TI 2012-034/1
Tinbergen Institute Discussion Paper

The Role of Performance Appraisals in Motivating Employees

Jurjen J.A. Kamphorst
Otto H. Swank

Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

The role of performance appraisals in motivating employees

Jurjen J.A. Kamphorst[1]
Erasmus School of Economics
and Tinbergen Institute
kamphorst@ese.eur.nl

Otto H.Swank
Erasmus School of Economics
and Tinbergen Institute
swank@ese.eur.nl

Abstract:

In many organizations, reward decisions depend on subjective performance evaluations. However, evaluating an employee's performance is often difficult. In this paper, we develop a model in which the employee is uncertain about his own performance and about the manager's ability to assess him. The manager gives an employee a performance appraisal with a view of affecting the employee's self perception, and the employee's perception of the manager's ability to assess performance. We examine how performance appraisals affect the employee's future performance. The predictions of our model are consistent with various empirical findings. These comprise (i) the observation that managers tend to give positive appraisals, (ii) the finding that on average positive appraisals motivate more than negative appraisals, and (iii) the observation that the effects of appraisals depend on the employee's perception of the manager's ability to assess performance accurately.

JEL codes: M52, M54, D82, D83

Keywords: Subjective Performance Appraisal, Credibility, Cheap Talk

[1]We thank Jasmijn Bol, Francis Bloch, Josse Delfgaauw, Kris de Jaegher, Botond Koszegi, Victor Maas, Menno Middeldorp, Sander Onderstal, Ronald Peeters, Canice Prendergast, Arno Riedl, Stephanie Rosenkranz, Roland Strausz, Roland van Weelder, Utz Weitzel, Bastian Westbrock, as well as participants of the Workshop on Economic Theory and Game Theory (december 2011) and the ESE Brown Bag seminar for their helpful comments.

1 Introduction

Much of the economics literature on rewards and incentives focuses on the problem of designing compensation schemes in different environments. A rich literature shows how, for example, the employee's risk aversion, the extent to which output is verifiable, the existence of multiple tasks, or the presence of team production, should affect compensation schemes. The existing economics literature pays relatively little attention to how an employee's performance is measured. Generally, it is taken for granted that some (imperfect) measures of performance are available. In practice, it is often the case that employees receive annual performance evaluations from their supervisors. These evaluations usually form the basis for setting bonuses or promotions.

There is a diverse business literature on performance evaluations. One strand in this literature examines what kind of evaluations supervisors give. One well-known finding is that many supervisors tend to give (too) positive assessments.[2] This phenomenon is known as the leniency bias. Another finding is that some managers tend to compress performance ratings.[3] This is known as the centrality bias [Motawidlo and Borman (1977)].

A second strand of the business literature shows how performance evaluations affect employees' future performances [see, for example, Balcazar et al. (1986), Kluger and DeNisi (1996), and Alvero et al. (2001)]. Positive evaluations are generally found to motivate employees. Negative evaluations, on the other hand, sometimes improve performance and sometimes deteriorate it. Steelman and Rutkowski (2004) show that the credibility of the supervisor affects the sign and the size of the effect of negative feedback on an employee's future performance. More generally, there is ample evidence that employees tend to reject feedback that is inconsistent with their own beliefs.[4]

The main objective of this paper is to develop a model that explains the two

[2] Medoff and Abraham (1980) report that of 7,000 performance ratings 95 % were in just two categories: Good and Outstanding. See also Prendergast (1999) and Jawahar and Williams (1997).

[3] Moers (2005), for example, finds that performance ratings on subjective dimensions are closer to the median rating than performance ratings on objective dimensions.

[4] Many scholars emphasize the importance of the supervisor's credibility (see, e.g., Lawler (1971), Meyer (1975), Early (1986) and Longenecker (1997). Gibbs et al. (2004) formulated the credibility issue as follows: "If subordinates do not trust their evaluators to make informed and unbiased performance assessments, then the result could be employee frustration, demotivation, and turnover".

biases in performance ratings, and at the same time explains how performance appraisals affect employees' future performances. Although this paper does not address the problem of the design of an optimal incentives scheme, it does illustrate that a better understanding in the performance appraisal process is likely to contribute to a better understanding of the working of incentive schemes.

The model we develop has four key characteristics. First, at the beginning of the game, both the supervisor and the employee form a perception of the employee's past performance. We model this formation of perceptions by assuming that the two agents receive private signals.[5] Second, we assume that supervisors differ in their abilities to assess the employee's performance correctly. The motivation of this assumption is that supervisors have been found to vary in their beliefs about their skills to appraise their subordinates [see, for example, Napier and Latham (1986), and Tziner et al. (2001)]. Third, we assume an environment where employees are rewarded on the basis of their performance evaluations. So, appraisals are linked to rewards.[6] Finally, the employee's ability and his effort are complements. The implication of this last characteristic is that the more the employee is confident about his ability, the more effort he exerts.

We derive several results. Our first set of results pertains to a situation where the employee knows his own ability. In this extreme situation, performance appraisals only provide information about the supervisor's ability to assess the employee's ability correctly. A supervisor who gives an incorrect assessment of an employee's performance loses credibility. A direct implication is that a supervisor who knows an employee's performance has no incentive to rate it incorrectly. This would only damage his credibility. The employee would doubt whether his future performance would be correctly assessed. For a supervisor who does not observe an employee's performance three forces are at work. First, she has an incentive to give an appraisal that is most likely to be consistent with the employee's perception. This force may

[5] In our model, the assumption that the employee receives a signal about his performance amounts to assuming that the employee has imperfect knowledge about his own abilities. The psychological literature offers a huge body of evidence that this assumption is valid (see, among others, Sedikes and Strube, 1995; Klar et al., 1996; Baumeister, 1998; Kruger, 1999; and Ackerman et al., 2002).

[6] This means that we assume that the supervisor is able to commit herself to a compensation scheme based on subjective performance evaluations. In a situation where the supervisor is the residual claimer, this assumption is strong, as ex post she will have an incentive to pretend that the employee did a poor job. However, in many situations supervisors are not residual claimers.

explain the centrality bias of performance evaluations. Second, as the employee's ability and his effort are complements, it is more important for the supervisor that her evaluation is correct in case the employee is more able. This force leads to a positive bias in performance appraisals. Finally, a less able supervisor wants to come across as able. This gives her an incentive to give an appraisal that able supervisors relatively frequently give. We show that this force tends to dampen the total effect of the first two forces.

The second set of results are derived from the version of the model in which we relax the assumption that the employee knows his own ability and thereby his past performance. In this setting, apart from the incentives discussed above, a supervisor has an incentive to give positive appraisals. The reason for this incentive is that the employee's effort is an increasing function of his belief about his ability. This result explains the leniency bias often found in the empirical literature on performance rating. The idea that supervisors give positive appraisals to boost employees' perceptions of their abilities to make them work harder is not novel. Bénabou and Tirole (2003), for instance, show that giving a challenging task to an employee signals confidence and thereby motivates. New is that simple cheap-talk messages may motivate employees.[7] Essential in our model is that apart from boosting an employee's confidence, the supervisor wants to show that she is capable of assessing the employee's performance. This weakens her incentive to give positive appraisals when the employee is perceived to perform poorly. If either the supervisor were always capable of assessing the employees' performance correctly or the employee had absolutely no clue about his ability, the supervisor's incentive to come across as able would vanish. In such a situation, she would always provide positive feedback. As a result, feedback would contain no information about an employee's actual performance.

Apart from the business literature on performance appraisals, this paper is most closely related to the literature on subjective performance appraisals [important early papers are Bull (1987) and Gibbs et al. (1994); see Prendergast (1999) and Bol (2009) for reviews of the literature]. Key notion in this literature is that most people

[7] Crutzen, Swank and Visser (2007) show that *comparative* cheap-talk messages may reveal meaningful information about employees' performance levels. However, they also show that supervisors tend to abstain from differentiating among employees.

do not work in jobs where all aspects of an employee's performance are verifiable. Repeated interaction may allow for an implicit contract in which rewards are based on unverifiable information. In practice, the problem with incentive contracts is not only that an employee's performance is not verifiable for a third party. Often, it is difficult to assess an employee's performance in the first place. Then, measuring performance requires expertise. Moreover, disagreement about the true performance may exist. Our paper does not focus on the determination of the optimal contract in these situations. Rather, it tries to shed light on the communication between supervisors and employees given a specific incentive scheme. We believe that this approach makes sense, because the persons who are responsible for performance appraisals do not always have a say in the design of the compensation scheme.

This paper is also related to Prendergast (1993) who shows that when firms use subjective performance evaluations an employee may have an incentive to conform to the opinion of his supervisor. In our model, however, it is the supervisor who has an incentive to guess the worker's opinion about his performance. By guessing correctly, the supervisor signals that she can assess the worker's future performance accurately.

Finally our paper is related to Prendergast and Topel (1993, 1996), who also start from the premise that a manager's appraisal is not fully trustworthy. In their model, the performance appraisal may deviate from the true performance because the manager is biased with respect to the employee. In our model, any deviation of the performance appraisal from the true performance level is because the manager lacks the necessary expertise to judge the worker's performance.

2 The Feedback Model

Our model describes the interaction between a worker (he) and his supervisor, the manager (she). The worker faces two kinds of uncertainties. First, he is uncertain about how his effort affects his performance, and second, he is uncertain about the manager's ability to assess his performance correctly. The worker chooses effort, e, to produce output y. The extent to which effort translates into output depends on the worker's ability, a: specifically $y = ae$. There are two types of workers, $a \in \{l, h\}$. The prior probability that the worker is of the high ability type equals

$\alpha : \Pr(a = h) = \alpha$ and $\Pr(a = l) = (1 - \alpha)$. The worker does not know his type. However, he receives a private signal about a, $s \in \{l, h\}$. With probability ζ, the worker's signal is fully informative, $s = a$. With probability $(1 - \zeta)$, s does not contain any information about a. Denote by η the probability that the worker believes that $a = h$ after he has received signal $s = h : \eta = \Pr(a = h|s = h) = \zeta + (1 - \zeta)\alpha$ and $1 - \eta = \Pr(a = l|s = h)$. Likewise denote by λ the probability that the worker believes that $a = l$ after he has received signal $a = l : \lambda = \Pr(a = l|s = l) = \zeta + (1 - \zeta)(1 - \alpha)$ and $1 - \lambda = \Pr(a = h|s = l)$.

There are two types of managers: $t = \{b, g\}$, with $\Pr(t = g) = \rho$. A good manager, $t = g$, observes both a and y. A bad manager, $t = b$, observes neither a nor y. The manager knows her type, but the worker does not know the manager's type. The manager takes two actions. First, before the agent chooses effort, the manager sends a message, $m \in \{l, h\}$, about her perception of the worker's ability. We assume a natural language in the sense that by sending $m = l$ the manager does not decrease the probability that the worker believes that $a = l$, while by sending $m = h$ the manager does not decrease the probability that the worker believes that $a = h$. Let $\hat{\alpha}(s, m)$ denote the probability that the worker believes $a = h$, conditional on s and m. The assumption of a natural language implies that $\hat{\alpha}(s, h) \geq \hat{\alpha}(s, l)$. We sometimes refer to the probability $\hat{\alpha}(s, m)$ as the worker's self-confidence. Second, after the worker has chosen effort, the manager assesses the output that the worker has produced. The key feature of our model is that the manager's feedback may contain information both about the worker's ability and about her own ability to assess the worker's performance correctly.

The worker's payoff equals $y - \frac{1}{2}\gamma e^2$ if $t = g$, and equals $\hat{y} - \frac{1}{2}\gamma e^2$ if $t = b$, where $\hat{y}$ is the worker's expected output, conditional on the information a bad manager possesses. A crucial assumption is that $\hat{y}$ is independent of the effort actually exerted by the worker, whereas y is increasing in effort. The manager is assumed to aim at maximizing the (expected) output the worker produces.

The timing of the game is as follows.

- Nature draws a and t. The manager observes t. A good manager also observes a.
- The worker receives a signal, s, about a.

- The manager sends a message, m, to the worker about a.
- The worker updates his beliefs about a and t.
- The worker chooses effort, e, leading to output $y = ae$
- A good manager observes y and pays the worker y. A bad manager does not observe y and pays the worker $\hat{y}$.
- Payoffs are realized.

All priors are common knowledge, as is the probability ζ.

Our model is a dynamic game with incomplete information. The effort strategy of the worker maps his signal about his ability and the message he received from the manager into an effort level $e(s, m) \in [0, \infty)$. A good manager's feedback strategy maps the worker's ability into a message m: $\mu_g(a) \in [0, 1]$ where $\mu_g(a)$ denotes the likelihood that a good manager sends message $m = h$, conditional on a. A bad manager's feedback strategy denotes the likelihood $\mu_b \in [0, 1]$ with which a bad manager chooses message $m = h$. Denote by $\hat{\rho}\left(s, m; \mu_g^*(a), \mu_b^*\right)$ the worker's posterior belief that the manager is good, $t = g$, conditional on s and m, and given the equilibrium feedback strategies, $\mu_g^*(a)$ and μ_b^*. We identify Perfect Bayesian Equilibria in which (i) given the posterior probabilities $[\hat{\alpha}(s, m)$ and $\hat{\rho}\left(s, m; \mu_g^*(a), \mu_b^*\right)]$ and feedback strategies of the two types of managers, the worker's effort choice maximizes his expected payoff; (ii) given the posterior probabilities and anticipating the worker's effort decision, the feedback strategy of each type of manager maximizes her expected payoff; and (iii) posteriors are based on Bayes' rule whenever possible. In our model, m is cheap talk. It is well-known that in models with cheap talk, babbling may occur. Throughout, our focus will be on equilibria without babbling.

Let us finish this section with three comments. First, an important assumption underlying our model is that the manager's ability to observe a is perfectly correlated with her ability to observe y (and so the ability to base the worker's reward on y). We could have avoided this assumption by adding a probation period to the model. If at the end of the probation period, only good managers were able to assess output and to infer ability, we would have a similar model as described above. Second, we assume that the manager is committed to reward the worker on the basis of the

output that has been produced. This is somewhat restrictive when the manager is a residual claimant. Whenever the manager is not a residual claimant, say the typical middle manager, we expect that she has few, if any, material incentives to avoid paying his worker the proper performance wage. Third, to drive home our results with respect to how the manager's feedback influences the worker's motivation in the simplest way, we have assumed a very simple production structure.

3 Equilibria

3.1 The worker knows his own ability: $\zeta = 1$

We start the analysis with examining the case that the worker knows his own ability, $\zeta = 1$ (implying $\eta = \lambda = 1$). In the resulting game, the worker tries to infer information about the manager's type, and the manager tries to convince the worker that she is good.

First consider the worker's effort decision. The worker anticipates that his effort only increases his reward in case the manager is good. His effort results from maximizing $\hat{\rho}\left(s, m; \mu_g^*(a), \mu_b^*\right) E(a|s,m)\, e - \frac{1}{2}\gamma e^2$ with respect to e, yielding

$$e^*(s,m) = \frac{\hat{\rho}\left(s, m; \mu_g^*(a), \mu_b^*\right) \{\hat{\alpha}(s,m)\, h + [1 - \hat{\alpha}(s,m)]\, l\}}{\gamma} \tag{1}$$

where $\hat{\alpha}(h,m) = 1$ and $\hat{\alpha}(l,m) = 0$, because $\zeta = 1$. Equation (1) shows that the worker's effort is an increasing function of the posterior that the manager is good.

Equation (1) implies that the manager wants the worker to believe that he can correctly assess the worker's ability. The assumption of a natural language implies that it is a dominant strategy for a good manager to reveal her perception of the worker, $\mu_g^*(h) = 1$ and $\mu_g^*(l) = 0$. Given the equilibrium strategy of a good manager,

Bayes' rule implies the following posterior beliefs[8]

$$\begin{aligned}
\hat{\rho}\left(h,l;\mu_g^*(a),\mu_b^*\right) &= 0 \\
\hat{\rho}\left(l,h;\mu_g^*(a),\mu_b^*\right) &= 0 \\
\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) &= \frac{\rho}{\rho+(1-\rho)(1-\mu_b)} \\
\hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right) &= \frac{\rho}{\rho+(1-\rho)\mu_b}
\end{aligned} \tag{2}$$

The posteriors show that guessing incorrectly ruins a manager's reputation, while guessing correctly improves it. The extent to which a correct message improves the manager's reputation depends on the probability with which a dumb manager sends that message. For instance, if a dumb manager rarely sends $m = l$ (μ_b close to 1), then $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right)$ is close to 1. More in particular, the higher is μ_b, the lower is $\hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right)$ and the higher is $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right)$. Note that if the manager is more likely to be competent, the strategy of a bad manager (μ_b) is less important.

Having established how much effort the worker exerts in equilibrium, the dominant strategy of a $t = g$ manager, and the posteriors, there remains to determine the optimal response of a bad manager. Using (1) and the posteriors, it is easy to see that always choosing $m = l$ ($\mu_b = 0$) is an optimal response of a bad manager if

$$(1-\alpha)\rho l^2 > \alpha h^2 \tag{3}$$

The left-hand side of (3) denotes the expected output when a worker receives $m = l$, given $\mu_b = 0$. The right-hand side gives the expected output when the worker receives $m = h$, given $\mu_b = 0$.[9] In an equilibrium with $\mu_b = 0$, the worker does not infer information from $m = l$, $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) = \rho$, but learns the manager's type if $m = h$: $\hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right) = 1$. One can show that this equilibrium requires that α is sufficiently smaller than $\frac{1}{2}$. The reason is twofold. First, a manager who sends the right message (so $m = s$) gains more credibility if $m = h$ than if $m = l$. So, relative to $m = l$, $m = h$ boosts the worker's confidence in the manager. We

[8] In the special case where $\mu_b = 0$ and $\mu_b = 1$, $\hat{\rho}\left(l,h;\mu_g^*(a),\mu_b^*\right)$ and $\hat{\rho}\left(h,l;\mu_g^*(a),\mu_b^*\right)$, respectively are off the equilibrium path. We assume that, also in this case, the 'wrong' feedback message is attributed to a Bad manager rather than to a Good manager.

[9] Recall that $e(h,l) = 0$

refer to this effect as the *confidence in manager*. Second, a high ability worker is more productive than a low ability worker. As a result, it is more productive to guess correctly if the worker is of the high ability type than if he is of the low ability type. This gives an incentive to a $t = b$ manager to send $m = h$. We call this the *productivity effect*. The only reason of a bad manager for sending $m = l$ is that it leads to a higher probability of being correct. We call this the playing the odds effect. The *playing the odds* effect works in favor of giving feedback to the most common worker type: $m = h$ if $\alpha > \frac{1}{2}$ and $m = l$ if $\alpha < \frac{1}{2}$.

Given (1), always sending $m = h$ ($\mu_b = 1$) is an optimal response of a bad manager if

$$\alpha\rho h^2 > (1-\alpha)\, l^2 \tag{4}$$

The left-hand side of (4) gives the expected output when a worker receives $m = h$, given $\mu_b = 1$. The right-hand side gives the expected output when the worker receives $m = l$, given $\mu_b = 1$. If in equilibrium $\mu_b = 1$, then the worker does not infer information from $m = h$, so that $\hat{\rho}\left(h, h; \mu_g^*(a), \mu_b^*\right) = \rho$, but learns the manager's type in case $m = l$, $\hat{\rho}\left(l, l; \mu_g^*(a), \mu_b^*\right) = 1$. An equilibrium in which $\mu_b = 1$ exists for a wider range of parameters than an equilibrium in which $\mu_b = 0$. The reason is that the productivity effect makes sending $m = h$ more attractive for a dumb manager than sending $m = l$. As a result, the playing the odds effect must be large to compensate the productivity effect.

Note that for $\mu_b = 0$ and $\mu_b = 1$ the confidence in manager effects are opposites: to boost the worker's confidence, a manager has an incentive to send $m = h$ if $\mu_b = 0$, but has an incentive to send $m = l$ if $\mu_b = 1$. The confidence in manager effect is responsible for the existence of an equilibrium in which the bad manager mixes. Such an equilibrium exists if both (3) and (4) are violated. A bad manager is indifferent between sending $m = l$ and sending $m = h$ if

$$\begin{aligned} (1-\alpha)\,\hat{\rho}\left(l, l; \mu_g^*(a), \mu_b^*\right) l^2 &= \alpha\hat{\rho}\left(h, h; \mu_g^*(a), \mu_b^*\right) h^2, \qquad \text{so that} \\ \mu_b &= \frac{\alpha h^2 - \rho l^2 + \alpha\rho l^2}{(1-\rho)\left(l^2 + \alpha h^2 - \alpha l^2\right)} \end{aligned} \tag{5}$$

One can check that μ_b is increasing in h and decreasing in l. These comparative static results reflect the productivity effect. The benefit of guessing right is higher

if the worker is of the high ability type than if the worker is of the low ability type. Moreover, μ_b is increasing in α. This is the playing the odds effect. The higher is the probability that the worker is of the higher ability type, the higher is the probability that by sending $m = h$ a bad manager guesses correctly. Finally, μ_b is increasing in ρ if and only if $\alpha > \frac{l^2}{l^2+h^2}$. The intuition for this last result is as follows. At $\alpha = \frac{l^2}{l^2+h^2}$ we have that $\mu_b = \frac{1}{2}$. Then, the confidence in manager effect favors neither feedback. For other α, we have that $\mu_b \neq \frac{1}{2}$. Then, the confidence in manager effect pushes μ_b (weakly) towards $\frac{1}{2}$. An increase in ρ dampens the confidence in manager effect: if ρ is high, the posterior beliefs of the worker hardly depends on μ_b. Consequently, an increase in ρ reduces the costs of a deviation of μ_b deviate from $\frac{1}{2}$. Hence, the larger is ρ, the lower is μ_b if $\alpha < \frac{l^2}{l^2+h^2}$, and the higher is μ_b if $\alpha > \frac{l^2}{l^2+h^2}$.

The next proposition summarizes the discussion above.

Proposition 1 *Suppose that in the feedback model $\zeta = 1$ and $\rho \in (0,1)$. Then, on the basis of the $t = b$ manager's strategy three equilibria can be distinguished:*
(i) if $(1-\alpha)\rho l^2 \geq \alpha h^2$, an equilibrium in pure strategies exists in which $\mu_b = 0$;
(ii) if $\alpha h^2 \geq \frac{(1-\alpha)l^2}{\rho}$, an equilibrium in pure strategies exists in which $\mu_b = 1$
(iii) if $\frac{(1-\alpha)l^2}{\rho} > ah^2 > (1-\alpha)\rho l^2$, an equilibrium exists in which μ_b is given by (5).
In equilibria $(i-iii)$, the worker's effort is given by (1), the $t = g$ manager's strategy is $\mu_g^*(h) = 1$ and $\mu_g^*(l) = 0$, and the posteriors are given by (2). The equilibrium probability with which a $t = b$ manager chooses $m = h$ is non-increasing in l, and non-decreasing in h and α.

3.2 The Worker Does Not Know His Ability, $\zeta = 0$

In case the worker's signal does not contain any information about his ability, the worker is not able to infer information about the manager's type from her feedback. Of course, in equilibrium the manager anticipates this. The main implication is that the manager needs not to consider how her feedback impacts on the worker's perception of her type. To put it differently, if $\zeta = 0$, the confidence in manager effect, the playing the odds effect and the productivity effect do not longer play a role.

Potentially, there is a new effect of feedback, however. If $\zeta < 1$, the manager has private information about the worker's ability. Consequently, feedback may affect

the worker's perception of his ability. We call this the *self-confidence* effect. We now argue that if $\zeta = 0$, in equilibrium, a manager always gives positive feedback: $\mu_g(h) = \mu_g(l) = \mu_b = 1$. Recall that the manager gives feedback with an eye to encouraging the worker to expend more effort. Our assumption of a natural language implies that providing negative feedback never induces a worker to have a more positive perception of his ability than positive feedback. As effort is increasing in the worker's belief about his ability, it is never optimal for the manager to send negative feedback. Of course, in equilibrium, the worker sees through the manager's attempt to boost his perception of his ability. As a result, the manager's positive feedback has no effect.

Proposition 2 *Suppose that in the feedback model $\zeta = 0$. Then, the unique equilibrium is a pooling one in which:*
(i) $\mu_g^*(h) = \mu_g^*(l) = \mu_b^* = 1$;
(ii) the worker's effort is given by (1);
(iii) the posteriors are equal to their priors.

3.3 The Worker's Signal Contains Some Information, $0 < \zeta < 1$

So far, we have distinguished four effects of feedback: the confidence in manager effect, the playing the odds effect, the productivity effect, and the self-confidence effect. If $0 < \zeta < 1$, potentially these four effects may simultaneously play a role. This illustrates that the effects of feedback may be quite complex.

We start the analysis by showing that if $0 < \zeta < 1$, feedback only matters in case the worker is uncertain about the manager's type. That is, we prove the following proposition

Proposition 3 *Suppose that in the feedback model $0 < \zeta < 1$ and $\rho \in \{0, 1\}$. Then, the unique equilibrium is a pooling equilibrium in which:*
(i) $\mu_g^*(h) = \mu_g^*(l) = \mu_b^* = 1$;
(ii) the worker's effort is given by (1);
(iii) the posteriors are equal to their priors.

First, suppose that the worker knows that the manager does not possess private information about his ability, $\rho = 0$. Then, feedback does not contain information, and it is optimal for the worker to ignore it. Now suppose that $\rho = 1$. Then, feedback does not provide information about the manager's ability, as the worker *knows* that the manager is able. The only effect that remains is the self-confidence effect. The worker may infer information from feedback about his own ability. However, as shown in the previous subsection, if the self-confidence effect is the only effect of feedback, the manager has an incentive to send $m = h$, irrespective of the worker's type. The main message of Proposition 3 is that informative feedback requires uncertainty about the manager's ability to assess the worker's ability and his performance. Bol (2011) presents evidence that managers are more inclined to provide biased, positive feedback to employees when their relationships are more longlasting. It seems plausible that when time elapses uncertainty about a manager's ability decreases. Against this background, Bol's finding is consistent with Proposition 3.

Having established equilibrium behavior for $\rho \in \{0, 1\}$, Proposition 4 describes equilibrium behavior for $0 < \rho < 1$.[10]

Proposition 4 *Consider the feedback model with* $0 < \rho < 1$. *Then, in any non-babbling equilibrium such that* $e^*(s, m = l) \neq e(s, m = h)$ *for some* $s \in \{l, h\}$ *we have:*
(I) $\mu_g^*(h) = 1 \geq \mu_b^* > \mu_g^*(l)$*; Moreover, if* $\mu_b^* < 1$*, then* $\mu_g^*(l) = 0$*;*
(II) the worker's effort is given by (1);
(III) $\hat{\alpha}(s, h) > \hat{\alpha}(s, l)\ \forall s \in \{h, l\}$*;*
(IV) $\hat{\rho}\left(l, l; \mu_g^*(a), \mu_b^*\right) \geq \hat{\rho}\left(h, h; \mu_g^*(a), \mu_b^*\right)$ *if* $\mu_b^* \geq \frac{(\zeta+(1-\zeta)\alpha)}{(1+\zeta)}$ *and* $\hat{\rho}\left(h, l; \mu_g^*(a), \mu_b^*\right) \geq \hat{\rho}\left(l, h; \mu_g^*(a), \mu_b^*\right)$ *if* $\mu_b^* \geq \alpha$.

Proof. See the appendix. ■

Proposition 4 presents a wide variety of results. First, consider Part I. It shows that a good manager who faces a high ability worker always provides positive feedback. If he were to provide negative feedback, he would damage his credibility (in expected terms) and would deteriorate the worker's self-confidence. A good

[10] Numerical examples of these equilibria for this case can be obtained from the authors.

manager, meeting a low ability worker, faces a trade-off. On the one hand, positive feedback improves the worker's self-confidence. On the other hand, negative feedback may enhance the worker's confidence in the manager. Finally, Part I of Proposition 4 shows that a bad manager has weaker incentives to provide positive feedback than a good manager who faces a high ability worker, but stronger incentives than a good manager who faces a low ability worker. Of course, the reason for this result is the playing the odds effect. The odds for a positive signal matching the signal of the worker are maximal if the manager knows that the worker is of the high-ability type, and minimal if the manager knows that the worker is of the low-ability type.

Part III and IV of Proposition 4 result from Bayes' rule. Part III shows that positive feedback boosts the worker's self-confidence. Part IV describes how feedback affects the worker's confidence in the manager. As in Section 3.1, the sign of the confidence in manager effect depends on the probability with which a bad manager gives positive feedback. If a bad manager predominantly provides positive (negative) feedback, providing negative (positive) feedback signals being a good manager. Together with Equation (1), Part II of Proposition 4 shows how the worker's self-confidence and his confidence in the manager determine effort.

To gain deeper insights into the variety of effects of feedback, it is convenient to assume that $\alpha = \frac{1}{2}$. In this case, the playing the odds effect is canceled out. The production effect and the self-confidence effect give incentives to a bad manager to provide positive feedback. The confidence in manager effect may temper these incentives, but never dominates them. Hence, for $\alpha = \frac{1}{2}$, $\mu_b > \frac{1}{2}$. Together with the result that a good manager, facing a high ability worker, always provides positive feedback, our model is able to explain the widely observed leniency bias: in general, managers tend to provide positive feedback. For $\alpha > \frac{1}{2}$, bad managers are even more inclined to provide positive feedback as a result of the playing the odds effect. Only if high ability workers are rare (low α), bad managers may lean to negative feedback. In line with our result that bad managers are more inclined to provide positive evidence, Bol (2011) presents evidence that managers for whom it is more costly to assess employee's performances are more lenient.

Our model highlights the importance of the interplay of the worker's self-perception and his perception of the manager's ability to assess his performance correctly. Neg-

ative feedback discourages a worker who thinks highly of himself. Such a worker would dismiss a manager who provides negative feedback as incompetent. Feedback that is consistent with the worker's self perception enhances the worker's confidence in the manager's ability to assess his performance. As a result, in our model negative feedback may encourage a worker who has a low self-perception. In line with our result, several studies have found that the effect of negative feedback on a worker's performance crucially depends on the manager's reputation for being able to assess the worker's performance correctly.

In our model, the manager can only send two messages. For this reason, our model cannot provide an explanation for the centrality bias. However, our model does suggest an explanation. It demonstrates that the effects of feedback depend on a worker's perception of his manager's ability to assess his performance correctly. A dumb manager anticipates this. He has an incentive to give feedback that is consistent with the worker's own perception. In a model with, say, three messages, a dumb manager may tend to give neutral feedback to avoid too large deviations between feedback and the worker's perception. Note that this explanation for the centrality bias requires that especially large inconsistencies between the manager's view and the worker's view on performance damage the manager's reputation for being able to assess the worker's performance correctly.

4 Conclusions

In many organizations, annual performance appraisals form the basis for the rewards employees get. In this paper, we have investigated how a manager's performance appraisal affects an employee's future performance. A key feature of our model is that both the manager and the employee have a perception of the employee's past performance. We have derived a couple of results. First, we have shown that even though a performance appraisal is cheap-talk, it may contain information that is relevant for the employee. Second, for a wide range of parameters the manager tend to give positive appraisals. Third, on average, a positive appraisal motivates an employee more than a negative appraisal. Fourth, the effect of appraisals on an employee's future performance depends on the employee's perception of the ability of the manager to assess his performance. Finally, our analysis suggests an explanation

for the centrality bias. The driving force behind most of our results is that the manager wants to come across as a person who is able to assess the performance of the agent correctly. This gives incentives to good managers to separate themselves from bad managers by giving informative feedback.

As usual, the results of our paper are derived from a model that is based on many assumptions. We have made some of these assumptions to drive home our results in a simple way. For instance, we have assumed that the good manager observes the employee's performance, while the bad manager does not have a clue. Qualitatively, assuming that a good manager is better in assessing the employee's performance than a bad manager would have sufficed.

A more restrictive assumption is that the manager is assumed to reward the employee on the basis of his perceived performance. As always, this is a restrictive assumption in case the manager is the residual claimant. Essential for our results is that the relationship between performance and pay depends on the manager's ability to assess the employee's performance.

A limitation of our model is that it does not consider long working relationships between the manager and the employee. This probably would make it hard for a bad mamager to maintain a good reputation. The employee would gradually learn the manager's type. As we have shown that performance appraisals only matter when the employee is uncertain about the manager's type, we expect that in a multi-period model the effects of performance appraisals diminish over time.

References

1. Ackerman, P.L., M.E. Beier and K.R. Bowen (2002), "What we really know about our abilities and our knowledge", *Personality and Individual Differences*, **33**, 587-605.

2. Alvero, A.M., B.R. Bucklin and J. Austin (2001), "An Objective Review of the Effectiveness and Essential Characteristics of Performance Feedback in Organizational Settings (1985-1998)", *Journal of Organizational Behavior Management* **21(1)**, 3-29.

3. Baker, G., R. Gibbons and K.J. Murphy (1994), "Subjective Performance Measures in Optimal Incentive Contracts", *The Quarterly Journal of Economics*

109, 1125-1156.

4. Balcazar, F., B.L. Hopkins and Y. Suarez (1985), "A critical, objective review of performance feedback", *Journal of Organizational Behavior Management* **7**, 65-89.

5. Baumeister, R.F. (1998). "The self", in: Gilbert, D.T., S.T. Fiske, and G. Lindzey (eds), *The Handbook of Social Psychology,* New York: McGraw Hill, 680-740.

6. Bénabou, R. and J. Tirole (2003), "Intrinsic and Extrinsic Motivation", *Review of Economic Studies* **70**, 489-520.

7. Bol, J.C. (2008), "Subjectivity in Compensation Contracting", *Journal of Accounting Literature* **27**, 1-32.

8. Bol, J.C. (2011), "The Determinants and Performance Effects of Manager's Performance Evaluation Biases", *The Accounting Review* **85(5)**, pp. 1549-1575.

9. Bull, C. (1987), "The Existence of Self-Enforcing Wage Contracts", *Quarterly Journal of Economics* **102**, 147-59.

10. Crutzen, B.S.Y., O.H. Swank and B. Visser (2007), "Confidence Management: On Interpersonal Comparisons in Teams", *Tinbergen Institute Discussion Papers* **07-040/1**, Tinbergen Institute.

11. Early, P.C. (1986), "Trust, Perceived Importance of Praise and Criticism, and Work Performance: An examination of feedback in the United States and England", *Journal of Management* **12(4)**, 457-473.

12. Gibbs, M, K.A. Merchant, W.A. van der Stede and M.E. Vargus (1994), "Determinants and Effects of Subjectivity in Incentives", *The Accounting Review* **79(2)**, 409-436.

13. Holmstrom B. (1979), "Moral Hazard and Observability, *The Bell Journal of Economics* **10(1)**, 74-91.

14. Jawahar, I.M. and C.R. Williams (1997), "Where All The Children Are Above Average: The performance appraisal purpose effect", Personnel Psychology 50, pp. 905-926.

15. Kahn, C. and G. Huberman (1988), "Two-Sided Uncertainty and "Up-or-Out" Contracts" *Journal of Labor Economics* **6(4)**, 423-444.

16. Klar, Y., A. Medding, and D. Sarel (1996), "Nonunique Invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments", *Organizational Behavior and Human Decision Processes*, **67**(2), 229-245.

17. Kluger, A.N. and A. DeNisi (1996), "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory", *Psychological Bulletin* **119(2)**, 254-284.

18. 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.

19. Lawler E.E. (1971), *Pay and Organizational Effectiveness: A psychological view*, McGraw-Hill.

20. Longenecker, C.O. (1997) "Why Managerial Performance Appraisals are Ineffective: Causes and lessons", *Career Development International* **2(5)**, 212-218

21. Medoff, J. and K. Abraham (1980), "Experience, Performance, and Earnings", *Quarterly Journal of Economics* **95(4)**, 703-36.

22. Meyer, H.H. (1975), "The Pay-for-Performance Dilemma", *Organizational Dynamics* 3(3), pp. 39-50

23. Moers, F. (2005), "Discretion and Bias in Performance Evaluation: The impact of diversity and subjectivity", *Accounting, Organizations and Society* **30**, 67-80.

24. Motowidlo, S.J. and W.C. Borman (1977), "Behaviorally anchored scales for measuring morale in military units", *Journal of Applied Psychology* **62**, 177-183.

25. Napier, N.K. and G.P. Latham (1986), "Outcome Expectancies Of People Who Conduct Performance Appraisals" *Personnel Psychology* **39**, 827-837.

26. Prendergast, C. (1993), "A Theory of Yes Men," *American Economic Review,* **83(4)**, 757-770.

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

28. Prendergast, C. and R. H. Topel (1993), "Discretion and Bias in Performance Evaluation", *European Economic Review* **37(2-3)**, 355-365.

29. Prendergast, C. and R. H. Topel (1996), "Favoritism in Organizations", *Journal of Political Economy* **104(5)**, 958-978.

30. Sedikides, C. and M. Strube (1995), "Introduction to Symposium", *Personality and Social Psychology Bulletin*, **21**(12), 1277.

31. Steelman, L.A., K.A. and Rutkoswki (2004), "Moderators of Employee Reactions toNegative Feedback" *Journal of Managerial Psychology* **19(1)**, 6-18.

32. Tziner, A., K.R. Murphy and J.N. Cleveland (2001) "Relationships between Attitudes towards Organizations and Performance Appraisal Systems and Rating Behavior", *International Journal of Selection and Assessment* **9(3)**, 226-239.

Appendix: Proof of Proposition 4

We prove each part of Proposition 4 in turn. For the proof of Part (I) we need several lemma's.

Lemmas for Part (I): First we point out that our assumption of a natural language – which we assume throughout the paper – implies $\mu_g^*(l) \leq \mu_g^*(h)$. Then we show that it is better for the manager to match the worker's private signal with her feedback than to give the other feedback message. To do that we need to derive the preference relations of the manager over the feedback messages, given how the worker responds to each combination of private signal and feedback. These preference relations are then also used in the two next lemmas which prove the

relationships between μ_b^* and respectively $\mu_g^*(h)$ and $\mu_g^*(l)$. Finally we observe that $\mu_b^* > 0$, which is the final Lemma necessary for the proof.

Lemma 1 *In any equilibrium we have* $\mu_g^*(l) \leq \mu_g^*(h)$.

Proof. By our assumption of a natural language we have $\hat{\alpha}(s,h) \geq \hat{\alpha}(s,l)$ for all $s \in \{h,l\}$. We will show that $\hat{\alpha}(s,h) \geq \hat{\alpha}(s,l)$ implies $\mu_g^*(h) \geq \mu_g^*(l)$.

$$\begin{aligned}
\hat{\alpha}(h,h) &= \frac{\alpha(\zeta+(1-\zeta)\alpha)\left(\rho\mu_g^*(h)+(1-\rho)\mu_b^*\right)}{\alpha(\zeta+(1-\zeta)\alpha)\left(\rho\mu_g^*(h)+(1-\rho)\mu_b^*\right)+(1-\alpha)(1-\zeta)\alpha\left(\rho\mu_g^*(l)+(1-\rho)\mu_b^*\right)} \\
\hat{\alpha}(h,l) &= \frac{\alpha(\zeta+(1-\zeta)\alpha)\left(\rho\left(1-\mu_g^*(h)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)}{\alpha(\zeta+(1-\zeta)\alpha)\left(\rho\left(1-\mu_g^*(h)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)+(1-\alpha)(1-\zeta)\alpha\left(\rho\left(1-\mu_g^*(l)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)} \\
\hat{\alpha}(l,h) &= \frac{\alpha(1-\zeta)(1-\alpha)\left(\rho\mu_g^*(h)+(1-\rho)\mu_b^*\right)}{\alpha(1-\zeta)(1-\alpha)\left(\rho\mu_g^*(h)+(1-\rho)\mu_b^*\right)+(1-\alpha)(\zeta+(1-\zeta)(1-\alpha))\left(\rho\mu_g^*(l)+(1-\rho)\mu_b^*\right)} \\
\hat{\alpha}(l,l) &= \frac{\alpha(1-\zeta)(1-\alpha)\left(\rho\left(1-\mu_g^*(h)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)}{\alpha(1-\zeta)(1-\alpha)\left(\rho\left(1-\mu_g^*(h)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)+(1-\alpha)(\zeta+(1-\zeta)(1-\alpha))\left(\rho\left(1-\mu_g^*(l)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)}
\end{aligned}$$

Then $\hat{\alpha}(h,h) \geq \hat{\alpha}(h,l)$ implies, after cross-multiplications of the denominators and simplification,

$$\begin{aligned}
\left(\rho\mu_g^*(h)+(1-\rho)\mu_b^*\right)\left(\rho\left(1-\mu_g^*(l)\right)+(1-\rho)\left(1-\mu_b^*\right)\right) &\geq \\
\left(\rho\left(1-\mu_g^*(h)\right)+(1-\rho)\left(1-\mu_b^*\right)\right)&\left(\rho\mu_g^*(l)+(1-\rho)\mu_b^*\right) \\
\rho\left(\mu_g^*(h)-\mu_g^*(l)\right)\left((1-\rho)\left(1-\mu_b^*\right)+(1-\rho)\mu_b^*+\rho\right) &\geq 0 \\
\rho\left(\mu_g^*(h)-\mu_g^*(l)\right) &\geq 0
\end{aligned}$$

and the result follows for $s = h$. The same steps will prove that $\hat{\alpha}(l,h) \geq \hat{\alpha}(l,l)$ implies $\mu_g^*(h) \geq \mu_g^*(l)$. ■

We now turn to the question whether a manager wants to match the private signal of the worker.

Lemma 2 *Consider a non-babbling equilibrium in which* $e(s,l) \neq e(s,h)$ *for some* $s \in \{h,l\}$, *then* $(e^*(l,l) - e^*(l,h)) > 0 \Leftrightarrow (e^*(h,h) - e^*(h,l)) > 0$ *and* $(e^*(l,l) - e^*(l,h)) < 0 \Leftrightarrow (e^*(h,h) - e^*(h,l)) < 0$.

Proof. We prove this by contradiction. Suppose not. Then without loss of generality there exist $k, k' \in \{h,l\}$, with $k \neq k'$, such that

$$(e^*(s=k,m=k) - e^*(s=k,m=k')) \leq 0 \leq (e^*(s=k',m=k') - e^*(s=k',m=k)).$$

By $e(s'', l) \neq e(s'', h)$ for some $s'' \in \{h, l\}$ at least one of these inequalities is strict. That implies that with positive probability $m = k'$ is strictly better than $m = k$, while $m = k'$ can never lead to a worse result. Thus any manager would strictly prefer $m = k'$ to $m = k$ and we have a babbling equilibrium: a contradiction. ■

Before we can proceed with the next lemma we need to derive the preference relations over the feedback messages by the managers, given $e^*(s, m)$, $s, m \in \{l, h\}$. Given the feedback strategies anticipated by the worker, μ_b^* and $\mu_g^*(a)$, we first consider the conditions for which a manager is willing to send $m = l$. Note that $\alpha = \Pr(s = h)$, $\eta = \Pr(a = h \mid s = h) = \Pr(s = h \mid a = h)$ and $\lambda = \Pr(a = l \mid s = l) = \Pr(s = l \mid a = l)$. The bad manager is willing to send $m = l$ only if

$$\begin{aligned}
\left\{ \begin{array}{c} \Pr(s = l)\, e^*(s = l, m = l)\, E(a|s = l) + \\ \Pr(s = h)\, e^*(s = h, m = l)\, E(a|s = h) \end{array} \right\} &\geq \left\{ \begin{array}{c} \Pr(s = l)\, e^*(s = l, m = h)\, E(a|s = l) + \\ \Pr(s = h)\, e^*(s = h, m = h)\, E(a|s = h) \end{array} \right\} \\
(1 - \alpha)\, e^*(l, l)\, E(a|s = l) + \alpha e^*(h, l)\, E(a|s = h) &\geq (1 - \alpha)\, e^*(l, h)\, E(a|s = l) + \alpha e^*(h, h)\, E(a|s = h) \\
(1 - \alpha)\, E(a|s = l)\, (e^*(l, l) - e^*(l, h)) &\geq \alpha E(a|s = h)\, (e^*(h, h) - e^*(h, l)) \qquad (6)
\end{aligned}$$

Similarly, if $a = h$, a good manager is willing send $m = l$ only if

$$\begin{aligned}
(1 - \eta)\, e^*(l, l)\, h + \eta e^*(h, l)\, h &\geq (1 - \eta)\, e^*(l, h)\, h + \eta e^*(h, h)\, h \\
(1 - \eta)\, (e^*(l, l) - e^*(l, h)) &\geq \eta\, (e^*(h, h) - e^*(h, l)) \qquad (7)
\end{aligned}$$

For $a = l$, a good manager facing a low ability worker is willing to send $m = l$ only if:

$$\lambda\, (e^*(l, l) - e^*(l, h)) \geq (1 - \lambda)\, (e^*(h, h) - e^*(h, l)) \qquad (8)$$

The bad manager is willing to adopt a mixed strategy if and only if (6) holds with equality. If (6) is violated, the bad manager strictly prefers to send $m = h$. The same applies for a good manager with respect to (7) if $a = h$ and with respect to (8) if $a = l$.

We can now show that a worker will put in more effort if the feedback message matches his private signal.

Lemma 3 *Consider a non-babbling equilibrium in which* $e(s,l) \neq e(s,h)$ *for some* $s \in \{h,l\}$. *Then* $(e^*(l,l) - e^*(l,h)) > 0$.

Proof. Suppose not. Then by Lemma 2 $(e^*(l,l) - e^*(l,h)) < 0$ and thus $(e^*(h,h) - e^*(h,l)) < 0$. As $(1-\lambda) < \eta$ we obtain that $\mu_g^*(l) < 1$ implies $\mu_g^*(h) = 0$. By Lemma 1 this implies that $\mu_g^*(l) = \mu_g^*(h)$ and $\mu_g^*(h) \in \{0,1\}$. It follows that $\mu_g^*(h) = \mu_b^*$, as the worker would believe that the manager is bad, whenever he observes a message which cannot be observed from a good manager. This would constitute a babbling equilibrium: a contradiction. ■

This enables us to prove the final three lemmas which together will prove Part (I).

Lemma 4 *Consider a non-babbling equilibrium in which* $e(s,l) \neq e(s,h)$ *for some* $s \in \{h,l\}$. *Then* $\mu_b^* > 0$ *implies* $\mu_g^*(h) = 1$.

Proof. Here we show, by contradiction that the good manager facing a high ability worker will strictly prefer $m = h$ if the bad manager is willing to send message $m = h$. Suppose not. Then (6) either holds with equality or is violated while (7) holds. By Lemma 3 this implies that

$$\begin{aligned} \frac{(1-\alpha)E(a|s=l)}{\alpha E(a|s=h)} &\leq \frac{(e^*(h,h) - e^*(h,l))}{(e^*(l,l) - e^*(l,h))}, \text{ and} \\ \frac{1-\eta}{\eta} &\geq \frac{(e^*(h,h) - e^*(h,l))}{(e^*(l,l) - e^*(l,h))}. \end{aligned}$$

Combining yields

$$(1-\alpha)E(a|s=l)\eta \leq \alpha E(a|s=h)(1-\eta)$$

Note that

$$\begin{aligned} E(a|s=l) &= l + (1-\zeta)\alpha(h-l) \\ E(a|s=h) &= l + (\zeta + (1-\zeta)\alpha)(h-l) \\ \eta &= \zeta + (1-\zeta)\alpha \end{aligned}$$

which gives us

$$\begin{aligned}(1-\alpha)\left(l+(1-\zeta)\,\alpha\,(h-l)\right)\left(\zeta+(1-\zeta)\,\alpha\right) &\leq \alpha\left(l+\left(\zeta+(1-\zeta)\,\alpha\right)(h-l)\right)(1-\zeta)\,(1-\alpha)\\ (1-\alpha)\,\zeta l &\leq 0\end{aligned}$$

By $\alpha<1$ and $\zeta, l>0$ this cannot hold. Thus, if $\mu_b^*>0$, then $\mu_g^*(h)=1$. ■

In a similar way, the following lemma can be derived.

Lemma 5 *Consider a non-babbling equilibrium in which* $e(s,l)\neq e(s,h)$ *for some* $s\in\{h,l\}$. *Then* $\mu_b^*<1$ *implies* $\mu_g^*(l)=0$.

Proof. Suppose not. Then (6) holds while (8) is either violated or satisfied with equality. Thus

$$\begin{aligned}\frac{\lambda}{1-\lambda} &\leq \frac{\left(e^*(h,h)-e^*(h,l)\right)}{\left(e^*(l,l)-e^*(l,h)\right)} \leq \frac{(1-\alpha)\,E(a|s=l)}{\alpha E(a|s=h)}\\ \alpha E(a|s=h)\,\lambda &\leq (1-\alpha)\,E(a|s=l)\,(1-\lambda)\\ \alpha\left(l+\left(\zeta+(1-\zeta)\,\alpha\right)(h-l)\right)\left(\zeta+(1-\zeta)\,(1-\alpha)\right) &\leq (1-\alpha)\left(l+(1-\zeta)\,\alpha\,(h-l)\right)(1-\zeta)\,\alpha\\ \alpha h\zeta &\leq 0\end{aligned}$$

By $\alpha,\zeta,l>0$ this cannot hold, which proves the lemma. ■

Now we only need to prove that μ_b^* is strictly positive, and the results will follow.

Lemma 6 *Consider a non-babbling equilibrium in which* $e(s,l)\neq e(s,h)$ *for some* $s\in\{h,l\}$. *Then* $\mu_b^*>0$.

Proof. If not, then either $\mu_b^*=\mu_g^*(l)=\mu_g^*(h)=0$ or $\mu_g^*(l)=\mu_b^*=0<\mu_g^*(h)$. In the former case, we would have a babbling equilibrium: a contradiction. In the latter case a bad manager could get the best possible result by sending feedback message h. She would convince the worker that she is a competent manager and convince the worker that he is able. Clearly sending message $m=l$ would have strictly inferior effects. Thus $\mu_b^*>0$. ■

Proof of Part (I): Lemmas 6 and 4 implies $0<\mu_b^*\leq\mu_g^*(h)=1$. By Lemma 5 we obtain $\mu_g^*(h)=1\geq\mu_b^*>\mu_g^*(l)$ and that $\mu_b^*<1$, then $\mu_g^*(l)=0$.

Proof of Part (II): this follows from the derivation of (1).

Proof of Part (III): Note that as feedback from an informed manager holds information ($\mu_g^*(h) > \mu_g^*(l)$) while the feedback of the bad manager contains no information, on average it is informative. Thus $\hat{\alpha}(s,h) > \hat{\alpha}(s,l)\ \forall s \in \{h,l\}$.

Proof of Part (IV): There are two cases. The first case is $\mu_g^*(l) > 0$, implying that the bad manager always sends message h. In that case the results follow immediately, as $m = l$ can only be sent by the good manager. Thus $\hat{\rho}\left(s, m = l; \mu_g^*(a), \mu_b^* = 1\right) = 1$, and the manager is seen as fully credible.

The second case is that $\mu_g^*(l) = 0$. We start by showing that $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) \geq \hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right)$ if $\mu_b^* \geq \frac{(\beta+(1-\beta)\alpha)}{(1+\beta)}$.

Note that $\Pr(s=h) = \beta\alpha + (1-\beta)\alpha = \alpha$. Using this we obtain the following probabilities and posteriors:

$$\begin{aligned}
\Pr(s = l \wedge m = l \wedge t = g) &= (1-\alpha)\rho(\beta + (1-\beta)(1-\alpha)) \\
\Pr(s = l \wedge m = l \wedge t = b) &= (1-\rho)(1-\alpha)(1-\mu_b^*) \\
\text{and thus} \quad \hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) &= \frac{(1-\alpha)\rho(\beta+(1-\beta)(1-\alpha))}{(1-\alpha)\rho(\beta+(1-\beta)(1-\alpha)) + (1-\rho)(1-\alpha)(1-\mu_b^*)};
\end{aligned}$$

and

$$\begin{aligned}
\Pr(s = h \wedge m = h \wedge t = g) &= \alpha\rho(\beta + (1-\beta)\alpha) \\
\Pr(s = h \wedge m = h \wedge t = b) &= (1-\rho)\alpha\mu_b^* \\
\text{and thus} \quad \hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right) &= \frac{\alpha\rho(\beta+(1-\beta)\alpha)}{\alpha\rho(\beta+(1-\beta)\alpha) + (1-\rho)\alpha\mu_b^*}.
\end{aligned}$$

Now we can rewrite $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) - \hat{\rho}\left(h,h;\mu_g^*(a),\mu_b^*\right) \geq 0$ as:

$$\begin{aligned}
\frac{(1-\alpha)\rho(\beta+(1-\beta)(1-\alpha))}{(1-\alpha)\rho(\beta+(1-\beta)(1-\alpha)) + (1-\alpha)(1-\rho)\left(1-\mu_b^*\right)} - \frac{\alpha\rho(\beta+(1-\beta)\alpha)}{\alpha\rho(\beta+(1-\beta)\alpha) + (1-\rho)\alpha\mu_b^*} &\geq 0 \\
\frac{\mu_b^*(1+\beta) - (\beta+(1-\beta)\alpha)}{\left(1-(1-\rho)\mu_b^* - \rho\alpha(1-\beta)\right)\left(-(1-\rho)\mu_b^* - \rho(\beta+(1-\beta)\alpha)\right)} &\leq 0
\end{aligned}$$

Observe that the denominator is negative as $(1-(1-\rho)\mu_b^* - \rho\alpha(1-\beta))$ is positive and $-(1-\rho)\mu_b^* - \rho(\beta+(1-\beta)\alpha)$ negative. Thus $\hat{\rho}\left(l,l;\mu_g^*(a),\mu_b^*\right) -$

$\hat{\rho}\left(h,h;\mu_g^*\left(a\right),\mu_b^*\right)\geq 0$ if and only if $\mu_b^*\left(1+\beta\right)-\left(\beta+\left(1-\beta\right)\alpha\right)\geq 0$. This holds if

$$\mu_b^*\geq\frac{\left(\beta+\left(1-\beta\right)\alpha\right)}{\left(1+\beta\right)}.$$

Now we show in the same way that $\hat{\rho}\left(h,l;\mu_g^*\left(a\right),\mu_b^*\right)\geq\hat{\rho}\left(l,h;\mu_g^*\left(a\right),\mu_b^*\right)$ if $\mu_b^*\geq\alpha$. The probabilities and posteriors are

$$\begin{aligned}\Pr\left(s=h\wedge m=l\wedge t=g\right) &= \left(1-\alpha\right)\rho\left(1-\beta\right)\alpha\\ \Pr\left(s=h\wedge m=l\wedge t=b\right) &= \left(1-\rho\right)\alpha\left(1-\mu_b^*\right)\\ \text{and thus}\quad \hat{\rho}\left(h,l;\mu_g^*\left(a\right),\mu_b^*\right) &= \frac{\left(1-\alpha\right)\rho\left(1-\beta\right)\alpha}{\left(1-\alpha\right)\rho\left(1-\beta\right)\alpha+\left(1-\rho\right)\alpha\left(1-\mu_b^*\right)};\end{aligned}$$

and

$$\begin{aligned}\Pr\left(s=l\wedge m=h\wedge t=g\right) &= \alpha\rho\left(1-\beta\right)\left(1-\alpha\right)\\ \Pr\left(s=l\wedge m=h\wedge t=b\right) &= \left(1-\rho\right)\left(1-\alpha\right)\mu_b^*\\ \text{and thus}\quad \hat{\rho}\left(l,h;\mu_g^*\left(a\right),\mu_b^*\right) &= \frac{\alpha\rho\left(1-\beta\right)\left(1-\alpha\right)}{\alpha\rho\left(1-\beta\right)\left(1-\alpha\right)+\left(1-\rho\right)\left(1-\alpha\right)\mu_b^*}.\end{aligned}$$

Thus $\hat{\rho}\left(h,l;\mu_g^*\left(a\right),\mu_b^*\right)\geq\hat{\rho}\left(l,h;\mu_g^*\left(a\right),\mu_b^*\right)$ if

$$\begin{aligned}\frac{(1-\alpha)\rho(1-\beta)\alpha}{(1-\alpha)\rho(1-\beta)\alpha+(1-\rho)\alpha\left(1-\mu_b^*\right)}-\frac{\alpha\rho(1-\beta)(1-\alpha)}{\alpha\rho(1-\beta)(1-\alpha)+(1-\rho)(1-\alpha)\mu_b^*} &\geq 0\\ \frac{\alpha-\mu_b^*}{\left(1-(1-\rho)\mu_b^*-\rho(\beta+(1-\beta)\alpha)\right)\left(-(1-\rho)\mu_b^*-\rho\alpha(1-\beta)\right)} &\geq 0\end{aligned}$$

Note that the denominator is negative as $\left(1-\left(1-\rho\right)\mu_b^*-\rho\left(\beta+\left(1-\beta\right)\alpha\right)\right)$ is positive and $\left(-\left(1-\rho\right)\mu_b^*-\rho\alpha\left(1-\beta\right)\right)$ is negative. Thus the condition becomes

$$\mu_b^*\geq\alpha$$

This concludes the proof.