Contracts for Agents with Biased Beliefs: Some Theory and an Experiment

Anja Sautmann*

Brown University, Department of Economics

anja_sautmann@brown.edu

May 2011 (first version July 2009)

Abstract

This paper experimentally tests the predictions of a principal-agent model in which the agent has biased beliefs about his ability. Overconfident workers are found to earn lower wages than underconfident ones because they overestimate their expected payoff, and principals adjust their offers accordingly. Moreover, the profit-maximizing contract distorts effort by varying incentives according to self-confidence, although only the most successful principals use this strategy. These findings have implications for the labor market; in particular, self-confidence is often correlated with gender, implying that principals would prefer to hire men over women simply because they are more overconfident.

*I would like to thank Guillaume Fréchette, Debraj Ray, and Andrew Schotter and the Center for Experimental Social Sciences at New York University. I am grateful to Pedro Dal Bó, Mark Dean, Wolf Ehrblatt, Matthew Embrey, Justin Kruger, Muriel Niederle, Louis Putterman, and Hugh Rabagliati for valuable comments and suggestions. This project was funded by the National Science Foundation under award no. 0849465.

It is a well-documented fact that people do poorly when assessing their own performance and abilities (Svenson (1981), Weinstein (1980), Lichtenstein et al. (1982); see Taylor and Brown (1988) for an overview). Most work has focused on overconfidence, which has been found to affect financial markets and managerial decision making (Odean (1999), Malmendier and Tate (2005), Malmendier and Tate (2008)), as well as market entry decisions in laboratory experiments (Camerer and Lovallo (1999), see their paper for further references). But individuals can also be underconfident, especially if they perceive a task as difficult (Clark and Friesen (2009), Kruger (1999), Moore and Cain (2007)). These belief biases are strong and persistent enough to lead to significant changes in outcomes and payoffs. For example, Barber and Odean (2001) find that financial traders, convinced that they can "outsmart" the market, make losses of up to 3.9% of annual income. Grubb (2009) shows that the typical cell phone plan menu is designed to screen customers who overestimate the precision of their demand prediction, making them pay more for their service.

This paper demonstrates that biased beliefs can also play an important role in the classical moral-hazard situation. I study a moral-hazard model in which the agents have biased beliefs about their own ability, and test its predictions in a laboratory experiment. To illustrate the basic idea, suppose that output depends positively on the agent's ability and effort, and incentive provision requires that a high output is rewarded with a high wage. Now assume that the employee is underconfident, that is, he underestimates his ability and therefore the chance that his output and wage are high. This means the principal must offer him a contract with higher expected wage than an unbiased agent would accept. Moreover, although it distorts effort, she will optimally shift some of the wage payments from the high to the low outcomes to reduce the loss from the belief difference. Conversely, if the agent is overconfident the principal provides very high incentives and pays a lower expected wage than an unbiased worker would receive for the same effort.

This paper tests experimentally if the agent's decision to accept a wage offer is affected by the self-confidence bias, and if principals adjust their contract offers accordingly. The experiment allows me to control for ability, self-confidence and effort, which are typically unobserved in labor market data. I find that subjects in the employee role accept lower (higher) expected wages if overconfident (underconfident). The profit-maximizing strategy in the experiment is to reduce the expected payment and raise incentives for overconfident agents and vice versa for underconfident ones. Subjects in the employer role correspondingly decrease the payments to overconfident agents and thereby raise their own profit, but increase the expected wage for underconfident agents. The most successful principals also adjust

incentives as predicted by the theory.

The results of this study have potentially important implications for the labor market. Incentive contracts are shown to entail redistributive effects between employers and employees. In addition, when effort and ability are complements the theoretical predictions imply that an overconfident (underconfident) employee works too hard (too little) for given incentives, and the profit-maximizing contract raises (lowers) incentives even further. This is in sharp contrast to an efficiency-minded social planner, who would choose flatter (steeper) incentives to correct for the distortion. Interestingly, the incentives chosen by the principals in the experiment resemble the social planner choice more closely than the model would predict, and the majority of principals do not make use of the profit-maximizing incentive adjustment. I will explore this issue in some detail in section 4.2, with particular attention to the possibility that the subjects play a pooling equilibrium in a signaling game.

A second implication of the model is that employment outcomes will systematically differ for populations that differ in self-confidence. For example there is robust evidence for gender differences in self-confidence in both the psychology and economics literatures. Barber and Odean's female traders are less overconfident and make fewer losses than their male counterparts; women underestimate their chances of success in tournaments, while men enter contests they are unlikely to win (Gneezy et al. (2003), Gneezy and Rustichini (2004), Niederle and Vesterlund (2007)[1]; see also Beyer (1990), Deaux and Farris (1977), Bengtsson et al. (2005)). This correlation implies that an employer who makes contract offers based on self-confidence will appear to discriminate between men and women. Indeed, I find that self-confidence in this experiment is correlated not only with gender but also with race, and that Asian (or Asian-American) and female subjects are sorted over-proportionally into the underconfident group. Conversely, note that the correlation of self-confidence with attributes like gender and race means that these attributes can serve as an indicator for the agent's self-confidence if beliefs are *not* directly observed. In a variant of statistical discrimination it is then actually optimal to offer different wages to men and women or to white and Asian employees.

The next section summarizes the related literature. Section 2 discusses the theoretical results in a general principal-agent model. Section 3 describes the experimental setup. The experimental results are presented in section 4, and section 5 concludes.

[1]Niederle and Vesterlund (2007) focus on the residual gender gap in tournament entry *after* controlling for beliefs as evidence for women's dislike of competition. But a sizable portion of the gap is accounted for by belief differences.

1 Related Literature

The theory predictions in this paper draw on an earlier working paper (Sautmann (2007)), but independent work by De la Rosa (2007) and Santos-Pinto (2008) is closely related. Both authors use models similar to the experimental setup used here, where output is discrete and the agent can be over- or underconfident about the probability of high output levels or about the effect of effort onto these probabilities. Santos-Pinto discusses the effect of an agent's bias on the principal's welfare, while De la Rosa focuses on the impact on effort and agents' welfare and the interaction of overconfidence with risk aversion. Adrian and Westerfield (2009) and Keiber (2006) study the effects of other types of belief biases on the allocation of risk between agent and principal. A related question, that of using contract offers to screen agents with heterogenous prior beliefs, has been investigated by Eliaz and Spiegler (2008) and Landier and Thesmar (2009).

The principal's contract choice problem studied here is also related to the literature on contracting under non-common priors, started by Morris (1994), which argues that individuals with different beliefs can mutually benefit from speculative trade by betting on outcomes to which they attach different probabilities. Eliaz and Spiegler (2007, 2009) look at the mechanism design aspect of such bets when the state of the world cannot be independently verified. The size of the bet is constrained because players may manipulate the bet after the state is realized (ex post), by playing a Nash equilibrium different from the outcome in the "bare" game. The principal solves a similar mechanism insofar as the wage difference between high and low outputs acts as a bet on the agent's ability, constrained only by the effect of the wage differential on the agent's unobservable effort. The effort distortion that stems from the principal's adjustment of incentives can be viewed from this angle: if the bet on output is too high, the agent is incentivized to (ex ante) manipulate the probability of the output that lets him win the bet.

To my knowledge this paper is the first to test contract choice with biased beliefs experimentally. It also adds a theoretical observation about the difference between the principal's contract choice and that of a social planner and the implications for effort and efficiency. Unlike in the literature on speculative trade, which can only be mutually welfare enhancing if the goal is to maximize subjective expected payoffs, it is assumed here that both the agent and the social planner value the actual expected outcome according to the unbiased ability distribution of the principal. This heightens the importance of the link between self-confidence and gender or race, because it means that overconfident (underconfident) agents can actually lose (win) from their belief bias.

2 Belief Biases in a Principal Agent Model

Consider a principal-agent setting in which the agent is hired to work with a technology owned by the principal. The agent produces good y with a stochastic output function, which depends on the agent's ability a and effort e:

$$y = f(a, e) + \theta,$$

where f is strictly increasing in a and e. Assume that f is smooth with bounded derivatives, and that the error term θ is independent of a and e and has mean zero. Effort costs the worker $C(e)$. For simplicity, assume that both parties are risk neutral.

Self-confidence enters the model through the agent's beliefs about his own ability. The principal, who is unbiased, holds beliefs about a that are identical to the true ability distribution P, but the agent's belief is given by a different distribution A. This might reflect that the principal has experience from contracting with many agents and is therefore less biased than the agent. The agent is *overconfident* if A dominates P and *underconfident* if P dominates A in the sense of strict first order stochastic dominance. The principal is aware of the agent's belief bias and can therefore adjust her wage offer accordingly. Note that the two parties can "agree to disagree" if they hold different priors regarding agent ability, even if they update their beliefs in a Bayesian fashion.

Since the principal cannot monitor the agent's effort, she must condition the worker's wage on output y. We restrict attention to linear wage contracts of the form $W(y) = ry + K$, where r is the piece rate and K is a lump-sum payment (a few words on this restriction later on). For a given contract and effort level, the agent therefore expects his payoff to be $r\mathrm{E}_A f(a, e) - C(e) + K$, and the principal's expected profit is $(1 - r)\mathrm{E}_P f(a, e) - K$. The subscripts on the expectations operator indicate that the same random variable, ability, is being evaluated under two different priors. After receiving a contract offer, the employee decides if he wants to accept based on his outside option and then chooses an effort level.

The Agent's Effort Choice. The benchmark for the analysis is the surplus-maximizing effort level according to the agent's beliefs, e^a, for which

$$\mathrm{E}_A f(a, e^a) - C(e^a) \geq \mathrm{E}_A f(a, e) - C(e) \text{ for all } e. \tag{1}$$

This is the effort level he would choose if he owned the production technology himself. If instead given a piece-rate r, the agent's effort choice $e(r)$ must satisfy the incentive constraint

$$r\mathrm{E}_A f(a, e(r)) - C(e(r)) \geq r\mathrm{E}_A f(a, e) - C(e) \text{ for all } e. \qquad (IC)$$

If $r = 1$, he chooses $e(r) = e^a$. For any other piece rate r, (1) and (IC) imply

$$\mathrm{E}_A f(a, e^a) - \mathrm{E}_A f(a, e(r)) \geq C(e^a) - C(e(r)) \geq r \left[\mathrm{E}_A f(a, e^a) - \mathrm{E}_A f(a, e(r))\right], \qquad (2)$$

and for an r greater than one this can be satisfied only if $e(r) \geq e^a$, and vice versa.

Lemma 1 *A piece rate $r > 1$ leads to an effort choice $e(r)$ greater than e^a, while a piece rate $r < 1$ implements an effort level $e(r)$ below e^a. The statement holds strictly if e^a is unique.*

The Principal. Now consider a profit-maximizing principal. For any given piece rate r and corresponding effort choice $e(r)$ she will choose K such that the participation constraint is satisfied with equality,

$$r\mathrm{E}_A f(a, e(r)) - C(e(r)) + K = U, \qquad (PC)$$

where U is the agent's outside option. Substituting for K in the principal's objective function (i.e. her expected profit), her optimization problem is

$$\max_r \left[(\mathrm{E}_A f(a, e(r)) - C(e(r))) + (1 - r)\left(\mathrm{E}_P f(a, e(r)) - \mathrm{E}_A f(a, e(r))\right) - U\right].$$

The first term of this expression equals the *agent's* expected net output. By selling the production technology to the agent, i.e. letting $r = 1$ and charging a lump sum that equals agent's expected net surplus, the principal can realize at least a profit of $\mathrm{E}_A f(a, e) - C(e) - U$. Note that this is the maximal possible profit if agent and principal have the same beliefs.

Now suppose the agent is underconfident, so that $\mathrm{E}_P f(a, e) - \mathrm{E}_A f(a, e) > 0$. Starting at $r = 1$, the principal increases her profit by replacing some of the agent's flexible pay with a fixed wage, i.e. lowering r and raising K such that the participation constraint remains satisfied. The increase in the lump-sum payment K is $(1 - r)\mathrm{E}_A f(a, e)$, compensation for the lower output share the agent expects to receive. But this is more than offset by the share in the expected profit that the principal now keeps for herself, $(1 - r)\mathrm{E}_P f(a, e)$. Even though there will be a small loss due to the downward distortion of effort, the net effect of this change on the principal's profit is positive. For a formal argument, note that at $r = 1$, a marginal increase in r changes the principal's profit by[2]

$$\frac{\mathrm{d}e(1)}{\mathrm{d}r}\left[\frac{\mathrm{d}\mathrm{E}_A f(a, e^a)}{\mathrm{d}e} - C'(e^a)\right] - \left[\mathrm{E}_P f(a, e^a) - \mathrm{E}_A f(a, e^a)\right].$$

[2]By the implicit function theorem there is a differentiable $e(r)$ describing the agent's effort choice. At $r = 1$, $\frac{\mathrm{d}}{\mathrm{d}r}e(1) = -\frac{\frac{\mathrm{d}}{\mathrm{d}e}\mathrm{E}_A f(a, e^a)}{\frac{\mathrm{d}^2}{\mathrm{d}e^2}\mathrm{E}_A f(a, e^a) - C''(e^a)}$.

The first term is zero since e^a maximizes $E_A f(a,e) - C(e)$, so that reducing r leads to a strict increase in profit. By contrast, raising it to $r > 1$ would not only distort effort away from e^a but it would also impose an additional cost on the principal. In the case of an overconfident agent, the argument is reversed. Now the principal pays a high piece-rate – inducing a higher effort level – and lowers K. This is profitable because it reduces the agent's expected wage from the principal's perspective.

Observation 1 *A profit-maximizing principal chooses $r^p > 1$ and induces an effort level $e^p > e^a$ if the agent is overconfident, and $r^p < 1$ and $e^p < e^a$ if he is underconfident.*

Their overly optimistic output estimate means that the overconfident agents lose from this employment relationship. From (PC) it is immediate that an overconfident agent's expected pay under P is less than his outside option U, even though he believes he will receive U. On the other hand, as long as r is positive, the underconfident agent actually gets more than U.

Observation 2 *Under the agent's beliefs, his expected payoff always equals U. But under the principal's belief it is less than U if the agent is overconfident and more than U if he is underconfident (unless $r^p < 0$).*

As a consequence, overconfident workers tend to be more attractive as employees. This is best seen when writing the principal's profit as $r\mathrm{E}_A f(a,e(r)) + (1-r)\mathrm{E}_P f(a,e(r)) - C(e(r)) - U$. In the special case where output is separable in effort and ability, so that the agent's response to incentives is independent of his beliefs and $e(r)$ is the same for both types of agents, it is clear that this expression is greater for any given r if the agent is overconfident, and therefore that an overconfident employee generates higher profits for the employer (as long as $r^p > 0$). More generally, the expression can only be lower for an overconfident agent if the distortion to $e(r)$ is much stronger for him, overcompensating the difference in $r\mathrm{E}_A f(a,e(r))$. Santos-Pinto (2008) shows in a discrete model that overconfident agents increase the principal's profit and underconfident agents decrease it, provided the optimal incentive scheme is increasing in output, and effort and self-confidence are complements (so the discrepancy in expected output according to agent's and principal's beliefs is higher when effort is higher).[3] For the purposes of the experiment, the following observation is sufficient.

Observation 3 *If output is separable in ability and effort, the profit from hiring the overconfident agent is strictly higher under the principal's beliefs as long as $r^p > 0$ for the underconfident worker.*

[3]Effort and self-confidence are complements here, but I only consider linear contracts.

In summary, the belief bias and the principal's response to it lead to systematic differences in payoffs, effort levels and incentives between over- and underconfident agents. An overconfident agent has a lower net payoff than an underconfident one, but works harder, and the principal makes higher profits from him. If men are more likely to be overconfident, this implies that they have a lower net payoff from working than women. It should be emphasized that their expected *wage* may still be higher than women's, since they also exert greater effort.[4]

The results so far have particular significance if ability and effort are complements in production, so that higher ability implies a higher marginal effect of effort. In this case we can compare the principal's contract choice with the surplus-maximizing effort level e^* a social planner would choose. The planner's beliefs are the same as the principal, so her problem is

$$e^* = \arg\max_e \mathrm{E}_P f(a,e) - C(e).$$

The assertion here is that social welfare is evaluated at the ability distribution P, without taking into account the agent's subjective ex-ante utility. Note that all previous results hold without reference to a "true" distribution of ability, whereas now we take a stand which beliefs to use for welfare judgements.

Intuitively, an overconfident agent overestimates the marginal return to effort and therefore works too hard, and the opposite holds for an underconfident worker. Unlike a profit-maximizing principal, the planner chooses incentives that correct this distortion. The first step to see this is to relate e^* to e^a. Note that

$$\mathrm{E}_P f(a,e^*) - \mathrm{E}_P f(a,e^a) \geq C(e^*) - C(e^a) \geq \mathrm{E}_A f(a,e^*) - \mathrm{E}_A f(a,e^a). \tag{3}$$

Now suppose the agent is overconfident and ability and effort are complements in output, so that $f(a_2,e_2) - f(a_2,e_1) \geq f(a_1,e_2) - f(a_1,e_1)$ for any $a_2 > a_1$, $e_2 > e_1$. If $e^* > e^a$,

[4]Note that I did not fully solve the model in order to focus on the main insights. In some cases, an optimum may not exist; e.g. if the agent is overconfident, the principal's payoff may approach infinity as $r \to +\infty$. For a finite optimal r it is sufficient that $\mathrm{E}_A f(a,e) - C(e) \to -\infty$ as $r \to +\infty(-\infty)$ but $E_A f(a,e) - E_P f(a,e) \to 0$; in other words, the belief discrepancy becomes unimportant at extreme effort levels. Alternatively, limited liability or risk aversion impose constraints on the piece rate. Although we have only considered linear contracts, the main insights carry over to other settings, see e.g. Santos-Pinto (2008). In general, the principal will raise relative wages for those output levels that have a higher probability weight under A than under P; in other words, overconfident agents are paid more for high outputs, thereby adding to their work incentives, and vice versa for underconfident employees.

the expression $f(a, e^*) - f(a, e^a)$ must be a positive, increasing function of a. But then (3) contradicts first order stochastic dominance of A over P, so it must be that $e^a \geq e^*$.

Lemma 2 *Suppose ability and effort are complements. If the agent is overconfident the efficient effort level from the agent's perspective is higher than that from the planner's (principal's) perspective; $e^a \geq e^*$. If the agent is underconfident then $e^a \leq e^*$. The inequalities are strict if a and e are strict complements and either e^a or e^* is unique.*

Suppose the social planner implements the social optimum by way of a linear wage contract, either by employing the agent herself, or by regulating the contract terms. A biased agent will not choose the socially optimal effort level at a piece rate of one, so the social planner corrects the effort distortion by lowering the piece rate for an overconfident type and raising it for an underconfident agent. Combining the lemma with observation 1 shows that the principal, by contrast, distorts effort away from the social optimum:

Observation 4 *If ability and effort are complements, $r^p > r^*$ and $e^p > e^*$ if the agent is overconfident, and $r^p < r^*$ and $e^p < e^*$ if the agent is underconfident.*

Observation 4 illustrates the importance of studying heterogenous beliefs. It contradicts the conclusions of the classical moral hazard model, where the principal implements the same effort level as a social planner would (realizing the first best outcome with risk neutral parties and the second best under risk aversion). Here, the principal does not imitate the planner to counteract the distortion from the agent's effort choice, but even adds to it. Since the principal maximizes profit, the resulting welfare loss is borne by the agent.

If effort and ability are substitutes, lemma 2 of course works in the opposite direction, and the agent's bias may reduce the efficiency loss from the principal's distortionary choice of incentives (yet even then the principal will in general not choose the socially optimal level of effort). Substitutability of effort and ability may occur for instance when the agent works towards a fixed quota or goal. However, there are many tasks in which ability and effort naturally complement each other. Take the example of time (effort) vs. cognitive skills or specialized knowledge (ability): in the same time, a more able employee will produce higher output than his less able colleague.

3 The Experimental Design

The experiment to test the predictions of the previous section consists of two stages and a questionnaire, all carried out on the computer. In stage 1, ability and self-confidence

are measured, and in stage 2 subjects interact as agents and principals. The questionnaire collects data on personal characteristics. Subjects can earn and spend points, which are converted into US dollars at the end of the session.

At stage 1, before receiving instructions on the rest of the experiment, participants take a 10-item, multiple-choice trivia quiz, modeled after the quizzes in Healy and Moore (2007). Afterwards they are asked to guess their own trivia score. They earn points both for correct answers to quiz questions and for guessing accuracy.[5] The difference between the guess and the true score is used as a measure of the subject's level of self-confidence. After the guess, stage 1 concludes and subjects are given instructions for stage 2.

At the start of stage 2, participants are assigned the role of employer or employee. To create two groups with discernibly overconfident and underconfident agents in all experimental sessions, the subjects in each session are ordered by level of self-confidence, and the highest and lowest quartiles become employees in group O(verconfident) and group U(nderconfident) (in the experiment, they were neutrally named group 1 and 2). The remaining subjects become employers and are randomly assigned to either group. Subjects are informed that the group assignment is based on the results from Stage 1 but not given more detail.

Stage 2 has 30 paying rounds, split evenly into treatment (T) and control (C) blocks. The order of T and C is randomized in each experimental session. The experiment therefore uses a mixed design: the group division is maintained throughout the experiment (between-subject design), while both groups undergo the treatment and control (within-subject design). This helps to control for unobserved differences between over- and underconfident agents, e.g. in risk aversion or overall optimism, as well as differences between treatment and control that are common to both groups, e.g. differences in the perceived level of risk.

In each round of stage 2, an employer and an employee from the same group are paired up to enact the principal-agent contracting situation. In the treatment T, "ability" (a) is represented by the trivia score of the employee. Agents have no information about the score besides their own guess, so they are subject to the self-confidence bias, reflected in the difference between that guess and their true trivia score. The principals, on the other hand, learn the trivia scores and score guesses of the employees in their group (although not that of the individual agent they are matched with). This means they are unbiased and aware of the agents' bias.

[5]The payoff for guess G when the actual score is S is $100-(G-S)^2$. Under risk-neutrality, the payoff maximizing guess is the expected value of S, i.e. the subject's ability expectation that we are interested in here.

The control C is identical, except that a is now a number between 0 and 10, randomly assigned to the employee by the computer, while the agent's test score (ability) does not affect the outcome. Employers and employees both learn the numbers assigned within the group (but not the number of each individual agent), so that everyone has the same, unbiased information about the distribution of a. The assigned numbers have the same distribution as the trivia scores of the agents in the group. Average ability is therefore constant between treatment and control, and the principals' beliefs are the same.[6] During instructions, all subjects see examples of the screens on which the employer and employee make decisions, so that it is common knowledge what information is available to each side.

The following is a detailed breakdown of a typical round.

Contracting in a Round in Stage 2: At the beginning of each round, a principal and an agent from the same group are randomly and anonymously matched. For the purpose of the experiment, a discrete version of the model with two output levels, high (H) or low (L), is implemented. If the outcome is H, the principal earns 95 points, if it is L, 60 points. The employer makes a wage offer to her employee by choosing wages $w_H \in \{0, \ldots, 95\}$ and $w_L \in \{0, \ldots, 60\}$ for H and L, respectively. There are three output draws, so that the principal can make profits between $3(95 - w_H)$ and $3(60 - w_L)$, and the theoretical maximum earning per round is 285.

After the employer submits the two wages, the employee is shown the offer on his screen. He has a budget of 20 points. If he rejects the offer, he receives an additional 100 points and the principal gets 0 points. Otherwise he chooses between investing the 20 points (in full) or investing nothing. The budget and the restriction on wages serve to separate past and present payoffs. Each subject is provided feedback about payoffs at the end of the round.

The probability of the high outcome as a function of the agent's investment (the equivalent of effort) and his score – either the trivia score in the treatment, or the assigned number in the control – is described by table 1 (available to the subjects in graphical form). Investing the 20 points increases the chance of outcome H in all three draws by 30%. For a risk-neutral employee, this is profitable whenever the wage gap $w_H - w_L$ is at least 23. The expected output increase is 31.5, so it is socially optimal, and profit-maximizing if the agent is unbiased, to induce the high effort level.[7]

Table 10 in the appendix lists the optimal contracts when the agent's self-confidence level

[6]The principals might notice that the distributions are similar, but the order in which scores/numbers are displayed is randomized.

[7]Note that this is true here, but does not hold in general when wages are bounded.

Table 1: Probability of the high outcome for both investment levels and different a.

Probability of H		Score/ability a			
		0-4	5-6	7-8	9-10
Investment	0	0.05	0.25	0.45	0.65
	20	0.35	0.55	0.75	0.95

is known, assuming optimal effort choices, profit-maximization and risk neutrality on both sides. The numbers in this experiment were chosen so that a decrease in the agent's self-confidence by one level (i.e. the score guess is, for example, 6, but the score is 7 or 8) makes it profitable to distort investment. The employer in fact optimally chooses a negative wage difference for the underconfident type. At the same time, the wage gap is maximized for the overconfident agent. More generally, the principal's profit rises with incentives if the agent is overconfident and falls if he is underconfident (see table 9 in appendix B). The data from the experiment is used to test the following general predictions:

1. The subjective expected payoff is the same for all agents.
2. The expected profit for the principal is higher if her employee is overconfident.
3. The expected payoff is higher for the underconfident than for the overconfident agent.
4. All else equal an overconfident agent receives a lower expected wage than an underconfident agent.
5. The strength of incentives ($w_H - w_L$) is lower for underconfident agents, and if the belief bias is large enough, they are induced to work less than overconfident agents.[8]

In section 4.2 I will also examine the (hypothetical) choice of a social planner, given the behavior of the agents in the experiment.

Note that predictions (2)-(4) rely on $w_H - w_L \geq 0$ for both types of agents. This is not optimal, since profit-maximization requires negative incentives for underconfident agents. In

[8]Observe that a high ability level leads to a greater weight on the marginal utility at the high over the wage and thus under risk aversion to a lower expected marginal utility of effort, so that over- and underconfident agents may exert different effort for equal incentives. This may confound the effort distortion from the self-confidence bias. As we will see this is not an issue here.

that case we would expect, for example, higher profits for the principals in the treatment compared to the control under both over- and underconfidence. However, as will be seen, principals in the experiment do not set negative incentives for underconfident employees, and this behavior was not entirely unexpected.[9] Instead of presenting the full set of conditional predictions I therefore focus here on the case where $w_H - w_L$ is always positive. Note also that prediction (5) can be evaluated separately from the others, in the sense that the principal may adjust the wage level to the self-confidence bias without also adjusting incentives. For example, if the agent is overconfident, the principal can benefit from simply lowering both the high and the low wage until (PC) is satisfied with equality, and we will see that this is what the principals in the experiment do. Of course, according to theory increasing incentives would allow them to lower the expected wage even further, and I will discuss this discrepancy in more detail in section 4.2.

The description of the experiment concludes with a few remarks on the connection between theory and experiment and the design choices for the latter. These choices draw in part on the insights from earlier laboratory experiments on moral hazard. The first full test of the hidden-action model was conducted by Berg et al. (1992), followed by Epstein (1992) and Keser and Willinger (2000).[10]

Preferences. The assumption of linear, separable preferences in the model is of course also a simplification. If the experimental subjects are risk-averse, the high risk inherent in large wage differences may put a limit on the principal's incentive adjustment in response to the agent's belief bias (see also De la Rosa (2007)). It can be shown, however, that the model predictions continue to hold qualitatively when some risk aversion is present (see appendix A for details).

Modified setup. The experimental design deviates in several points from the original principal-agent model. The simple binary specification for output, wages and investment was

[9] *Some* incentive adjustment was expected, but that the principals do not choose negative "rewards" for high outcomes was considered a distinct possibility. The choice for the experimental design was between restricting the principal's choices to contracts with only positive incentives or offering more than two investment levels – both at a loss of clarity and tractability for the subjects – , and a potential ambiguity in the theory predictions and results, and I opted for the latter.

[10] In addition, a range of experiments have studied agency situations in the context of preferences for fairness and reciprocity and intrinsic motivation (esp. Anderhub et al. (2002); see Fehr and Schmidt (2006) for an overview).

Table 2: Optimal contracts, all experimental sessions and groups.

session	group	scores	guesses	opt. wage offers C		opt. wage offers T		diff-in-diff incentives	diff-in-diff investment
				w_L	w_H	w_L	w_H		
1	O	3 3 3 4 5	5 6 5 6 7	32	55	0	73		
	U	5 5 6 7 8	3 4 4 5 4	27	51	44	2	116	1
2	O	1 4 4 5 7	5 5 5 6 8	32	55	0	73		
	U	5 5 7 7 8	2 3 3 5 4	27	51	45	0	119	1
3	O	2 5 5 7	4 6 5 7	34	33	34	33		
	U	6 6 6 7	4 3 4 5	27	51	44	2	66	1
4	O	3 3 5 5 5	5 5 5 5 6	32	55	0	73		
	U	5 6 7 8 8	3 4 6 7 6	27	51	34	33	75	1
5	O	4 4 5 6	6 5 7 7	32	55	0	73		
	U	4 6 7 8	3 4 5 4	34	33	44	2	91	0
weighted average								94.7	0.83

chosen to make outcome and action spaces easy to understand and display on a computer screen. The three output draws ensure that the effect of effort and ability creates real payoff differences. Expected output is separable in ability and effort, implying that the profit prediction of observation 3 is unambiguous and that any effort distortion is a consequence of the principal's response to the agent's bias. In other words, the self-confidence bias affects effort only through the principal's contract offer (obs. 1), whereas the agent's optimal effort level e^a equals e^* (i.e. lemma 2 holds only weakly). Observation 4 still applies to sufficiently underconfident agents, i.e. effort is distorted away from the optimal level.

Perhaps a more controversial deviation from the theory is that the principal in the experiment does not know which agent in her group she is dealing with. Rather, she has an equal chance to interact with one of four or five different agents. This choice was made on the one hand to prevent the principal from identifying individual agents and playing a repeated game with them, and on the other to reduce the possibility of the agents learning from the principal's contract offer. At the root of this is the difficulty of inducing different priors in an experiment. While different ability priors are plausible in real life principal-agent situations, where both employers and employees have a rich history of interactions and little knowledge of the sources for each others beliefs, this is not always so in an experiment.

Fang and Moscarini's (2005) paper on belief biases and incentives illustrates the problem.

The authors study the role of overconfidence for wage compression and show that, if a firm makes differentiated contract offers based on its estimate of worker ability, workers can infer their true productivity from the offer. This means their 'morale' is destroyed if they initially hold overly optimistic beliefs, and it can be optimal to pool all contract offers. By a similar argument, imagine the principal knows the exact test score of each agent and this is known to all subjects. If the principal offers low or negative incentives to an underconfident agent, the agent may conclude that his beliefs are incorrect. However, when an employer faces a group of agents, each employee can at most partly update his beliefs, since the other agents may have very different test scores and score guesses.

I will return to the possibility of a common prior (and the signaling/pooling equilibria arising in a game with common priors) when analyzing the principal's choices of incentives in detail in section 4.2. But even within the original framework of heterogenous priors we must verify that the theory predictions hold in the new setup. Table 2 calculates the optimal contract offers for both groups in treatment and control for the actual experimental sessions, assuming risk neutrality and taking into account that the principal does not know which of his group's agents he is interacting with. For example, in session 3, group O, the principal knows that the agent has a score of 2, 5, 5, or 7. Large intra-group differences can make it optimal to choose a contract with no incentives, and (3 O) is such a case: the risk of rejection from the lowest-ability agent (with score 2 and in T belief 4) requires high wages, but the 0.25 chance of meeting the high-ability agent instead, who gets the high wage very often, makes an incentivized contract too costly. In this case, the contract closest to the flat wage of 33.33 is optimal. This occurs several times in the experiment, but in all but one instance (investment in session 5) the qualitative predictions above are unaffected.

Training and feedback. Previous experimenters, starting with Bull et al. (1987), found that principals sometimes have difficulty in choosing payoff-maximizing strategies and make persistently suboptimal choices if agents' effort is not observed (Keser and Willinger (2000)). The principals therefore learn the agent's investment choice in each round. Since training as an agent increases efficient contract choices (Berg et al. (1992)), all subjects are shown a typical round from the agent perspective before roles are assigned. Principals can also test what their wage offer looks like to an employee before submitting it. To help subjects understand the setup it is framed as an employment situation (see Cooper et al. (1999)). Finally, the first two rounds in T and C are trial rounds.

Fairness and reciprocity. The experimental literature suggests that agents punish "unfair" offers even at their own disadvantage, implying that principals may be reluctant to make

such offers in the first place. The agents therefore do not learn the principal's payoffs for H and L (a placeholder is used in the instructions). This does not affect their optimization problem, but it prevents them from judging an offer as "unfair". It is also a fairly common feature in typical employer-employee relationships.

4 Results

The data analysis will focus on the differential effect of the belief bias on contracting and outcomes for over- and underconfident agents. Most results are presented as difference-in-difference OLS estimates from individual random effects regressions, with standard errors clustered by experimental session. The regression equation is

$$y_{i,t} = \alpha + \beta_{O\times T}(O \times T) + \beta_O O + \beta_T T + u_i + e_{i,t}$$

for each dependent variable y. The independent variables are indicators for the overconfident group (O), the treatment periods (T), and the interaction of the two ($O \times T$), which equals one if a subject in group O is in a treatment period and zero otherwise. The coefficient of interest is the one on $O \times T$: it measures the effect of an overconfidence bias relative to the effect of an underconfidence bias, after controlling for group differences and for any level effects of the treatment that are the same in both groups.

The experiment was programmed with the software Z-Tree (Fischbacher (2007)) and conducted in the computer laboratory of the Center for Experimental Social Sciences at New York University. Most subjects were NYU undergraduates. There were five sessions, two with 16 and three with 20 participants. In three of the sessions (with 56 subjects) stage 2 started with the treatment T, and in two of them (36 subjects) it started with the control C. Subjects received a US$10 participation fee, and their earnings from the experiment were converted into US$ at an exchange rate of 0.005. Total profits ranged from 648 to 3043 points for the principals, with a mean of 2001 and a standard deviation of 587, and for the agents from 3335 to 6208, with mean 4369 and standard deviation 623.

Table 3 lists average ability, score guess, and self-confidence level (guess minus actual score) for the agents in the two groups. On average, the difference between the estimated and true trivia score was 1.43 for agents in group O, ranging from 0.75 to 1.6 in the different sessions, and -2.26 for group U (with range -3 to -1.6). In all but session 4, average self-confidence in the underconfident group was below -2, that is, score guesses were more than two points too low, making it optimal to distort effort downwards (see table 2). Note also that there is

Table 3: Ability and self-confidence of agents in O and U.

Group O (N=23)	**Mean**	**Std. Dev.**	**Group U (N=23)**	**Mean**	**Std. Dev.**
Ability (trivia score)	4.261	1.453	Ability (trivia score)	6.391	1.196
Score guess	5.696	0.974	Score guess	4.13	1.18
Self confidence	1.435	0.992	Self confidence	-2.261	0.964

Table 4: Group averages in treatment and control ([a]accepted contracts only).

	(O,C)	(O,T)	(U,C)	(U,T)
Exp. profit for principals[a]	84.1	89.9	94.0	83.6
Exp. profit for principals	61.7	61.0	76.3	66.6
Realized profit principals[a]	89.5	89.7	95.2	88.3
Realized profit principals	63.5	60.3	74.8	68.3
Exp. profit for agents	139.6	133.7	152.7	159.3
Realized profit agents	138.6	132.7	153.5	157.8
Exp. profit, agents' beliefs	138.4	142.7	152.8	147.4
Incentives $w_H - w_L$[a]	26.6	30.0	17.0	20.6
Incentives $w_H - w_L$	28.1	30.0	18.6	21.1
Exp. wage[a]	139.3	133.2	151.3	161.0
Exp. wage	124.6	115.3	145.2	149.7
Investment (proportion)[a]	0.62	0.65	0.55	0.59
Offer rejection (proportion)	0.27	0.32	0.19	0.20

an ability difference of more than two points between groups. It is not surprising that high-performing individuals will tend to underestimate their outcome and vice versa, assuming that the actual success rate is subject to some stochastic variation independent of ability. Sorting subjects by self-confidence therefore led to some sorting by test score as well.[11] The difference-in-difference design of the experiment will help to filter out this group effect of ability on outcomes.

[11]This is reminiscent of Healy and Moore (2007) who point to a robust negative correlation between self-confidence and task difficulty as evidence for Bayesian inference.

Table 5: Random effects regressions: expected profits for principals ([a]accepted contracts only) and for agents under true probabilities and under agents' beliefs.

	Exp. profit P	**Exp. profit P[a]**	**Exp. profit A**	**Exp. pr. A's beliefs**
	coef./(stde.)	coef./(stde.)	coef./(stde.)	coef./(stde.)
O×T	8.888	12.701**	-12.537**	9.640**
	(7.357)	(4.055)	(2.996)	(2.308)
O	-14.583	-8.567	-13.077	-14.372
	(12.277)	(13.906)	(11.144)	(10.331)
T	-9.642	-7.537**	6.542**	-5.407**
	(7.264)	(2.849)	(2.460)	(0.964)
Intercept	76.291**	94.159**	152.726**	152.820**
	(10.747)	(12.625)	(9.693)	(9.523)
No. obs.	1380	1042	1380	1380

Significance levels: †: 10% ∗: 5% ∗∗: 1%

4.1 Group Averages and Difference in Difference Estimates

Table 4 lists the averages of all relevant variables by groups and conditions. Tables 5, 6 and 7 report the corresponding difference-in-difference estimates. For principals, all (offered) and accepted contracts are shown separately to distinguish differences in principals' choices from variation in what agents were willing to accept.

The expected wage payment and principals' and agents' profits are calculated using the actual probability of success, as determined by the agent's ability a and her investment choice. These are the same probabilities used by the computer program in the experiment to determine the final outcomes. For completeness, the averages of realized profits are shown in table 4 as well.[12] Finally, the table reports the expected profit according to the agent's belief about his ability. In the treatment, this belief is assumed to be the agent's score guess from the first stage of the experiment, and in the control it is the group average of assigned scores, known to both agents and principals.

Expected Profits for Principals and Agents. Table 4 shows that the switch from C to T

[12]Despite the large number of draws, some of the average realized profits are quite different from expected profits. Careful checks of the randomization procedure in the experimental program and the resulting outcome draws did not yield an obvious explanation, so I am bound to assume that it was indeed chance.

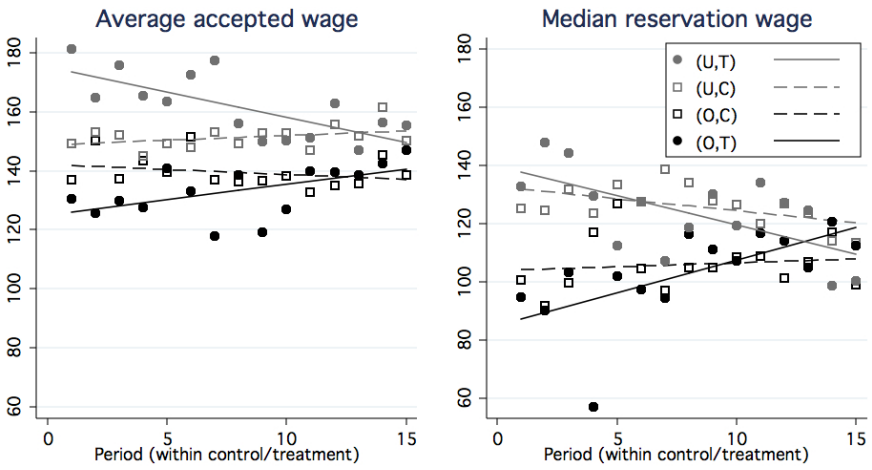


Figure 1: Left - period average of accepted expected wage, right - period median of expected wages between highest rejected and lowest accepted wage.

causes principals' profits to fall significantly and agents' to rise in group U, whereas this is not true in group O. Table 5 confirms that across all periods the relative effect of dealing with an overconfident agent on the employer's expected payoffs is positive, but negative for the agents.

But before studying these variables in more detail, consider the averages and the regression for the agent's subjective expected profit in tables 4 and 5. Unsurprisingly, given how it was calculated, agents' subjective profit expectation in the control is very close to actual expected profits, whereas in the treatment agents in O overestimate and in U underestimate their profit by about 10 points each. However, according to their own beliefs overconfident agents expect higher profits in the treatment than in the control, and underconfident agents expect lower profits. This "overshooting" is noteworthy: according to the model, principals adjust wages to make the agents' subjective expected utility equal to their reservation utility at all times, and we would expect only a small difference between T and C (i.e. the coefficient on O×U should be small and/or insignificant). Instead subjective expected profit seems to increase by 9.64 points for overconfident relative to underconfident employees.

As it turns out, this result is most likely a consequence of learning on the part of the agents. By observing the three outcomes per round in the treatment, agents can make inferences about their ability as the experiment proceeds. Their belief bias will therefore gradually decline, and the score guess from the beginning overestimates average real beliefs in O and underestimates them in U. This interpretation is confirmed by figure 1, which depicts two measures of the agents' acceptance decisions over time. On the left is the average accepted

Table 6: Random effects regressions: expected profits for principals ([a]accepted contracts only) and for agents under true probabilities and under agents' beliefs; first eight periods of each treatment-group combination (periods 3-10 and 19-27).

	Exp. profit P	**Exp. profit P**[a]	**Exp. profit A**	**Exp. pr. A's beliefs**
	coef./(stde.)	coef./(stde.)	coef./(stde.)	coef./(stde.)
O×T	23.671*	22.600**	-19.274**	4.140
	(11.902)	(5.882)	(5.645)	(4.023)
O	-18.656†	-11.244	-11.084	-12.752
	(10.994)	(14.081)	(11.471)	(10.407)
T	-16.125	-13.536**	11.522*	-1.573
	(10.019)	(4.104)	(5.219)	(2.504)
Intercept	74.886**	96.180**	150.220**	150.887**
	(9.240)	(11.502)	(8.634)	(8.426)
No. obs.	736	532	736	736

Significance levels: †:10% *:5% **:1%

expected wage in each period. Since the average may be "contaminated" by the principals' offer decisions, the right panel uses the median of all wages that lie between the highest rejected and the lowest accepted wage of that period as an estimate of the reservation wage. Both panels indicate an increase in the acceptance threshold in (O,T), but a decrease in (U,T), and the change is much larger than the equivalent changes in C. This is a clear sign of learning.[13] In what follows I will therefore focus on the first half of T and C in each experimental session, where learning effects are less strong. Repeating the regressions for profits for the first eight periods (table 6) shows that agents' subjective profit is unaffected by the self-confidence bias. Prediction (1) holds despite the positive coefficient in the full regression.

Result 1 *Given their (biased) beliefs, agents' subjective payoffs are unaffected by the self-confidence bias, consistent with the prediction that they will be paid their reservation utility*

[13]Any alternative theory would require that the treatment has different effects on the agent's subjective expected utility from the same contract offer in group O and U. A prime candidate would be a differential change, from C to T, to the riskiness of the contract offer in the two groups, but the change to incentives is similar – and on average fairly small – in O and U.

in all conditions.

Table 6 also shows that the principal can on average extract 23.7 points from an overconfident employee relative to an underconfident one, a gain of 36% over the average profit of 65.6 points in C. This is not just an effect of different agent decisions, because it holds for all contracts that the principals offer, including rejected ones. An overconfident agent, on the other hand, makes relative losses of almost the same size, namely 19.3 points on average, or 13% of the average profit of an agent in C. Since incentives are positive in all group-treatment combinations, this result strongly supports predictions (2) and (3).

Result 2 *The principal's expected profit increases when contracting with an overconfident agent, and decreases when hiring an underconfident agent, compared to an unbiased employee.*

Result 3 *The self-confidence bias decreases the expected profit of an overconfident agent, but increases that of an underconfident agent.*

Group and Treatment Effects. The significant coefficients on T in tables 5 and 6 indicate that a share of profits is shifted from principals to agents in the treatment. This is because self-confidence is 1.44 in O, but -2.26 in U, so the average agent underestimates his payoffs in T. The average true expected payoff for agents in the treatment must therefore be higher than in the control, and that for the principals consequently lower. The negative coefficients on O for the profits of agents and principals are most likely a result of the ability difference between O and U. Higher agent ability yields a larger pie to split between the contract partners in group U.

Wage Offers and Agents' Decisions. Table 7 shows the difference-in-difference estimates for expected wages, incentives, and agents' decisions in the first eight periods (see table 13 in appendix B for results for all periods). The self-confidence bias has a large and significant effect on the expected wage.[14] The relative effect of agent overconfidence on the wage payment is -20.6 points, 16% of the average expected wage in C. Expected wages decrease in O and increase in U from C to T. Moreover, the effect is neither accompanied by a jump in investment rates (higher investment rates would call for higher wages to compensate agents for their costs) nor by a relative change in rejection rates (which might indicate, for

[14]The coefficients in the regressions for the individual wages (not reported here) are both negative and have almost the same size, -5.23 for w_H and -5.53 for w_L. They are significant at the 10% level for accepted contracts, but for a one-sided test only when including all contracts. Recall that the expected wage is the result of three outcome/wage draws.

Table 7: Random effects regressions: wages and agent decisions, first eight periods.

	Exp. w.	**Exp. w.***	**Incentives**	**Incentives***	**Invest.**	**Reject.**
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)
O×T	-20.618*	-24.878**	0.293	-0.777	-0.052	-0.033
	(9.629)	(3.707)	(5.159)	(4.395)	(0.081)	(0.099)
O	-19.453	-10.253	6.734	8.817	0.070	0.103
	(15.267)	(15.296)	(6.235)	(5.904)	(0.086)	(0.066)
T	9.447**	14.814*	3.364	3.439	0.080	0.033
	(2.696)	(6.140)	(3.080)	(2.881)	(0.138)	(0.101)
Intercept	142.560**	149.791**	19.668**	19.000**	0.580**	0.217**
	(11.476)	(12.896)	(4.321)	(4.452)	(0.141)	(0.052)
No. obs.	736	532	736	532	532	736

Significance levels: †:10% ∗:5% ∗∗:1%

example, that principals follow the "wrong" offer strategy, but overconfident agents reject low offers less often, so that the wage difference is purely driven by agents' decisions), and it is present when looking at all wage offers as well as only accepted ones. This suggests that principals actively respond to the presence of a self-confidence bias by adjusting the wage *level*, increasing their own profits at the expense of overconfident agents, but yielding to the higher wage demands of underconfident agents.

Result 4 *All else equal overconfident agents are offered – and accept – lower expected wages.*

At the same time, however, the principals are not using the wage *difference* to modify the effect of the self-confidence bias on their profits. Agents' investments consequently do not change. The coefficient on $O \times T$ for incentives in table 7, as well as for investment, is close to zero and insignificant.

Result 5 *The self-confidence bias has no effect on the wage gap $w_H - w_L$. There is consequently no change to the agent's effort choice.*

This result is quite robust across periods and sessions, and I will investigate its possible causes in more detail in the next section.

Robustness Checks. The analysis was repeated on subsets of the data to check for order or session effects. There were not obvious outliers when analyzing the data session by session,

and results did not depend on whether the treatment or control came first. To test for spillover effects between treatment and control, three experimental sessions were conducted that administered either only the treatment, where ability is given by the test score (one session/16 subjects), or only the control, with ability assigned by the experimenter (two sessions/36 subjects).[15] The results were essentially unchanged.

4.2 Principals' Contract Choices and Optimal Contracts

Although wage levels vary, the principals in the experiment do not seem to adjust incentives optimally. The offers seen in the experiment may differ from the theory simply because the principals do not make profit-maximizing choices. But it is also possible that agents' behavior deviates from the theory predictions, altering the optimal response for the principal. We would therefore like to find out what the principals' best contract choice is, *given* the agents' behavior in the experiment. To do so I will (a) look at the choices of the highest-performing principals and (b) use probit estimates of agent-subjects' decisions to identify the profit-maximizing contract in the experiment.

Profit-maximizing strategies. I first restrict attention to the principals with the highest total profit (within each group) and use their contract offers as a proxy for the best strategy *in the experiment*. Table 8, part (1) shows averages of all variables for the principals in the 90th percentile of total profits (three subjects and 90 observations per group). The last column (DD) lists the difference-in-difference estimates, given by the coefficient on $O \times T$ in a random-effects regression. The results show that the most successful principals did adjust incentives in the right direction, with a corresponding effect on investment.[16] Note, however, that there are many rejections and that the sum of agents' and principals' profits is smaller in C than in T, implying that the principals are not realizing the highest possible output. This indicates that we are still not looking at the principals' best possible strategies here.

In a second approach I use probit estimates for rejection and investment probabilities to predict the principal's expected profit for all observed wage offers and pick out the profit-maximizing contracts. The probit results are reported in table 12 of the appendix, and they confirm that the agents make decisions along expected lines: they reject lower expected wages more often (although less so under overconfidence, echoing result 4), and invest more

[15] In addition, principals' output levels (profits) in these sessions were raised by 30 points to give them more flexibility and to close the payoff gap between principals and agents.

[16] Total profits are added across T and C, to avoid comparing principals from different sessions. They are not necessarily maximal within each condition T and C.

Table 8: Profit-maximizing and output-maximizing contracts, first eight periods only.

	(O,C)	(O,T)	(U,C)	(U,T)	DD[a]
(1) 90^{th} percentile of total profit (top 3 subjects, by group), avg. outcomes					
Incentives $w_H - w_L$	12.3	15.6	20.8	17.4	6.67†
Avg. exp. profit principal	83.2	82.6	84.0	95.7	-12.29
Average exp. profit agent	140.2	143.3	124.6	131.9	-4.29
Avg. exp. profit, agent's beliefs	140.7	147.8	125.9	123.0	10.07*
Proportion rejected	0.08	0.08	0.33	0.73	0.17
Proportion investment	0.41	0.73	0.94	0.8	0.46*
(2) Highest expected profit (by group and condition), probit estimates					
Incentives $w_H - w_L$	30	90	20	-32	(112)
Wages $w_H/\ w_L$	60/30	90/0	45/25	18/50	(57/-55)
Exp. profit principal	92.1	106.2	100.6	105.0	(9.7)
Exp. payoff agent	137.3	121.5	120.4	125.3	(-20.7)
Exp. payoff, agent's beliefs	137.3	196.7	120.4	157.4	(22.4)
Probability of rejection (fitted)	0.05	0.01	0.24	0.16	(0.04)
Probability of investment (fitted)	0.79	0.91	0.65	0.08	(0.69)
(3) Highest net output (by group and condition), probit estimates					
Incentives $w_H - w_L$	80	90	32	45	(-3)
Exp. profit principal	61.5	93.6	56.8	7.0	(81.9)
Exp. payoff agent	176.0	143.1	202.8	255.0	(-85.1)
Exp. net output	237.5	236.7	259.5	262.0	(-3.3)
Probability of rejection (fitted)	0.00	0.00	0.02	0.00	(0.02)
Probability of investment (fitted)	0.98	0.91	0.97	0.93	(-0.03)

a: Coefficient on O×T in OLS regression in (1), simple diff.-in-diff. otherwise.
Significance levels: †:10% ∗:5% ∗∗:1%

if incentives are high. Part (2) of table 8 reports the profit-maximizing contracts given these decisions (i.e., given the choice probabilities for investment and rejection as predicted by the probit estimate). The optimal strategy involves a large adjustment to incentives and a negative wage difference for underconfident agents. Neither a higher risk of rejection nor other idiosyncrasies in the agents' response to a contract offer seem to affect the original prediction of the theory.

Result 5 (2) *The (hypothetical) profit-maximizing strategy adjusts the wage gap in response to a self-confidence bias, with negative incentives for underconfident agents.*

To summarize, all but one empirical result follow the theoretical predictions. But why do the principals not maximize profits, and in particular not set negative incentives for the underconfident agents?

Interpreting the Principals' Choices. One explanation may be nonstandard preferences. Principals might for example have a predilection for equity or fairness and feel uncomfortable "punishing" an agent for high output. There is some indication that contract offers cluster near the point $(30, 47.5)$, that is, some employers seem to attempt to equitably share their payoffs. Since the agents do not actually learn the principal's profit, this would suggest that employers have a genuine preference for equal splitting. The fairness argument cannot fully explain, however, why those principals do not choose an incentive compatible contract and then share the (higher) resulting payoff, or why the wage level varies systematically with self-confidence. A second possibility is a taste for efficiency on the part of the principals. Table 8 part (3) uses the probit estimates to find the output-maximizing contracts, equivalent to what the social planner would choose. These contracts maximize investment rates by offering strong incentives, and they show no effect of the self-confidence bias on the level of incentives. In that respect principals' choices in the experiment resemble those of the social planner. Yet output maximization also requires a very large payoff to the agent in order to minimize rejection rates, and this is not what the principals do: overall rejection rates are around 20% and 30% in U and O, respectively.

Result 6 *Net output is maximized, i.e. investment rates are high and rejection probabilities low, if effort incentives are very strong and expected wages high for all agents. The principals do choose high incentives throughout, but unlike the social planner they do not offer contracts that minimize rejection rates.*

A final possible explanation is that both sides behave optimally, but that the different-prior assumption made in the theory is not accurate. Consider for instance the simplified case with

only one, risk-neutral, agent and assume that agent and principal have originally the same prior, but the principal observes both A and a signal about a. Upon learning the principal's beliefs P, the agent would then adopt those beliefs, since P incorporates the principal's prior information and his observation of A. Similarly, if, as in the experiment, the principal observes only a noisy signal about A and a, the agent will form beliefs about the true a following Bayes' rule, and the principal in turn forms beliefs about the agent's posterior. In this case employers and employees may view the experiment as a signaling game, since the contract offer can transmit information about P. A pooling equilibrium here implies that the principal offers a contract that would just satisfy the participation constraint if the agent was unbiased and P was equal to A. In other words, the principal would adjust the wage level to the agent's bias, but never choose (meaningfully) different incentives. This suggests that what we see in the experiment may be a pooling equilibrium akin to Fang and Moscarini (2005), in which all types of agents are offered the same incentives.[17]

Appendix C discusses the simplified case with only one, risk neutral, agent and gives examples of possible (partial) pooling equilibria. It is shown that a pooling equilibrium can only exist if the belief bias for all underconfident agents is relatively small, since their bias implies relative losses to the principal. If there are some sufficiently underconfident agents (and agents are risk neutral and purely rational), the principal will prefer to offer them the same wage for high and low outcomes, reducing the wage costs from the belief bias in exchange for low effort. Many more pooling equilibria can be ruled out by a similar argument using the intuitive criterion for equilibrium selection.

That said, in a broader sense the logic of a pooling equilibrium – where off-equilibrium beliefs sustain equilibrium strategies – may still apply to the observed contract offers. Experimenting with "unusual" offers is very costly to the principals, and it may be that the majority of them so firmly expect contracts with low incentives to be rejected that they simply never choose them. This reluctance may be heightened by preferences for efficient outcomes or "fair" incentives which do not punish effort. Indeed, only 2.7% of contracts overall have a negative wage difference, offered by only 6 out of the 46 subjects in the principal role, even though those few observed offers suggest that the agents do *not* actually reject

[17]Observe that a separating equilibrium would be played only once, in the first round; afterwards we would expect no further learning (see appendix C for examples of separating equilibria). Since the principal's information is incomplete there would still be belief differences, and the predictions from the theory section earlier would apply. A pooling equilibrium, on the other hand, could persist in all rounds.

them more often (see also table 12 in Appendix C).

From this experiment it cannot be deduced if and why the principals believe that negative incentives will be rejected. As in the classical signaling model they may think that the agent will hold unfavorable beliefs after observing such an offer (i.e. negative incentives are interpreted as an indicator of very high ability), reducing the expected value of the contract below their outside option. The principals may also follow a more "behavioral" line of reasoning and anticipate e.g. that agents may retaliate for contract offers with perverse incentives, or turn down offers which induce an internal conflict between efficient and payoff-maximizing effort. These possible pathways remain a topic for further research.

5 Conclusion

This paper develops a model of a moral hazard situation in which the agent can be subject to a self-confidence bias, and then tests the main theoretical predictions in an experiment. Group and treatment effects are controlled for by using a difference-in-difference design.

In line with theoretical predictions, the expected profit for the principals in the experiment is relatively higher if the agent is overconfident, due to a lower expected wage. Conversely, the expected profit (under true probabilities) for an overconfident agent decreases, and that for an underconfident agent increases. At the same time, the agent's subjective profit, calculated using the score guess as his ability beliefs, is not affected by the self-confidence bias as long as learning is accounted for.

The experiment shows that principals are to an extent able to incorporate the belief bias into their decisions. They reduce the wage payment for overconfident agents, extracting some of the agents' profits, and respond to the demand for higher compensation from underconfident agents. In addition, it is shown that the profit-maximizing strategy in the experiment adjusts incentives to the self-confidence bias to realize additional profits. However, only the most successful principals follow this strategy, and there is consequently no significant effort distortion. This may be due to a belief on the part of the principals that negative incentives will not be accepted by the agents. These beliefs may be enforced by the fact that negative incentives run counter to standard employment contracts, and that the principals either prefer not to offer such contracts or fear that agents may reject them on account of efficiency or fairness concerns. As a consequence, the outcomes of the experiment resemble the choices of a social planner more closely than what the theory would have predicted.

Among the subjects in this experiment, both race and gender are informative signals about

an agent's self-confidence. Table 11 in appendix B shows that white male students are the most overconfident, while women and Asian and African-American subjects relatively underestimate their scores. This seems to lead to some, albeit statistically not significant, sorting among agents: there are 38.1% white and 52.4% Asian agents in O, but 31.6% and 68.4% in U, respectively. When including the experimental sessions conducted for robustness checks, there is also sorting by gender.

Self-confidence is here directly observed and matching is anonymous, so principals cannot – and need not – use race or gender to make inferences about the agents' beliefs. The correlations found here may also not extend to all task domains. Yet at equal abilities a principal in this experiment would prefer to hire men over women and white subjects over Asian or Black/African American ones, simply because they tend to be more overconfident. The importance of this issue for the labor market is clear, and differences in self-confidence as a possible source for differences in job market outcomes by gender and race warrant further research.

References

Adrian, T. and Westerfield, M. M. (2009), 'Disagreement and learning in a dynamic contracting model', *The Review of Financial Studies* **22**(10), 3873–3906.

Anderhub, V., Gächter, S. and Königstein, M. (2002), 'Efficient contracting and fair play in a simple principal-agent experiment', *Experimental Economics* **5**, 5–27.

Barber, B. M. and Odean, T. (2001), 'Boys will be boys: Gender, overconfidence, and common stock investment', *Quarterly Journal of Economics* **116**(1), 261–292.

Bengtsson, C., Persson, M. and Willenhag, P. (2005), 'Gender and overconfidence', *Economics Letters* **86**, 199–203.

Berg, J. E., Daley, L. A., Dickhaut, J. W. and O'Brien, J. (1992), Moral hazard and risk sharing: Experimental evidence, *in* M. R. Isaac, ed., 'Research in Experimental Economics', Vol. 5, JAI Press, pp. 1–34.

Beyer, S. (1990), 'Gender differences in the accuracy of self-evaluations of performance', *Journal of Personality and Social Psychology* **59**(5), 960–970.

Bull, C., Schotter, A. and Weigelt, K. (1987), 'Tournaments and piece rates: An experimental study', *Journal of Political Economy* **95**(1), 1–33.

Camerer, C. and Lovallo, D. (1999), 'Overconfidence and excess entry: An experimental approach', *American Economic Review* **89**(1), 306–318.

Clark, J. and Friesen, L. (2009), 'Overconfidence in forecasts of own performance: an experimental study', *Economic Journal* **119**, 229–251.

Cooper, D. J., Kagel, J. H., Lo, W. and Gu, Q. L. (1999), 'Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers', *American Economic Review* **89**(4), 781–804.

De la Rosa, L. E. (2007), 'Overconfidence and moral hazard', *Danish Center for Accounting and Finance (D-CAF) Working Paper* (24).

Deaux, K. and Farris, E. (1977), 'Attributing causes for one's own performance: The effects of sex, norms, and outcome', *Journal of Research in Personality* **11**, 59–72.

Eliaz, K. and Spiegler, R. (2007), 'A mechanism-design approach to speculative trade', *Econometrica* **75**(3), 875–884.

Eliaz, K. and Spiegler, R. (2008), 'Consumer optimism and price discrimination', *Theoretical Economics* **3**, 459–497.

Eliaz, K. and Spiegler, R. (2009), 'Bargaining over bets', *Games and Economic Behavior* **66**(1), 78–97.

Epstein, S. (1992), Testing principal-agent theory, *in* M. R. Isaac, ed., 'Research in Experimental Economics', Vol. 5, JAI Press, pp. 35–60.

Fang, H. and Moscarini, G. (2005), 'Morale hazard', *Journal of Monetary Economics* **52**, 749–777.

Fehr, E. and Schmidt, K. M. (2006), The economics of fairness, reciprocity and altruism - experimental evidence and new theories, *in* S.-C. Kolm and J. Mercier Ythier, eds, 'Handbook on the Economics of Giving, Reciprocity and Altruism', Vol. 1, Elsevier, chapter 8, pp. 615 – 691.

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

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

Gneezy, U. and Rustichini, A. (2004), 'Gender and competition at a young age', *American Economic Review* **94**(2), 377–381.

Grubb, M. (2009), 'Selling to overconfident consumers', *American Economic Review* **99**(5), 1770–1807.

Healy, P. J. and Moore, D. A. (2007), 'Bayesian overconfidence', *Working paper* .

Keiber, K. L. (2006), Managerial compensation contracts and overconfidence.

Keser, C. and Willinger, M. (2000), 'Principals' principles when agents' actions are hidden', *International Journal of Industrial Organization* **18**, 163–185.

Kruger, J. (1999), 'Lake wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments', *Journal of Personality and Social Psychology* **77**(2), 221–232.

Landier, A. and Thesmar, D. (2009), 'Financial contracting with optimistic entrepreneurs', *Review of Financial Studies* **22**(1), 117–150.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982), Calibration of probabilities: The state of the art to 1980, *in* D. Kahneman, P. Slovic and A. Tversky, eds, 'Judgement Under Uncertainty: Heuristics and Biases', Cambridge University Press.

Malmendier, U. and Tate, G. (2005), 'CEO overconfidence and corporate investment', *Journal of Finance* **60**(6), 2661–2700.

Malmendier, U. and Tate, G. (2008), 'Who makes acquisitions? CEO overconfidence and the market's reaction', *Journal of Financial Economics* **89**(1), 20–43.

Moore, D. A. and Cain, D. M. (2007), 'Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition', *Organizational Behavior and Human Decision Processes* **103**, 197–213.

Morris, S. (1994), 'Trade with heterogeneous prior beliefs and asymmetric information', *Econometrica* **62**(6), 1327–1347.

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

Odean, T. (1999), 'Do investors trade too much?', *American Economic Review* **89**, 1279–1298.

Santos-Pinto, L. (2008), 'Positive self-image and incentives in organisations', *The Economic Journal* **118**(531), 1315–1332.

Sautmann, A. (2007), 'Self-confidence in a principal-agent relationship', *mimeo* .

Svenson, O. (1981), 'Are we all less risky and more skillful than our fellow drivers?', *Acta Psychologica* **47**(2), 143–148.

Taylor, S. E. and Brown, J. D. (1988), 'Illusion and well-being: A social psychological perspective on mental health', *Psychologica Bulletin* **1003**(2), 193–210.

Weinstein, N. D. (1980), 'Unrealistic optimism about future life events', *Journal of Personality and Social Psychology* **39**(5), 806–820.

A Risk Aversion

The theoretical discussion of the effect of self-confidence assumed risk neutrality, even though the classical principal-agent problem is chiefly concerned with the trade off between the optimal risk allocation and optimal incentives. Two issues arise in this context: first, risk aversion on the part of the agent or the principal may change the predictions qualitatively through an interaction with self-confidence. Second, there may be a correlation between the degree of risk aversion and the level of self-confidence of the agent. In this case, the effects of varying risk aversion may confound the effect of differences in self-confidence.

For a simple discussion of risk aversion, let $P(a, e) = a + e$ be the probability of a high outcome H versus a low outcome L. Let $e \in [0, b]$ be the investment (effort) level chosen by the agent, where $a + b \in (0, 1)$ for all possible ability levels a, and $b > 0$ (to simplify the discussion, e is assumed continuous, but the results carry easily over to discrete settings).

For a given wage scheme, the agent chooses effort e to maximize

$$U(w_L, e)(1 - \hat{a} - e) + U(w_H, e)(\hat{a} + e),$$

where $\hat{a}$ is his ability expectation. If utility is separable into a wage utility and an effort cost and the agent is risk neutral, this amounts to maximizing $w_L + (\hat{a} + e)(w_H - w_L) - ce$. The agent chooses $\hat{e} = b$ if $(w_H - w_L) \geq c$, and $\hat{e} = 0$ otherwise (assuming he does invest if he is indifferent). The optimal effort level depends only on the cost parameter c and the strength of incentives $w_H - w_L$, and beliefs do not enter the agent's problem but through the participation constraint. Given his effort choice, the agent accepts the job only if it is at least as attractive as his outside option:

$$\text{(PC)} \qquad w_L + (\hat{a} + \hat{e})(w_H - w_L) - c\hat{e} \geq U$$

If the agent is unbiased, the principal maximizes her profit (and net surplus) by choosing the incentive level $w_H - w_L = H - L$ ($r = 1$). Note that this model maps into that from the theory section in the main text. Let $\theta = 0$, and rewrite the wage scheme as $r = \frac{w_H - w_L}{H - L}$ and $K = 3(w_L - rL)$. Now choosing incentives $w_H - w_L$ and the "base wage" w_L is equivalent to choosing r and K.

Suppose the agent is risk averse, so that his preferences are expressed by a concave utility function. In the most general formulation, U is a bivariate function in wage w and effort e as above; but the commonly used variants assume either that U is additively separable in wage and effort, or that it is univariate, with $U(w, e) = V(w - e)$. In the experiment, e is an investment, lending justification to the second specification. But the fact that e is an upfront

payment, while the wage is uncertain and realized only after e has been chosen, justifies a formulation like $U(w,e) = V(w) - C(e)$. While any specific choice of utility function is open to criticism, this section will focus on these two versions. The goal is in any case not an exhaustive analysis of the effects of risk aversion on the model, but an illustration of some of the possible changes to the predictions for the risk neutral case.

Note that in the additively separable case, the first derivative of U with respect to e is $-C'(e)$, the second derivative is $-C''(e)$, and the cross derivatives are zero; in the univariate case these are $-V'(w-e)$, $V''(w-e)$, and $-V''(w-e)$, respectively.

Assuming that we are at the interior of $[0,b]$, the first order condition for a utility maximizing choice of e for the agent is

$$[U(w_H,e) - U(w_L,e)] + (1-\hat{a}-e)U_e(w_L,e) + (\hat{a}+e)U_e(w_H,e) = 0$$

Differentiating the left-hand side with respect to e, w_L, and w_H gives

$$\begin{aligned}
\frac{\mathrm{d}FOC}{\mathrm{d}e} &= 2[U_e(w_H,e) - U_e(w_L,e)] + (1-\hat{a}-e)U_{ee}(w_L,e) + (\hat{a}+e)U_{ee}(w_H,e) \\
\frac{\mathrm{d}FOC}{\mathrm{d}w_H} &= U_w(w_H,e) + (\hat{a}+e)U_{ew}(w_H,e) > 0 \\
\frac{\mathrm{d}FOC}{\mathrm{d}w_L} &= -U_w(w_L,e) + (1-\hat{a}-e)U_{ew}(w_L,e)
\end{aligned}$$

The second and third terms in the first equation are negative under both utility specifications, but the first term may be positive in the univariate case. This is more likely the larger the wage gap is. As long as $\frac{\mathrm{d}FOC}{\mathrm{d}e}$ is negative, an increase in w_H leads unambiguously to higher effort. A decrease in w_L, on the other hand, can lead to lower effort, if $(1-\hat{a}-e)r_A(w_L,e) > 1$, where $r_A(w_L,e)$ denotes the coefficient of absolute risk aversion at (w_L,e). This might be the case if risk aversion is strong, self-confidence small, and the effort level fairly low. Finally, note that $\hat{a}$ enters the first order condition through the probability weights on U_e. If $U_e(w_L,e) > U_e(w_H,e)$, as is the case with the univariate utility function, $\frac{dFOC}{d\hat{a}} > 0$.

To summarize, an increase in the strength of incentives is expected to lead to higher effort, except possibly if self confidence and effort are very low and the agent is highly risk averse. An overconfident agent may, for the same incentives and utility function, exert higher effort than an underconfident agent, but at high wage gaps, the effect of an increase in incentives and the positive effect of self-confidence on the response to incentives may be reversed.

Variations in Risk Aversion Take a utility function of the second type and let U_1 and U_2 be two utility functions with $U_2 = g(U_1)$ and g concave. U_2 represents a more risk averse

agent.[18] The first order condition for an agent with utility U_2 can therefore be written as

$$\frac{g(V(w_H-e))-g(V(w_L-e))}{g'(V(w_L-e))}-(1-\hat{a}-e)V'(w_L-e)-(\hat{a}+e)V'(w_H-e)\frac{g'(V(w_H-e))}{g'(V(w_L-e))}.$$

Concavity of g implies that

$$\begin{aligned} & g'(V(w_L-e))\left(V(w_H-e)-V(w_L-e)\right) \\ < \;& g(V(w_H-e))-g(V(w_L-e))<g'(V(w_H-e))\left(V(w_H-e)-V(w_L-e)\right), \end{aligned}$$

so that at the optimal effort choice under U_1, the FOC for U_2 is bounded by

$$(1-\hat{a}-e)V'(w_L-e)\left(\frac{g'(V(w_H-e))}{g'(V(w_L-e))}-1\right)<FOC<(\hat{a}+e)V'(w_H-e)\left(1-\frac{g'(V(w_H-e))}{g'(V(w_L-e))}\right)$$

For small increases in risk aversion, the first order condition for U_2 is close to zero at the optimal choice for U_1, and continuity implies that the optimal level of e is close by. On the most general level, since $g'(V(w_H-e))<g'(V(w_L-e))$, the left bound is negative and the right one is positive, and the direction of change is not determined. But note that the bounds for the first order condition depend positively on $\hat{a}$ and e. If effort and self-confidence are low to begin with, a more risk averse agent is likely to exert less effort in response to the same incentives (assuming that $\frac{dFOC}{d\hat{a}}<0$).

In summary, for the same incentives, a more risk averse agent chooses an effort level "close" to that of the less risk averse agent. But if risk aversion is negatively correlated with self-confidence (that is, underconfident agents are more risk averse), it may reinforce the self-confidence effect: an underconfident agent cuts his efforts down even further.

Risk Aversion and Incentives How does risk aversion interfere with the result that the principal wants to shift wages to outcomes whose probability is overestimated by an agent with a belief bias? For an informal argument, suppose that both the principal and the agent are risk averse to some degree, and in a slight abuse of notation, let the agent's utility function be $U(w,e)$ and the principal's $V(\pi)$. Suppose the principal shares the agent's belief $\hat{a}$, and assume the contract (w_L,w_H) is implementing the optimal e under these conditions.

Now change the belief of the principal to $a\neq\hat{a}$. Note first that under the same contract, e still maximizes the principal's payoff. So what happens if the principal changes w_H, along with a change in w_L so that the participation constraint of the agent continues to hold?

[18] Note that this is not easily extended to the first type of utility function without making additional assumptions as to how C relates to V. One possibility would be to write $U(w,e)=cV(-e)+V(w)$, thereby assuming that utility is time separable with some form of time discounting expressed by c. Higher risk aversion is then represented again by a concave transformation of V.

For a small change $\mathrm{d}w_H$, $\mathrm{d}w_L \approx -\frac{U_w(w_H,e)(\hat{a}+e)}{U_w(w,e)(1-\hat{a}-e)}\mathrm{d}w_H$. The resulting change in profit is given by

$$\left[V'(L-w_L)(1-a-e)\frac{U_w(w_H,e)(\hat{a}+e)}{U_w(w_L,e)(1-\hat{a}-e)} - V'(H-w_H)((a+e)\right]\mathrm{d}w_H - \varepsilon,$$

with ε representing the loss from a suboptimal choice of e induced by the altered incentives (note that this loss is small, since e was chosen optimally for the original wage scheme). The term in brackets is greater than zero if

$$\frac{U_w(w_H,e)}{U_w(w_L,e)}\frac{\hat{a}+e}{1-\hat{a}-e} > \frac{V'(H-w_H)}{V'(L-w_L)}\frac{a+e}{1-a-e}$$

We clearly have that $\frac{\hat{a}+e}{1-\hat{a}-e} > \frac{a+e}{1-a-e}$. If risk aversion is moderate, the principal still benefits from increasing $w_H - w_L$, in most cases causing the effort choice to be inefficiently high. Conversely, if the agent is underconfident, the principal will lower $w_H - w_L$ and reduce effort. Thus, as in the original model, the principal chooses stronger incentives and induces a higher effort for an overconfident agent than he would if his beliefs were correct, and vice versa if the agent is underconfident. The expected utility is identically equal to the outside option for both types of agents, but lower for the overconfident agent under the principal's beliefs, and the expected wage is likely to be lower for him, too. Equivalently, the expected payoff for the principal is higher when dealing with an overconfident agent.

B Appendix: Experimental Design and Results

Table 9: Wages w_L set so that (PC) is satisfied for each $w_H - w_L$ (under risk neutrality and optimal effort choice). Self-confidence ($\hat{a}$) by one level removed from the true a, i.e. a in 5-6 implies $\hat{a}$ between 0-4 (U) or 7-8 (O).

		Underconfident $\hat{a} < a$				Overconfident $\hat{a} > a$			
$w_H - w_L$	Investment	Trivia score a				Trivia score a			
		0-4	5-6	7-8	9-10	0-4	5-6	7-8	9-10
-20	0	-	118.3	139.3	160.3	73.3	94.3	115.3	-
0	0	-	106.3	127.3	148.3	85.3	106.3	127.3	-
23	20	-	104.0	125.0	146.0	110.6	131.6	152.6	-
40	20	-	93.8	114.8	135.8	120.8	141.8	162.8	-

Table 10: Payoff-maximizing contract choices under risk neutrality.

	Underc. $\hat{a} < a$				Overc. $\hat{a} > a$			
	Trivia score a				Trivia score a			
	0-4	5-6	7-8	9-10	0-4	5-6	7-8	9-10
$w_H - w_L$		-33	-45	-61	73	52	40	
w_L		35	45	61	0	1	2	
w_H		2	0	0	73	53	42	
Payoff P		134	155	180	145	147	155	
Payoff A		80	74	64	57	69	76	

Profits and incentives. Table 9 lists the principal's expected profit as a function of incentives and the self-confidence level of the agent under the assumption that the agent's beliefs are known. Principals' payoffs are increasing in $w_H - w_L$ if the agent is overconfident and decreasing if he is underconfident (except for a nonmonotonicity at 22: there are wage gaps just below 23 in which profits are lower than at 23, because here the inefficient effort choice is not cancelled out by the gains from the belief bias). At the same ability level, overconfident employees tend to be more attractive, because their effort choice is efficient. Table 10 lists the optimal contract choices. Note that the optimal contract would always involve $w_H = 0$ for an underconfident agent and $w_L = 0$ for an overconfident one if there were no indivisibilities.

Table 11: OLS - self-confidence as a function of individual attributes.

	Coef.	(Std. Err.)	Coef.	(Std. Err.)
Female	-0.410*	(0.191)	-0.461†	(0.273)
Asian	-0.563**	(0.214)	-0.020	(0.299)
Black or African-American	-0.977†	(0.558)	-0.474	(0.795)
Hispanic or latino	0.031	(0.502)	0.057	(0.718)
Other	0.310	(0.643)	0.514	(0.919)
Ability	-0.783**	(0.069)		
Intercept	4.201**	(0.437)	-0.180	(0.291)
R^2	0.531		0.033	
No. obs.	128		128	

Significance levels: $\dagger$:10% $*$:5% $**$:1%

Self-confidence and individual attributes. Table 11 reports OLS regressions of self-confidence on individual characteristics. Observe that the coefficient on ability is large and negative: as an individual's test score increases, she becomes more likely to underestimate it. When ability is omitted, the coefficients on "Asian" and "Black or African American" are smaller and insignificant. This is because the trivia quiz is not neutral with respect to to gender and race. For example, Asian subjects had lower scores than white subjects (4.95 versus 5.63).

Agents' decisions. Table 12 reports marginal effects of a probit regression for an agent's probability of investing and of rejecting the principal's offer. In the investment probit I include both the wage gap $w_H - w_L$ and a dummy indicating wage gaps over 23, the boundary for incentive compatibility. In addition, the wage gap is interacted with group and treatment indicators, since self-confidence may interact with incentives under risk aversion (see above). As expected, an increase in incentives has a strong positive effect on investment probability.

Table 12: Agent's decisions

	(1)		**(2)**		**(3)**		**(4)**	
Incentives	0.100**	(0.014)	0.102**	(0.015)	0.007	(0.007)	0.004	(0.007)
Incent.× O	-0.068**	(0.015)	-0.066**	(0.015)	-0.050**	(0.010)	-0.046**	(0.010)
Incent.× T	-0.060**	(0.014)	-0.060**	(0.014)	-0.019*	(0.008)	-0.018*	(0.009)
Incent.× O× T	0.042*	(0.016)	0.040*	(0.016)	0.039**	(0.012)	0.037**	(0.013)
I(Incent.≥23)	0.465**	(0.169)	0.410*	(0.175)				
O	1.228**	(0.375)	1.184**	(0.376)	5.258**	(1.332)	5.161**	(1.322)
T	0.956**	(0.293)	0.951**	(0.291)	0.610	(0.635)	0.591	(0.636)
O×T	-0.561	(0.369)	-0.514	(0.369)	-3.534*	(1.445)	-3.460*	(1.435)
I(Incent.<0)			0.703	(0.535)			-0.568	(0.442)
Exp. wage					-0.025**	(0.005)	-0.025**	(0.005)
Exp.w.×O					-0.042**	(0.011)	-0.042**	(0.011)
Exp.w.×T					-0.003	(0.005)	-0.003	(0.005)
Exp.w.×O×T					0.029*	(0.012)	0.029*	(0.012)
I(Exp.w.≥100)					-0.500**	(0.177)	-0.539**	(0.179)
Intercept	-1.711**	(0.299)	-1.743**	(0.300)	2.491**	(0.614)	2.570**	(0.618)
χ^2	199.964		201.173		239.480		242.313	
No. obs.	1042		1042		1380		1380	

Significance levels: †:10% *:5% **:1%

(1) and (2): probability of investment, (3) and (4) probability of rejecting the contract (probit).

Table 13: Random effects regressions: wages and agent decisions, all periods (see table 7).

	Exp. w.	**Exp. w.***	**Incentives**	**Incentives***	**Invest.**	**Reject.**
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)
O×T	-13.773*	-12.937**	-0.614	-0.577	0.008	0.041
	(5.770)	(2.433)	(2.795)	(2.796)	(0.055)	(0.063)
O	-20.624	-12.957	9.443	10.218†	0.057	0.078**
	(13.594)	(15.489)	(5.856)	(5.428)	(0.081)	(0.029)
T	4.517*	6.543	2.490	2.580	0.018	0.014
	(2.058)	(4.229)	(2.121)	(2.160)	(0.089)	(0.069)
Intercept	145.218**	151.398**	18.623**	18.173**	0.564**	0.188**
	(12.357)	(13.706)	(4.175)	(4.194)	(0.135)	(0.037)
No. obs.	1380	1042	1380	1042	1042	1380

Significance levels: †:10% ∗:5% ∗∗:1%

The rejection decision is assumed to depend on the expected wage (calculated here using the agent's subjective success probability, but excluding the investment decision), and the wage difference between high and low outcomes (as a proxy for risk). The outside option of 100 points is captured by a dummy. Again I allow wage expectation and incentives to interact with the self-confidence bias. In line with standard predictions, the probability of rejection depends negatively on the expected wage. As indicated by the coefficient on $O \times T$, there is also a strong and significant negative effect of an overconfidence bias (conditional on expected wage), confirming result 4.

(2) and (4) include a dummy for negative incentives, to check if agents make unexpected decisions in response to negative wage differences, but the coefficients are insignificant (and their signs suggest that negative incentives are, if anything, more attractive). Note that the effect of the wage difference on the rejection probability is negative in the overconfident group, but much smaller in U and slightly positive in (U,C). This might indicate group differences in risk attitudes, with overconfident agents more risk loving. In table 8, regressions (1) and (3) are used.

C Belief Signalling

Suppose the principal holds posterior beliefs P with expectation a after observing the agent's signal about ability $\hat{a}$ and his beliefs A. Rewrite the agent's expected success probability as $(0.05+0.3I(invest)+0.2\hat{a})$, with $\hat{a} \in [0,3]$ (each integer corresponding to the four trivia score categories between 0 and 10). Letting $w_H - w_L = 23$, the lowest incentive compatible offer that theoretically satisfies the participation constraint is $\hat{w}_L = 31.95 - 4.6\hat{a}$. The principal's type is given by the deviation of his beliefs from the agent's, $x = a - \hat{a}$. At $(\hat{w}_H, \hat{w}_L)$, the principal's profit is $96.75 + 21\hat{a} + 7.2x$. One can calculate the constant wage which makes the principal indifferent between $\hat{w}$ and $(\hat{w}_L, \hat{w}_H)$ as $\hat{w} = 29.5 + 4.6x$ (taking into account that the agent will not invest at the flat wage).

Figure 2 illustrates the case where $\hat{a} = 2$ and $x = 1$, so that the agent is underconfident. The lowest incentive compatible wage pair is $(\hat{w}_L, \hat{w}_H) = (27.35, 50.35)$. The dotted lines are the agent's and principal's indifference curves in w_L-w_H-space for $\hat{a}$ (the lowest integer offer is slightly above these curves at (27,51)). The solid lines are their indifference curves through $(\hat{w}_L, \hat{w}_H)$ at the principal's beliefs $\hat{a}+x$. Note that the principal's indifference curve is discontinuous at the incentive compatibility line, since output increases as a result of higher effort, and that the agent's indifference curves intersect at the constant wage of 100/3, the flat wage which makes an agent indifferent between accepting the contract and taking the outside option ((34, 33) with only integer wages). The principal's indifference curve under P crosses the 45-degree line at $(34.1, 34.1)$.

Now suppose there is a pooling equilibrium in which the principal offers the incentive contract corresponding to the agent's beliefs, here $(27.35, 50.35)$, and consider any contract in the area ABCD. These contracts are (weakly) preferred by both the principal and the agent under $\hat{a} + x$, but they are strictly worse for a principal with belief $\hat{a}$. By the intuitive criterion, the agent should not believe that an offer in this area was made by the $\hat{a}$ type, and there cannot be a pooling equilibrium. In fact, all contracts in this area to the right of the grey dotted line are preferred by the agent for *any* belief between $\hat{a}$ and $\hat{a} + x$. By a similar argument pooling at contracts with $w_H - w_L \neq 23$ cannot be sustained. Moreover, even without using the intuitive criterion, in the figure there is a subarea of wages in ABCD which would always be (weakly) preferred by both principal and agent, in particular including the constant wage $(34, 34)$.The only restriction on this argument is that the principal's indifference curve has to lie to the right of the agent's, and this is the case only if $x \geq \frac{5}{6}$, i.e. if at least some agents are sufficiently underconfident.

This game has many equilibria, some of which involve partial pooling. As an example I will

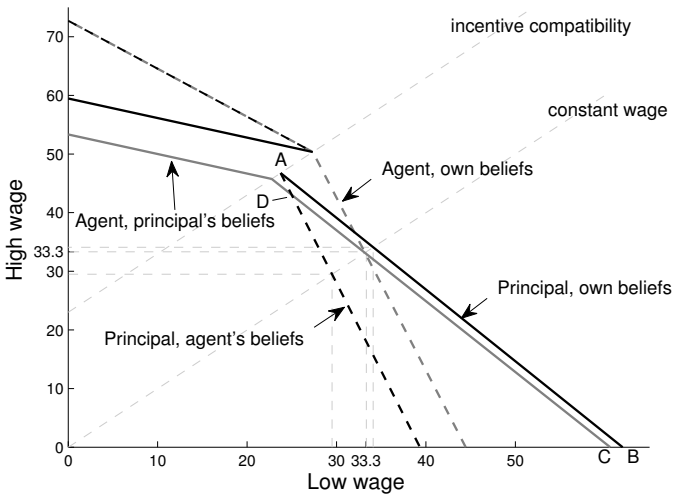


Figure 2: Indifference curves in w_L-w_H-space.

characterize a full separating equilibrium.

For any two principal types $a' = \hat{a} + x'$ and $a'' = \hat{a} + x''$, the contracts chosen in equilibrium must be such that none has an incentive to mimic the other. Assume that $w_H - w_L \geq 23$; it must hold that

$$3(0.35 + 0.2a_1)w_H^1 + 3(0.65 - 0.2a_1)w_L^1 \geq 3(0.35 + 0.2a_1)w_H^2 + 3(0.65 - 0.2a_1)w_L^2$$

for a_1 equal to a' and a'' and (w_L^i, w_H^i) the contract associated with a^i, $i = 1, 2$. For this to hold w_L and w_H must be strictly increasing and decreasing in a, respectively, i.e. incentives fall with ability. A sufficient condition is for example that $\frac{\mathrm{d}w_L}{a} < 0$ and $\frac{\mathrm{d}w_H}{\mathrm{d}w_L} = -\frac{0.65-0.2a}{0.35+0.2a}$.

D Instructions

This section provides the instructions for a typical experimental session in chronological order. All instructions were read out loud and displayed on the computer screen. Subjects also received a paper version of the instructions. The instructions here are for a session where the treatment is administered first. Square brackets [] indicate a new "page" on the computer screen.

Stage 1: Trivia Quiz

[screen 1]

Welcome!

Thank you for participating in this study. Please read and sign the letter of consent.

Switch off your cell phones and other electronic devices, and put your personal belongings away.

Please do not speak with each other during the experiment. If you have questions at any point, please raise your hand and wait until I come over, to speak to me in private.

During this experiment you will have opportunities to earn or spend points. All your decisions and earnings will be recorded under the computer number (labxx). At the end of the experiment, the number of points associated with each computer number will be converted into US dollars.

For every two points that you earn you will receive $0.01.

At the end of the experiment I will enter your earnings in a payment form, which you can then redeem at the bursar's office at 25 West 4th Street (petty cash).

[screen 2]

Stage 1

The experiment will be conducted in three stages. Instructions will be given along the way.

This is Stage 1. You will now be asked 10 trivia questions.

With each correct answer you earn 20 points.

Please click "Start" to begin the trivia quiz.

[quiz questions - 1 per screen, multiple choice answers (correct answer in italics)]

1. What is South America's highest peak? Mt. Simón Bolívar; Mt. Ancohuma; *Mt. Aconcagua*; Mt. Pumasillo

2. John Adams and Thomas Jefferson, the 2nd and 3rd Presidents of the United States, both died on what day? October 30, 1801; March 4, 1809; January 25, 1840 *July 4, 1826*

3. The Italian village of Pompeii was destroyed in 79 AD by what type of natural disaster? Flood; Earthquake; *Volcano*; Wildfires

4. Jim Morrison (lead singer of The Doors), Elvis Presley, and Jimi Hendrix all died from what? Heart attack; Car accident; Suicide; *Drug overdose*

5. Which team won the first Super Bowl? *Green Bay Packers*; Baltimore Colts; New York Jets; Kansas City Chiefs

6. What actor holds the record for having been nominated most frequently for the "Best Actor" Academy Award (9 times)? Dustin Hoffman; *Spencer Tracy*; Paul Newman; Jack Nicholson

7. What is the largest species of whale? Sperm whale; *Blue whale*; Orca (Killer whale); Bowhead whale

8. Laudanum is a form of what drug? Chloral hydrate; Valium; *Opium*; Mescaline

9. Who was the first African American to win an Academy Award for best actress? Angela Bassett; Whoopi Goldberg; Jennifer Hudson; *Halle Berry*

10. What is the capital city of Germany? Bonn; Frankfurt; *Berlin*; Amsterdam

[score guess]

You answered all 10 questions.

Now please guess as accurately as possible how many of them you think you answered correctly.

You earn 100 points minus the square of the difference between your guess and your actual score. In other words, the further your guess is away from your actual score, the less points you earn, and the decrease is quadratic.

For example, if you have 7 right answers, but your guess is 4, you earn 100-9=91 points. If you guess 8, you earn 100-1 = 99 points.

Stage 2: Contracting Rounds

[screen 1]

Stage 2

You will now enter Stage 2 of the experiment. Stage 2 consists of two parts, each with 2 trial rounds and 15 paying rounds.

Imagine that you are participating in a job market. Based on the results of Stage 1, you will be divided into two groups (Group 1 and Group 2), and within each group some of you are going to be employers, and some of you are going to be employees.

Your role will be the same for the entire experiment. Employers and employees will only interact through the computer and never be in direct contact.

At the beginning of each round, each employer is paired up randomly with one employee from the same group. The employer is going to make a job offer, that is, (s)he offers the employee a certain amount of points for completing a job that consists of three "tasks". Each task has two possible outcomes, H(igh) and L(ow). If the outcome is H, then the employee makes a high number of points for the employer. If the outcome is L, then the employer earns less.

The employee does not have to accept the job offer. If (s)he does accept the offer, (s)he can increase the probability of getting high outcomes in the three tasks by investing some of his/her own points.

To show you how this works, we will go through a typical round, first from the employee's, then from the employer's perspective.

[screen 2]

Stage 2

As an *employee* you start each round with a budget of20 points. You first receive a job offer from your employer of that round. The employer makes a job offer to the employee by choosing two "wages": what (s)he pays for each task in which the outcome is H, and what (s)he pays for each task with outcome L.

You do not have to accept the offer. If you *reject* it, you receive an additional 100 points and the round ends.

If you want to *accept* the offer, you also have to decide whether you want to invest your budget in the job or not. By investing, you increase the probability of receiving a high outcome in each task.

Once you have accepted the offer and decided on your investment, the computer will determine the outcome of the job (that is, whether each of the three tasks had high or low outcomes).

Importantly, the probability of a high outcome depends not only on the investment, but also on your trivia score from Stage 1. Your trivia score decides how likely the high outcome is for each task, regardless of how much you invested in a given round.

If two employees make the same investment decision, the person with the higher score is more likely to get high outcomes.

Once the outcomes of the three tasks have been determined, you will receive the appropriate wage for each, and it will be added to your total points. If you did not invest, the 20 budget points are added to your total as well.

[Subjects are shown the employee decision screen and have time to familiarize themselves with it.]

[Four questions to test understanding; subjects can go back to the employee screen to answer them.]

1. Suppose an employee thinks his score is 5. What is the probability that the outcome of each task is H if he *does not* invest? (Answer: 0.25)

2. Suppose an employee thinks her score is 9. What is the probability that the outcome of each task is H if she *does* invest? (Answer: 0.95)

3. Suppose this employee does not invest. How often, approximately, can she expect that her outcome will be H? (round to whole number) (Answer: 2)

4. If her wages are X=200 and Y=100, how much can she expect to earn (rounded to whole number)? (Note that these wages are an example - they are not actually possible in this experiment) (Answer: 520)

[screen 3]

Stage 2

As an *employer* you make a job offer to the employee by choosing two "wages": what you pay for each task in which the outcome is H, and what you pay for each task with outcome L.

If the employee *rejects* the offer, you earn 0 points and the round ends.

If the employee *accepts* the offer, your earnings depend on the outcomes of the 3 tasks. In turn, the probability of outcome H in each task depends on the employee's ¡N— !text: 0="trivia"; 1="assigned";¿ score and his/her investment.

To make it easier for you to choose the wage, you have access to the employee screen where you can test what your own wage offer looks like. Moreover, you can see the trivia scores and guesses of

the employees in your group (although you will not get information about individual employees). Once your employee has decided, the computer will determine the outcomes of the three tasks. You will receive the earnings associated with them, minus the wage that you offered your employee for each.

[All subjects now see the employer screen.]

[screen 4]

After the employer and the employee made their decisions, they learn how many of the three tasks had outcome H or L, and how much they earned in this round; unless of course the offer was rejected (in which case the employee receives 120 points, the employer earns nothing).

This ends the round, and a new round begins.

Please wait until the experimenter starts Part 1 of Stage 2.

[screen 5: group and role assignment]

You have been assigned to
Group [1/2]
For the remainder of this experiment, you are an
[**Employer/Employee**].

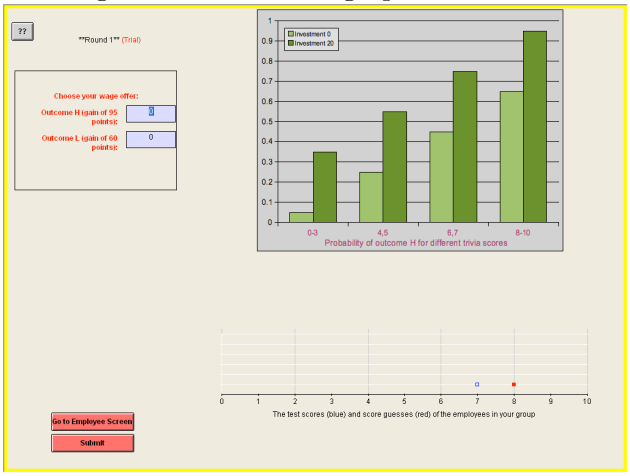
When you are ready, you will enter Part 1 of Stage 2, consisting of 2 trial rounds and 15 paying rounds.

In each round you will be randomly matched with one employee from your group.

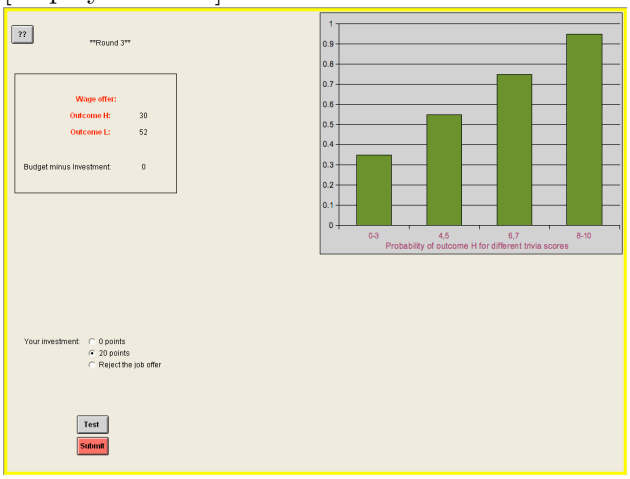
Remember that you can use the help button for explanations.

Please click when you are ready to start.

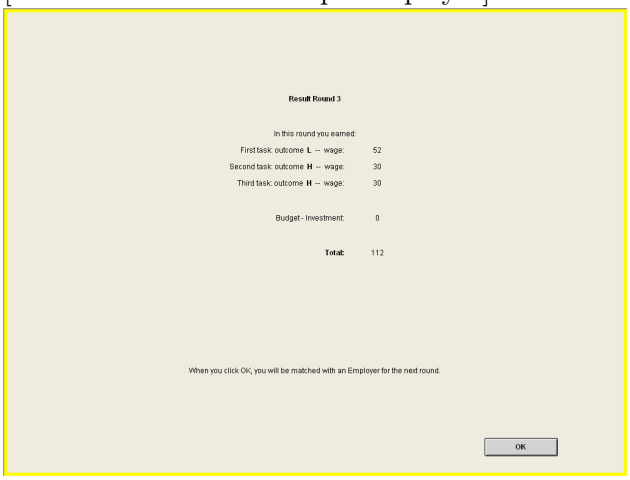
[The 17 first rounds start. Below is the employer screen; note that in the experiment there are at least 4 agents shown in the graph of test scores and score guesses]



[employee screen]



[resolution screen - example employee]



[After 17 rounds, Part 2 of Stage 2 starts]

Stage 2 - Part 2

You have completed the first part of Stage 2. In Part 2 there will be again 2 trial rounds and 15 paying rounds. You remain an [employer/employee] in group [1/2].

The only thing that changes from Part 1 is the score that determines the probability of outcome H. In Part 1, it was the trivia score from Stage 1.

In Part 2 it will be a number between 0 and 10 selected at random for each employee.

As an employee, you won't learn your own score, and as an employer you will not know your employee's score. But on the employee screen you can see a graphic display of the scores of all the employees in your group.

Please click when you are ready to start.

Questionnaire

[there were some questions that are not used in the analysis. they are omitted here.]

[personal characteristics (multiple choice)]

- My gender
- Are you Spanish/Hispanic/Latino?
- Race (check all that apply): (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, White)

[final screen - payoffs]

Thank you! You have completed the experiment.

You earned [points from Stage 1, points from Stage 2, total points]

Total earnings in US$ (incl. participation base compensation) [total earnings]