.04 (0.10) | -0.25 (0.22) | | |
| $\hat{\Delta}^i > 0$ * *PRINCIPAL* | | | 0.28 (0.22) | 0.22 (0.23) |
| $\hat{\Delta}^i < 0$ * *PRINCIPAL* | | | 0.001 (0.14) | * 0.44 (0.23) |
| *PRINCIPAL* | ** -5.76 (2.27) | * -4.52 (2.47) | ** -6.75 (2.38) | ** -5.30 (2.55) |
| clusters | 16 | 16 | 16 | 16 |
| *n* | 2400 (full) | 2376 (wins.) | 2400 (full) | 2376 (wins.) |
| $R^2$ | 14.02% | 13.61% | 17.26% | 17.57% |

## Table 5
## OLS Regressions of Lying on the Potential Gains from Lying

Table 5 presents OLS regressions for the absolute value of deviations of reports from the truth in the PF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\lvert\hat{\Delta}^i - \Delta\rvert$ | **(1)** | **(2)** | **(3)** |
|---|---|---|---|
| *CONSTANT* | *** 5.97 (0.84) | *** 4.17 (0.56) | ** 3.97 (1.31) |
| *GAIN* | *** 0.43 (0.01) | *** 0.43 (0.01) | *** 0.44 (0.02) |
| *ROUND* | | * 0.23 (0.10) | 0.25 (0.21) |
| *GAIN * ROUND* | | | -0.00 (0.00) |
| clusters | 8 | 8 | 8 |
| *n* | 960 (full) | 960 (full) | 960 (full) |
| $R^2$ | 81.18% | 81.37% | 81.38% |

## Table 6
## OLS Regressions of Reported on Actual Output Difference in the PF Condition

Table 6 presents OLS regressions for reported output differences in the PF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\hat{\Delta}^i$ | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| *CONSTANT* | -1.06 (0.87) | -1.06 (0.87) | -1.06 (0.87) | -1.06 (0.87) |
| $\Delta$ | *** 0.12 (0.02) | *** 0.16 (0.03) | *** 0.12 (0.02) | *** 0.16 (0.03) |
| $\Delta$ * *LOWMACH* | | | *** 0.10 (0.02) | *** 0.11 (0.02) |
| $\Delta$ * *HIGHMACH* | | | -0.04 (0.02) | * -0.04 (0.02) |
| $\Delta$ * *ROUND* | | -0.004 (0.003) | | -0.004 (0.003) |
| clusters | 8 | 8 | 8 | 8 |
| *n* | 960 (full) | 960 (full) | 960 (full) | 960 (full) |
| $R^2$ | 14.61% | 14.93% | 15.14% | 15.48% |

Note: Round and Machiavelli score were also included as non-interaction terms, but the coefficients are never significant and therefore omitted from Table 5

## Table 7
## OLS Regressions Effort Response PF Condition Stage 2

Table 7 presents OLS regressions for second stage effort choices in the PF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\hat{e}_2$ | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| *CONSTANT* | ***47.34 (2.55) | *** 48.06 (3.19) | *** 50.00 (3.27) | *** 49.29 (3.09) |
| $\lvert\hat{\Delta}^i\rvert$ | *** 0.45 (0.09) | *** 0.47 (0.10) | *** 0.82 (0.16) | |
| $\hat{\Delta}^i > 0$ | | | | -0.18 (0.15) |
| $\hat{\Delta}^i < 0$ | | | | *** 1.12 (0.16) |
| *LOW DIFFERENCE* | | ** -0.99 (0.38) | ** -1.26 (0.40) | ** -1.18 (0.36) |
| *ROUND* | ** -0.56 (0.23) | 0.25 (0.34) | 0.16 (0.35) | 0.21 (0.31) |
| clusters | 8 | 8 | 8 | 8 |
| *n* | 960 (full) | 896 (full) | 872 (wins.) | 872 (wins.) |
| $R^2$ | 7.21% | 8.63% | 7.40% | 11.32% |

Note: There are 896 (872) observations in the full (winsorized) sample when the lagged variable *LOW DIFFERENCE* is included. Additional results for the full sample are very similar, but are omitted from Table 8 for reasons of conciseness.

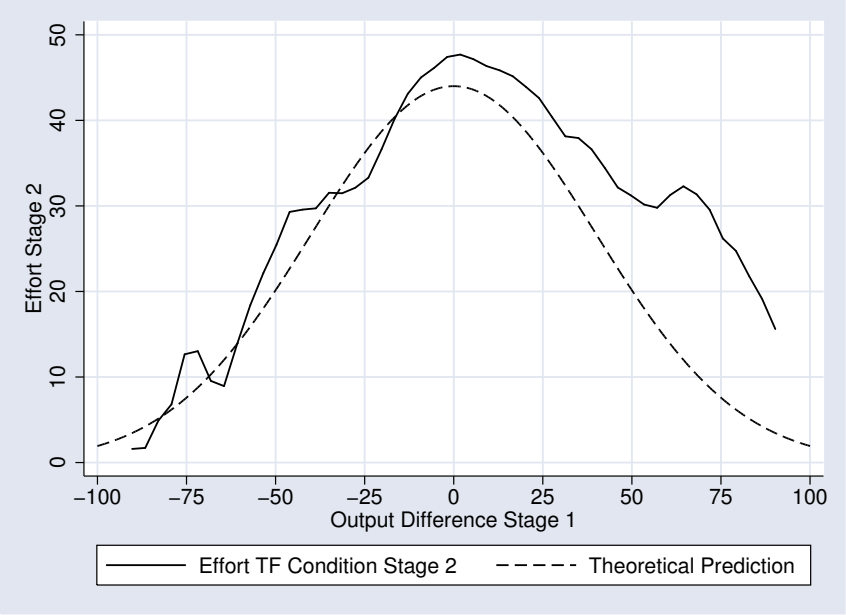
## Table 8
## OLS Regressions Effort Response PF Condition Stage 1

Table 8 presents OLS regressions for first stage effort choices in the PF condition. Clustered standard errors are reported in brackets. Statistical significance at the ten, five and one percent level is indicated by *, ** and ***.

| $\hat{e}_1$ | **(1)** | **(2)** | **(3)** |
|---|---|---|---|
| *CONSTANT* | *** 39.13 (1.36) | *** 39.63 (1.68) | *** 39.53 (1.48) |
| *LOW DIFFERENCE* | | ** -1.07 (0.37) | ** -2.34 (0.99) |
| *ROUND* | * -0.47 (0.21) | -0.15 (0.27) | -0.25 (0.23) |
| clusters | 8 | 8 | 8 |
| *n* | 960 (full) | 896 (full) | 896 (full) |
| $R^2$ | 1.65% | 3.45% | 4.36% |

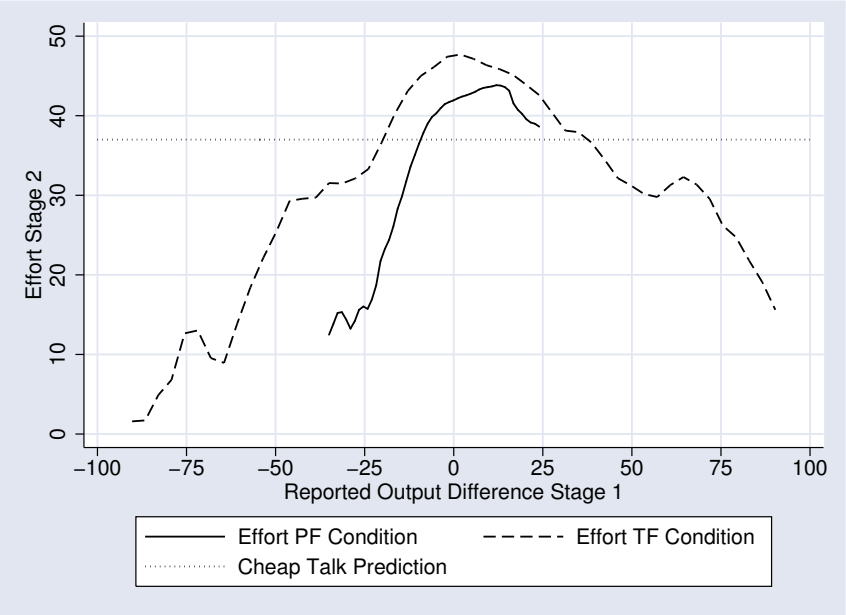
# Figure 1
# Predicted Effort Levels and Kernel Regression TF Condition Stage 2

Figure 1 presents predicted and actual average effort levels for Stage 2 in the TF condition using Epanechnikov Kernel regression.



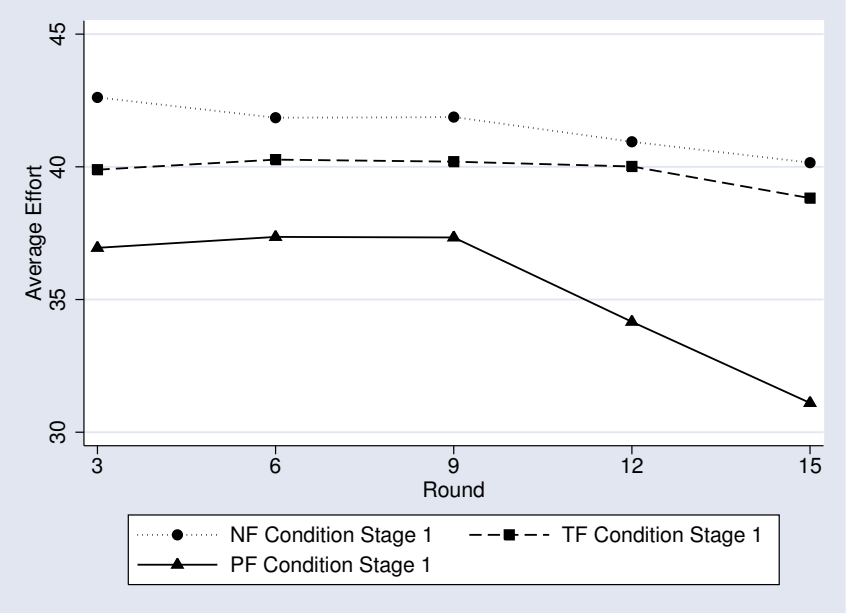

# Figure 2
# Predicted and Actual Effort Levels TF & PF Condition Stage 2

Figure 2 presents predicted and actual average effort levels for Stage 2 in the TF and PF conditions using Epanechnikov Kernel regressions.



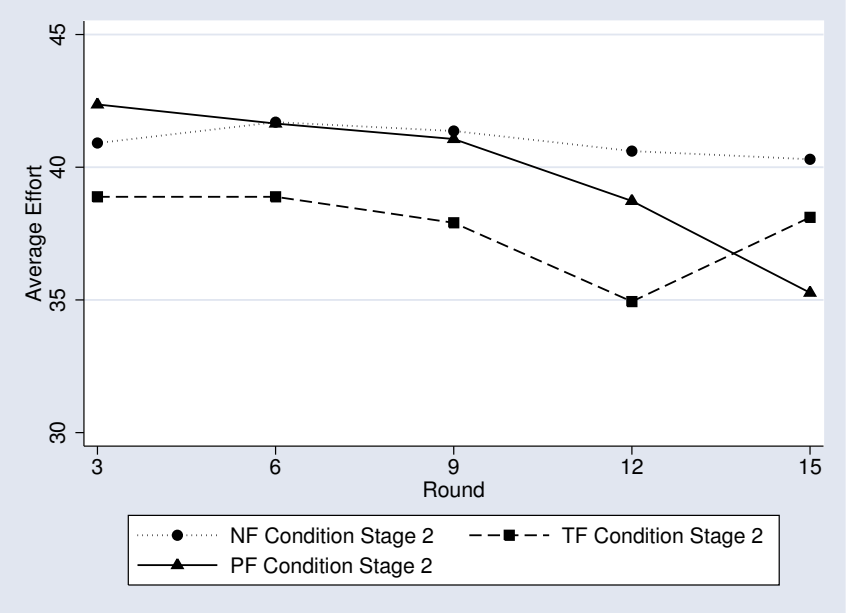
## Figure 3
## Average Effort Levels in NF, TF and PF Condition Stage 1

Figure 3 presents first-stage average effort levels by round for the NF, TF and PF condition.



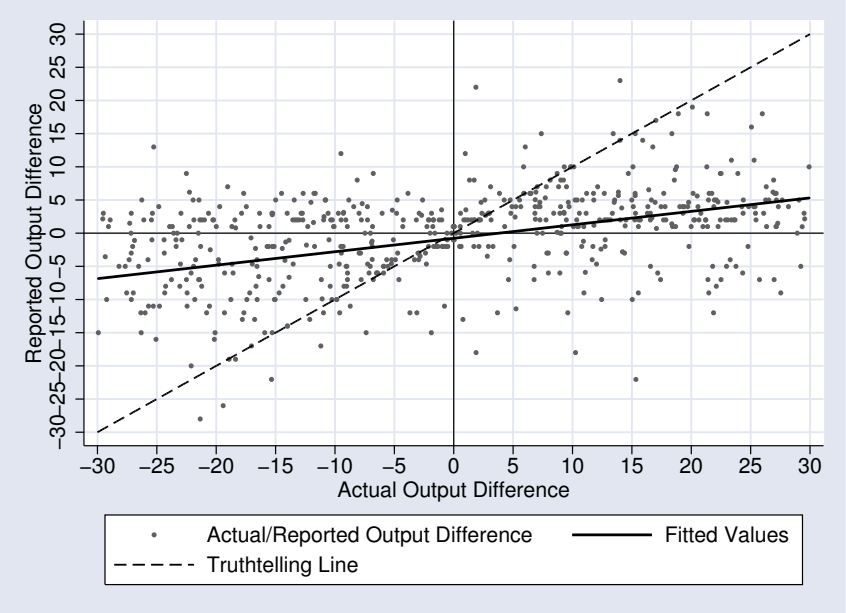

## Figure 4
## Average Effort Levels in NF, TF and PF Condition Stage 2

Figure 4 presents second-stage average effort levels by round for the NF, TF and PF condition.



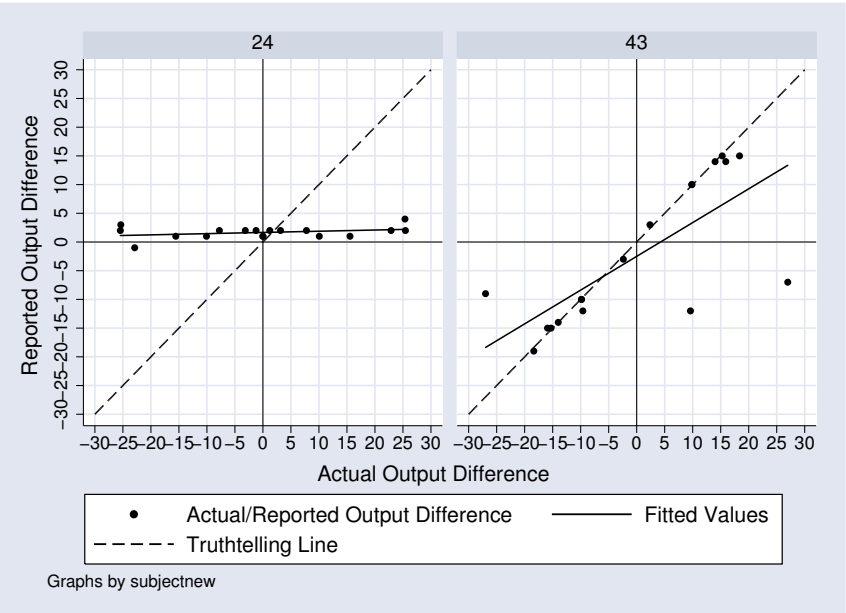
## Figure 5
## Actual/Reported Output Difference and Fitted Values in PF Condition

Figure 5 presents actual and reported output difference pairs and fitted values in the PF condition.




## Figure 6
## Actual/Reported Output Difference and Fitted Values for Subject #24 and #43

Figure 6 presents actual and reported output difference pairs and fitted values in the PF condition for subject #24 and #43.



# Figure 7
# Optimal and Actual Efforts PF Condition Stage 2

Figure 7 presents optimal and actual effort levels for Stage 2 in the PF condition. The optimal effort levels are chosen such that an agent chooses effort optimally using the correct posterior distribution of actual output differences for a given reported output difference.