The role of performance appraisals in motivating employees

Jurjen J.A. Kamphorst
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

October 7, 2015

Abstract:
Employees’ career perspectives often depend on how their supervisors perceive their performance. However, evaluating an employee’s performance is often difficult. We develop a model in which an employee is uncertain about his own performance and about the manager’s ability to assess him. The manager gives the employee a performance appraisal aiming to affect both the employee’s self perception and his own credibility in assessing the performance. We examine how performance appraisals affect the employee’s future performance. Our model’s predictions are consistent with empirical findings, among which are that (i) managers give, on average, ’too’ positive appraisals; and (ii) both positive and negative feedback can (de)motivate workers.

JEL codes: M52, M54, D82, D83

Keywords: Subjective Performance Appraisal, Credibility, Cheap Talk

*We thank Jasmijn Bol, Francis Bloch, Josse Delfgaauw, Silvia Dominguez-Martinez, Robert Dur, Kris de Jaegher, Botond Köszegi, Victor Maas, Menno Middeldorp, John Morgan, Sander Onderstal, Ronald Peeters, Canice Prendergast, Arno Riedl, Stephanie Rosenkranz, Dana Sisak, Wim van der Stede, Roland Strausz, Roland van Weelder, Utz Weitzel, and Bastian Westbrock, as well as participants of the UMR-GAEL seminar (February 2013), EALE 2012, EEA-ESEM 2012, the Workshop on Economic Theory and Game Theory (December 2011), and the ESE Brown Bag seminar for their helpful comments. Also, we gratefully acknowledge the financial support of this research by the NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) through grant no. 400-07-122.
1 Introduction

Measuring performance is an essential part of any compensation system. Objective indicators of all aspects of a complex jobs are rarely available. For those jobs, firm often have to use subjective performance evaluation. Subjective performance evaluation, by its very nature, requires that supervisors form perceptions. The accuracy of these perceptions depends on the expertise of the manager as well as other factors. Against this background, it is hardly surprising that supervisors have been found to vary in their skills to appraise employees [see, for example, Napier and Latham (1986) and Tziner et al (2001)]. This finding has potentially important implications when employees’ rewards or career perspectives depend on how their supervisors perceive their performance. For example, doubts about a supervisor’s ability to assess performance accurately weaken an employee’s incentives to exert effort. As a result, when providing performance appraisals, the supervisor’s reputation for being capable of correctly assessing performance is at stake. In this paper, we develop a model in which supervisors differ in their ability to appraise employees’ performances. We use this model to better understand how supervisors appraise their workers, and how workers respond to these performance appraisals.

Though the empirical economics literature on performance appraisals is rather limited, there is a diverse business literature on this topic. One strand in the business literature examines what kind of evaluations supervisors give. In most organizations evaluation systems comprise five rating levels. One well-known finding is that many supervisors tend to give (too) positive assessments. 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)]. This phenomenon is known as the leniency bias. Another finding is that some supervisors tend to compress performance ratings. This is known as the centrality bias [Motawidlo and Borman (1977)]. In terms of a five point scale, the centrality bias means that the ratings 1 and 5 are rarely used. In the eighties, at Merck, 97 percent of workers were offered ratings of 3 or 4 (Prendergast, 1999, footnote 36). More recently, Boll (2008) reported evidence for both the leniency and centrality bias for a financial service provider in the Netherlands. Interestingly, she found that in particular less informed supervisors bias their ratings.

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)]. This literature finds that feedback can both

\[^{1}\text{Moers (2005) finds that performance ratings on subjective dimensions are closer to the median rating than performance ratings on objective dimensions.}\]
motivate and demotivate workers. However, on average it tends to have a positive effect, especially if the feedback was positive. For the case of negative feedback, Steelman and Rutkowski (2004) show that the credibility of the supervisor affects the sign and the size of the effect of the 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.²

The contribution of this paper is twofold. First, to our knowledge our paper offers the first model that explains both the empirical results on supervisors’ behavior on providing appraisals, and how workers respond to performance appraisals in a single setting. Second, this paper shows that cheap-talk messages to motivate employees can contain information when the supervisor is concerned about her reputation for being able to assess performances correctly. Indirectly our paper contributes to the literature on compensation schemes, as measuring performance is an important aspect of compensation schemes. We refer to an earlier version of our paper for an analysis of how imperfect evaluations of employees should affect compensation schemes. Important for the results of the present paper is that, if recognized by his supervisor, a better performance benefits an employee.

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.³ 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. Third, we assume an environment where the employee’s career perspectives depend on his supervisor’s assessment of his performance. For example, the employee’s supervisor may have a say in promotion decisions or assignment of tasks. As mentioned above, the important assumption is that a better performance benefits the employee if it is recognized by the manager. Finally, the employee’s ability and his effort are complements. The implication of this last characteristic is that the more confident the employee is about his ability, the more effort he exerts.

We derive several results. Our first set of results pertains to a situation where the

²Many scholars emphasize the importance of the supervisor’s credibility (see, e.g., Lawler (1971), Meyer (1975), Early (1986) and Longenecker (1997). Gibbs et al. (1994) 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.” [p. 415; emphasis added].

³In 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)].
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 smart supervisor, who observes 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 mediocre 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 explain the central- ity 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 give relatively frequently. We show that this third 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 have relaxed the assumption that the employee knows his own ability. 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. 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 employee’s performance correctly or the employee had absolutely no clue about his

---

4 In Bénabou and Tirole (2003), talk to motivate employees does not work. It is a dominant strategy for the manager to send the message that boosts the employee’s self-confidence. In our model, the manager’s concern about his reputation for being able to assess the worker’s performance correctly may induce her to reveal her private information.

It is worth noting that reputational concerns may also induce managers to manipulate information [see, for example, Suurmond, Swank and Visser (2004)]. 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.

5 Daley and Green (2012) also study a model in which the receiver observes both a message chosen by the sender as well as an independent signal. However, the messages in Daley and Green are costly.
ability, the supervisor’s incentive to come across as able would vanish. In such a situation, if the worker would still believe that some information is contained in the feedback, the supervisor would always provide positive feedback. As a result, feedback could not contain any information about an employee’s actual performance or the supervisor’s type.

For our final result, we extend the model by introducing a third ability level. Moreover, we allow good managers to make small mistakes in their assessments. Consequently good managers can make small mistakes in their feedback. We show that this can generate a centrality bias. The reason is that a bad manager is afraid to make big mistakes as big mistakes reveal their incompetence. In contrast, small mistakes are also made by good managers and are therefore credible. ‘Average’ feedback thus becomes a safe haven to the bad manager.

Our results on supervisors’ incentives to give particular appraisals are important for understanding workers’ responses to appraisals. For example, we show that a positive appraisal motivates a worker more when he has a positive perception of his own performance. A negative appraisal, by contrast, motivates a worker with a negative perception of his performance more than a worker with a positive perception.

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), Baker et al. (1994) and Gibbs et al. (1994); see Prendergast (1999) and Bol (2008) for reviews of the literature, and Ederhof, Rajan and Reichelstein (2011) for a synthesis of the recent literature on discretionary rewards with an emphasis on the accounting literature]. Key notion in this literature is that most people do not work in jobs where all aspects of an employee’s performance are verifiable. Contracting upon subjective performance appraisals is problematic as supervisors have incentives to save labor costs by underreporting performance. However, repeated interaction may allow for an implicit contract in which rewards are based on unverifiable information.

Zábojnýk (2011) models the interaction between an objective and a subjective performance measure. Like us, he assumes that the supervisor has superior information about the employee’s ability. Unlike us, the supervisor is the residual claimant. As a result, the supervisor has an incentive to underreport performance. Zábojnýk (2011) presents very interesting results about the pros and cons of committing to specific distributions of evaluations. His model is less relevant for situations where the supervisors are not residual claimants. This also holds for Suvorov and Van de Ven (2009) who examine how the size of a reward may signal information about an employee’s ability. In both Zábojnýk (2011) and Suvorov and Van de Ven (2009), having to pay a bonus for good performance make
positive feedback credible. In our model, the supervisor’s incentive to show that he is able to assess performance makes positive feedback credible.

Our paper resembles MacLeod (2003) in the sense that both the supervisor and the worker receive a private signal about the worker’s performance. MacLeod (2003) shows that if these signals are not perfectly correlated, the optimal contract lumps good enough performances together. This can result in most performances being rated as ’above average’. In this way, the leniency bias is explained. In contrast to our paper, in MacLeod (2003) there is uncertainty neither about the type of the supervisor nor about the type of the worker.

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 it is the supervisor rather than the employee who has an incentive to conform. By appraising 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 supervisor’s appraisal is not fully trustworthy. In their model, however, the performance appraisal may deviate from the true performance because the supervisor is biased with respect to the employee. In our model, the appraisal can deviate from the true performance because the supervisor lacks the necessary expertise to judge the worker’s performance. Moreover, the supervisor has incentives to adjust the appraisal in order to manage the self-confidence of the worker.

2 The Feedback Model

Background

In an appraisal interview an employee and a manager exchange their views about the employee’s performance. An appraisal interview is usually meant to improve the employee’s future performance. We model the link between past performance and future performance through the employee’s ability. To keep things simple, in our model a performance appraisal is an ex ante statement about the employee’s ability. Implicit is an earlier period in which the employee’s performance contains information about the employee’s ability, and in turn about the employee’s future potential. Under the assumption that the employee is uncertain about his ability, in this setting a performance appraisal may affect an employee’s future work effort.

Another important feature of our model is that the employee’s payoff depends on
the manager’s view about his performance. The idea is that in many environments an employee’s career perspectives depend on how the manager perceives the employee’s performance. We do not explicitly model how a better perceived performance enhances the employee’s career perspectives. We simply assume that the employee’s payoff depends linearly on the employee’s perceived performance. Essential is that the employee must have confidence in the manager’s ability to assess his performance accurately. Otherwise he will not be induced to exert effort in order to further his career. To minimize on notation, we assume that some managers can perfectly assess employees’ performances, whereas other managers are completely unable to assess employees’ performances. Relaxing this assumption does not qualitatively affect our results.

The Formal Model

Our model describes the interaction between two risk neutral players: a worker (he) and his supervisor, the manager (she). 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, $a \in \{l, h\}$, where $h > l > 0$. 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 $\zeta$, the worker’s signal is accurate, $s = a$. With probability $(1 - \zeta)$, $s$ does not contain any information about $a$. In that case, $\Pr (s = k|a = k') = \Pr (a = k) \forall k, k' \in \{l, h\}$. Denote by $\eta$ the probability that the worker assigns to $a = h$ 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 $\lambda$ the probability that the worker believes that $a = l$ after he has received signal $s = l : \lambda = \Pr (a = l|s = l) = \zeta + (1 - \zeta) (1 - \alpha)$ 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.\footnote{In our model no supervisor wants to reveal himself as a bad manager. Therefore, allowing for a richer message space (with more cheap talk messages than worker types) does not affect the results.} 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 depends on whether or not the manager observes his performance. It 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 \( \gamma > 0 \) and \( \hat{y} \) is equal to the worker’s expected output, conditional on the information a bad manager possesses. The crucial assumption is that if the manager is bad, the employee’s effort does not affect his career perspectives, while if the manager is good, the employee’s effort does. 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 \( y \). A bad manager does not observe \( y \). The payoff to the worker depends on the manager’s type.

All priors are common knowledge, as is the probability \( \zeta \). Signal \( s \), effort \( e \) and performance \( y \) are not verifiable.

Our model is a dynamic game with incomplete information. Let \( \hat{\alpha}(s, m) \) denote the worker’s belief that he is able, \( a = h \), conditional on \( s \) and \( m \). Similarly, denote by \( \hat{\rho}(s, m) \) the worker’s belief that the manager is good, \( t = g \), conditional on \( s \) and \( m \). 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 \( m = h \). We focus on equilibria with a natural language in the sense that (i) 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 \( a = h \); and (ii) the manager credibility is not lower if the worker’s private signal matches the feedback than if the private signal does not match the feedback. Formally, these two requirements are that (i) \( \hat{\alpha}(s, h) \geq \Pr(a = h | s) \geq \hat{\alpha}(s, l) \) for any \( s \in \{h, l\} \); and (ii) \( \hat{\rho}(s = m, m) \geq \hat{\rho}(s \neq m, m) \) for any \( m \in \{h, l\} \). We sometimes refer
to the probability \( \hat{\alpha} (s, m) \) as the worker’s self-confidence. We identify Perfect Bayesian Equilibria in which (i) given the posterior probabilities \([\hat{\alpha} (s, m) \text{ and } \hat{\rho} (s, m)]\) 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.

### 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 increases his reward only in case the manager is good. His effort results from maximizing \( \hat{\rho} (s, m) E (a|s, m) e - \frac{1}{2} \gamma e^2 \) with respect to \( e \), yielding

\[
e^*(s, m) = \frac{\hat{\rho} (s, m) \{\hat{\alpha} (s, m) h + [1 - \hat{\alpha} (s, m)] l\}}{\gamma}
\]

(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 belief 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\(^7\)

\[
\begin{align*}
\hat{\rho}(h, l) &= 0 \\
\hat{\rho}(l, h) &= 0 \\
\hat{\rho}(l, l) &= \frac{\rho}{\rho + (1 - \rho)(1 - \mu_{\text{ant}}^b)} \\
\hat{\rho}(h, h) &= \frac{\rho}{\rho + (1 - \rho)\mu_{\text{ant}}^b}
\end{align*}
\]  

(2)

where \(\mu_{\text{ant}}^b\) is the bad manager’s feedback strategy as anticipated by the workers.

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 bad manager sends that message. For instance, if a bad manager rarely sends \(m = l\) (\(\mu_b\) close to 1), then \(\hat{\rho}(l, l)\) is close to 1. More in particular, the higher is \(\mu_b\), the lower is \(\hat{\rho}(h, h)\) and the higher is \(\hat{\rho}(l, l)\). Note that if the manager is more likely to be competent, the strategy of a bad manager \((\mu_b)\) is less important.

We have established how much effort the worker exerts in equilibrium, the dominant strategy of a \(t = g\) manager, and the posteriors. Now we analyze the optimal response of a bad manager given these strategies and beliefs. Using (1) and the posteriors, the bad manager’s expected payoff is then

\[
y(m) = \begin{cases} 
\alpha \frac{\rho}{\rho + (1 - \rho)\mu_{\text{ant}}^b} \gamma h^2 & \text{if } m = h \\
(1 - \alpha) \frac{\rho}{\rho + (1 - \rho)(1 - \mu_{\text{ant}}^b)} \gamma l^2 & \text{if } m = l
\end{cases}
\]  

(3)

It is easy to see that always choosing \(m = l\) (\(\mu_b = 0\)) is an equilibrium response of a bad manager\(^8\) if

\[
(1 - \alpha) \rho l^2 \geq \alpha h^2
\]  

(4)

The left-hand side of (4) denotes the expected output when a worker receives \(m = l\), given \(\mu_{\text{ant}}^b = 0\). The right-hand side gives the expected output when the worker receives \(m = h\), given \(\mu_{\text{ant}}^b = 0\). In an equilibrium with \(\mu_b = 0\), the worker does not infer information from \(m = l\), \(\hat{\rho}(l, l) = \rho\), but learns the manager’s type if \(m = h\): \(\hat{\rho}(h, h) = 1\). One can show

\(^7\)In the special case where \(\mu_b = 0\) or \(\mu_b = 1\), \(\hat{\rho}(l, h)\) respectively \(\hat{\rho}(h, l)\) 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.

\(^8\)Note that in such an equilibrium \(\mu_{\text{ant}}^b = \mu_b = 0\). Thus (3) becomes \((1 - \alpha) \rho \frac{\gamma h^2}{\gamma}\). By sending \(m = h\) if \(\mu_{\text{ant}}^b = 0\), the bad manager would convince all able workers that she is good: \(\hat{\rho}(h, h) = 1\). Thus such workers would put in effort \(\frac{h}{\gamma}\), producing \(\frac{h^2}{\gamma}\). Thus a bad manager would prefer \(m = h\) unless \((1 - \alpha) \rho \frac{l^2}{\gamma} > \alpha \frac{h^2}{\gamma}\).
that this equilibrium requires that $\alpha$ 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 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 equilibrium response of a bad manager if

$$\alpha \rho h^2 \geq (1 - \alpha) l^2 \quad (5)$$

The left-hand side of (5) gives the expected output when a worker receives $m = h$, given $\mu_b^{\text{ant}} = 1$. The right-hand side gives the expected output when the worker receives $m = l$, given $\mu_b^{\text{ant}} = 1$. If in equilibrium $\mu_b = 1$, then the worker does not infer information from $m = h$, so that $\hat{\rho}(h, h) = \rho$, but learns the manager’s type in case $m = l$, $\hat{\rho}(l, l) = 1$. 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 bad 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 (4) and (5) are violated. A bad manager is indifferent between sending $m = l$ and sending $m = h$ if

$$(1 - \alpha) \hat{\rho}(l, l) l^2 = \alpha \hat{\rho}(h, h) h^2,$$  

so that

$$\mu_b = \frac{\alpha h^2 - (1 - \alpha) \rho l^2}{(1 - \rho) (\alpha h^2 + (1 - \alpha) l^2)} \quad (6)$$

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

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 $\mu_h = 1$

(iii) if $\frac{(1 - \alpha) l^2}{\rho} > \alpha h^2 > (1 - \alpha) \rho l^2$, an equilibrium exists in which $\mu_h$ is given by (6).

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 $\alpha$.

### 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 only the self-confidence effect is important, no information can be revealed by the feedback in equilibrium. Suppose the worker would learn something about her ability from the feedback, then $\hat{\alpha} (s, h) > \hat{\alpha} (s, l)$. As $\hat{\rho} (s, m) = \rho$ for all $s$ and $m$, a manager would always give positive feedback: $\mu_g (h) = \mu_g (l) = \mu_b = 1$. Consequently, nothing can
be learned from the feedback. This contradiction implies that no informative equilibrium exists.

In the resulting babbling equilibrium we will have $\mu^*_g(h) = \mu^*_g(l) = \mu^*_b$. The first equality is necessary as otherwise the worker would be able to learn about his ability from the feedback. The latter equality prevents the worker from learning more about the manager.

**Proposition 2** Suppose that in the feedback model $\zeta = 0$. Then no informative equilibrium exists, i.e. in any equilibrium:

(i) $\mu^*_g(h) = \mu^*_g(l) = \mu^*_b$;
(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. In the previous two sub-sections, the self-confidence effect of feedback did not exist in equilibrium. In the present section, we show that if $0 < \zeta < 1$, potentially all four effects simultaneously play a role.

We start the analysis by showing that if $0 < \zeta < 1$, feedback is uninformative in case the worker is certain 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 no informative equilibrium exists, i.e. in any equilibrium:

(i) $\mu^*_g(h) = \mu^*_g(l)$;
(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

---

9Note that $\mu^*_h \neq \mu^*_g(l)$ is possible, as no new information regarding the manager’s type can be revealed.
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) $\mu_y^* (h) = 1 \geq \mu_b^* > \mu_y^* (l)$; Moreover, if $\mu_b^* < 1$, then $\mu_y^* (l) = 0$;

(II) the worker’s effort is given by (I);

(III) $\hat{\alpha} (s, h) > \hat{\alpha} (s, l) \forall s \in \{h, l\}$;

(IV) $\hat{\rho} (l, l) \geq \hat{\rho} (h, h)$ if and only if $\mu_b^* \geq \frac{(\gamma + (1 - \gamma)\alpha)}{(1 + \gamma)}$ and $\hat{\rho} (h, l) \geq \hat{\rho} (l, h)$ if and only if $\mu_b^* \leq \alpha$.

**Proof.** See Appendix A. ■

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 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.

Parts 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

10Numerical examples of these equilibria for this case can be obtained from the authors.
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 $\alpha$), 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. Negative 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, Steelman and Rutkowski (2004) experimental findings show that the effect of negative feedback on a worker’s performance crucially depends on the supervisor’s credibility in assessing the worker’s performance correctly.

When we consider the distribution of feedback messages, the productivity effect and the self-confidence effect suggest that feedback will tend to be compressed at the high feedback message. However, this compression is more a symptom of a leniency bias rather than of a centrality bias. To disentangle the centrality bias from the leniency bias we would need a model in which the feedback could be compressed around a feedback message which is more neutral. In that sense, to study the centrality bias well we need to extend our model to have at least three ability types. This we do in the next section.
4 The Centrality Bias

The centrality bias represents the phenomenon that managers avoid extreme ratings. In case of a three point scale, the centrality bias means that managers tend to report rating two. In the basic feedback model, a manager can send two messages. This model is therefore not equipped to explain the centrality bias.

In this section, we extend the basic feedback model in two ways. First, we assume three types of workers, rather than two: \( a \in \{l, n, h\} \) with \( h > n > l \). To keep things simple, we assume that \( \Pr (a = l) = \Pr (a = n) = \Pr (a = h) = \frac{1}{3} \). In this way, we eliminate the playing-the-odds effect discussed in Section 3.1. As we will show later, in the extended model another kind of the playing-the-odds effect arises. In line with allowing for three ability types, we allow the manager to send three messages, \( m \in \{l, n, h\} \). The second extension is that we relax the assumption that a good manager always observes the worker’s ability correctly. In this section, we assume that a good manager receives a partially informative signal about the worker’s ability, \( s_g = \{l, n, h\} \). A good manager may make small mistakes, but never makes big mistakes. Specifically, we assume that

\[
\begin{align*}
\Pr (s_g = l|a = l) &= \Pr (s_g = h|a = h) = \pi; \\
\Pr (s_g = n|a = l) &= \Pr (s_g = n|a = h) = 1 - \pi; \\
\Pr (s_g = n|a = n) &= 2\pi - 1; \\
\Pr (s_g = h|a = n) &= \Pr (s_g = l|a = n) = 1 - \pi.
\end{align*}
\]

where \( \pi \in (\frac{2}{3}, 1) \). The lower bound of \( \pi \) ensures that, given \( s_g \), the most likely ability level of the worker is equal to \( s_g \). As in the basic model, we assume that a bad manager does not receive a signal about the worker’s ability. We maintain the assumption that at the end of the game, the worker’s output is recognized by a good manager, but not by a bad manager.

The main objective of this section is twofold. First, we want to show that equilibria of the extended feedback model exist in which bad managers choose \( m = n \). Second, we want to demonstrate that the productivity effect and the confidence in management effect, identified in the basic feedback model, are also present in the extended feedback model. The productivity effect helps to explain the leniency bias. Hence, the extended feedback model explains both the leniency bias and the centrality bias in a single setting. It also shows under which circumstances each bias occurs. Not surprisingly, several results derived from the basic feedback model also hold in the extended feedback model. To avoid repetitions we therefore do not present a full analysis of the extended model, but focus on
a few interesting outcomes.

As before, a PBE equilibrium consists of a set of beliefs, a strategy of the worker, a strategy of the bad manager and a strategy of the good manager. As in the present model the worker knows his own ability, his optimal strategy is to choose effort \( e = \frac{\hat{\rho}(a,m)a}{\gamma} \). This expression clearly shows that, as in Section 3, the worker’s motivation crucially depends on his perception about the manager’s competence. To minimize on notation, we assume that \( \gamma = 1 \). Concerning the good manager, we focus on equilibria in which the good manager honestly reveals her signal. The focus of the analysis is therefore on the strategy of the bad manager, \( \theta_b^k = \Pr (m = k | t = b) \). In English, what are the probabilities that in equilibrium a bad manager give positive, neutral and negative feedback?

When a good manager always honestly reveals her signal, the posterior probabilities that a manager is good are\(^\text{11}\)

\[
\begin{align*}
\hat{\rho}(l,l) &= \frac{\rho \pi}{\rho \pi + (1 - \rho) \theta_b^k} \\
\hat{\rho}(n,l) &= \frac{\rho (1 - \pi)}{\rho (1 - \pi) + (1 - \rho) \theta_b^k} \\
\hat{\rho}(h,l) &= 0 = \hat{\rho}(l,h) \\
\hat{\rho}(l,n) &= \frac{\rho (1 - \pi)}{\rho (1 - \pi) + (1 - \rho) \theta_b^k} = \hat{\rho}(h,n) \\
\hat{\rho}(n,n) &= \frac{\rho (2\pi - 1)}{\rho (2\pi - 1) + (1 - \rho) \theta_b^m} \\
\hat{\rho}(n,h) &= \frac{\rho (1 - \pi)}{\rho (1 - \pi) + (1 - \rho) \theta_b^k} \\
\hat{\rho}(h,h) &= \frac{\rho \pi}{\rho \pi + (1 - \rho) \theta_b^h}
\end{align*}
\]

Two features of the above posteriors are worth mentioning. First, as \( \pi > \frac{2}{3} \), given \( a, m = a \) yields a higher posterior probability than \( m \neq a \). This feature gives a manager an incentive to try to provide feedback that is consistent with the employee’s self perception. Second, \( \hat{\rho}(a,m) \) is decreasing in \( \theta_b^m \). This feature is responsible for the confidence in manager effect. As a result of this effect, in equilibrium bad managers may follow mixed strategies.

Now consider the problem a bad manager faces. For the three alternative messages

\(^\text{11}\)In case, \( \theta_b^k = 1 \), we assume \( \hat{\rho}(h,l) = 0 = \hat{\rho}(l,h) \) as out-of-equilibrium beliefs.
her expected payoff equals

\[
\frac{1}{3} \left[ \hat{p}(l,l) l^2 + \hat{p}(n,l) n^2 \right] \quad \text{if } m = l \\
\frac{1}{3} \left[ \hat{p}(l,n) l^2 + \hat{p}(n,n) n^2 + \hat{p}(h,n) h^2 \right] \quad \text{if } m = n \\
\frac{1}{3} \left[ \hat{p}(n,h) n^2 + \hat{p}(h,h) h^2 \right] \quad \text{if } m = h
\]

(7)

We are now ready to establish the possibility of the centrality bias.

**Proposition 5** Suppose an equilibrium of the extended feedback model in which a good manager sends \( m = s^g \) and a bad manager follows a pure strategy. Then, \( \theta_b^m = 1 \). Moreover, for a range of parameters there exists an equilibrium such that \( m = s^g \) and \( \theta_b^n = 1 \).

**Proof.** See Appendix B. ■

Proposition 5 is a direct implication of the feature of our model that good managers may make small errors but never make large errors. By giving neutral feedback, a bad manager never ruins her reputation. Feedback \( m = n \) is a safe haven. Proposition 5 can also be interpreted as a variation of the playing-the-odds effect. In Section 3, the playing the odds effect reflects a bad manager’s inclination to send a message that is likely correct. In the present model, a bad manager has an incentive to send \( m = n \) to avoid that his message is completely wrong.

Proposition 5 illustrates bad managers’ inclination to avoid extreme ratings. In the model, there are two forces that may drive bad managers away from sending \( m = n \). These forces emerge from the productivity and confidence in manager effect. First, with \( \theta_b^m = 1 \), \( m \neq n \) boosts the worker’s confidence in the manager unless \( m \) is the opposite of \( a \). This gives a bad manager an incentive to deviate from \( m = n \). The lower is the probability that a manager is good, the stronger is the confidence in manager effect. For this reason, \( \theta_b^n = 1 \) can only be part of an equilibrium if \( \rho \) is high. Second, the productivity effect reflects that it is especially important that higher ability workers expend effort. This gives bad managers an incentive to send \( m = h \) rather than \( m = l \). Note that if in equilibrium \( \theta_b^h > 0 \), the confidence in manager effect favors sending \( m = l \). For low values of \( \rho \), and \( l, n, \) and \( h \) close to each other, an equilibrium exists in which a bad manager sends all messages with positive probability. Due to the productivity effect we must have, however, that \( \theta_b^h > \theta_b^l \) in those equilibria.
5 Conclusions

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 tends 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 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 manager 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


### A 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 $\mu^*_b$ 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 $\mu^*_g (h) \geq \mu^*_g (l)$.

\[
\hat{\alpha} (h, h) = \frac{\alpha (\zeta (1-\zeta) \alpha (\rho \mu^*_g (h) + (1-\rho) \mu^*_b))}{\alpha (\zeta (1-\zeta) \alpha (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)))}
\]
\[
\hat{\alpha} (h, l) = \frac{\alpha (\zeta (1-\zeta) \alpha (\rho \mu^*_g (h) + (1-\rho) \mu^*_b))}{\alpha (\zeta (1-\zeta) \alpha (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)))}
\]
\[
\hat{\alpha} (l, h) = \frac{\alpha (\zeta (1-\zeta) \alpha (\rho \mu^*_g (h) + (1-\rho) \mu^*_b))}{\alpha (\zeta (1-\zeta) \alpha (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)))}
\]
\[
\hat{\alpha} (l, l) = \frac{\alpha (\zeta (1-\zeta) \alpha (\rho \mu^*_g (h) + (1-\rho) \mu^*_b))}{\alpha (\zeta (1-\zeta) \alpha (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)))}
\]

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

\[
(\rho \mu^*_g (h) + (1-\rho) \mu^*_b) (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)) \geq (\rho (1-\mu^*_g (h)) + (1-\rho) (1-\mu^*_b)) (\rho \mu^*_g (h) + (1-\rho) (1-\mu^*_b))
\]
\[
\rho (\mu^*_g (h) - \mu^*_g (l)) ((1-\rho) (1-\mu^*_b) + (1-\rho) \mu^*_b + \rho) \geq 0
\]
\[
(\mu^*_g (h) - \mu^*_g (l)) \geq 0
\]

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, \( \mu_b^* \) 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

\[
\begin{align*}
\Pr (s = l) e^* (s = l, m = l) E (a \mid s = l) + \\
\Pr (s = h) e^* (s = h, m = l) E (a \mid s = h)
\end{align*}
\geq
\begin{align*}
\Pr (s = l) e^* (s = l, m = h) E (a \mid s = l) + \\
\Pr (s = h) e^* (s = h, m = h) E (a \mid s = h)
\end{align*}
\]

\[
(1 - \alpha) e^* (l, l) E (a \mid s = l) + \\
\alpha e^* (h, l) E (a \mid s = h) 
\geq \\
(1 - \alpha) e^* (l, h) E (a \mid s = l) + \\
\alpha e^* (h, h) E (a \mid s = h) 
\]

\[
(1 - \alpha) E (a \mid s = l) (e^* (l, l) - e^* (l, h)) 
\geq 
\alpha E (a \mid s = h) (e^* (h, h) - e^* (h, l)) \tag{8}
\]

Similarly, if \( a = h \), a good manager is willing send \( m = l \) only if

\[
(1 - \eta) e^* (l, l) h + \eta e^* (h, l) h 
\geq 
(1 - \eta) e^* (l, h) h + \eta e^* (h, h) h
\]

\[
(1 - \eta) (e^* (l, l) - e^* (l, h)) 
\geq 
\eta (e^* (h, h) - e^* (h, l)) \tag{9}
\]

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

\[
\lambda (e^* (l, l) - e^* (l, h)) 
\geq 
(1 - \lambda) (e^* (h, h) - e^* (h, l)) \tag{10}
\]

The bad manager is willing to adopt a mixed strategy if and only if (8) holds with equality. If (8) is violated, the bad manager strictly prefers to send \( m = h \). The same applies for a good manager with respect to (9) if \( a = h \) and with respect to (10) 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 \( \mu_b^* (l) < 1 \) 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 \( \mu_b^* (h) = \mu_b^* \), 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 \( \mu_h^* > 0 \) 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 (8) either holds with equality or is violated while (9) holds. By Lemma 3 this implies that

\[
\frac{(1 - \alpha) E(a|s = l)}{\alpha E(a|s = h)} \leq \frac{(e^*(h,h) - e^*(h,l))}{(e^*(l,l) - e^*(l,h))}, \text{ and}
\]

\[
\frac{1 - \eta}{\eta} \geq \frac{(e^*(h,h) - e^*(h,l))}{(e^*(l,l) - e^*(l,h))}.
\]

Combining yields

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

Note that

\[
E(a|s = l) = l + (1 - \zeta) \alpha (h - l)
\]

\[
E(a|s = h) = l + (\zeta + (1 - \zeta) \alpha) (h - l)
\]

\[
\eta = \zeta + (1 - \zeta) \alpha
\]

which gives us

\[
(1 - \alpha) (l + (1 - \zeta) \alpha (h - l)) (\zeta + (1 - \zeta) \alpha) \leq \alpha (l + (\zeta + (1 - \zeta) \alpha) (h - l)) (1 - \zeta) (1 - \alpha)
\]

\[
(1 - \alpha) \zeta l \leq 0
\]

By \( \alpha < 1 \) and \( \zeta, l > 0 \) this cannot hold. Thus, if \( \mu_h^* > 0 \), 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 \( \mu_h^* < 1 \) implies \( \mu_g^*(l) = 0 \).
Proof. Suppose not. Then (8) holds while (10) is either violated or satisfied with equality. Thus

\[
\frac{\lambda}{1 - \lambda} \leq \frac{\left( e^* (h, h) - e^* (h, l) \right)}{\left( e^* (l, l) - e^* (l, h) \right)} \leq \frac{(1 - \alpha) E (a | s = l)}{\alpha E (a | s = h)}
\]

\[
\alpha (l + (\zeta + (1 - \zeta) \alpha) (h - l)) (\zeta + (1 - \zeta) (1 - \alpha)) \leq (1 - \alpha) (l + (1 - \zeta) \alpha (h - l)) (1 - \zeta) \alpha
\]

\[
\alpha h \zeta \leq 0
\]

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

Now we only need to prove that \( \mu_b^* \) 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 \( \mu_b^* = \mu_y^* (l) = \mu_y^* (h) = 0 \) or \( \mu_y^* (l) = \mu_y^* (h) = 0 \). 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 \( 0 < \mu_b^* \leq \mu_y^* (h) = 1 \). By Lemma 5 we obtain \( \mu_y^* (h) = 1 \geq \mu_b^* > \mu_y^* (l) \) and that if \( \mu_b^* < 1 \), then \( \mu_y^* (l) = 0 \).

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 \( \mu_y^* (h) > \mu_y^* (l) \) 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 \( \mu_y^* (l) > 0 \). Part (I) then implies \( \mu_b^* = 1 \). By Bayesian updating this implies that \( \hat{\rho} (s, m = l) = 1 \) as \( m = l \) can only be sent by the good manager. Thus \( \mu_b^* \geq \max \left\{ \frac{(\zeta + (1 - \zeta) \alpha)}{(1 + \zeta)}, \alpha \right\} \) and \( \hat{\rho} (s, l) \geq \hat{\rho} (s, h) \).

The second case is that \( \mu_y^* (l) = 0 \). We start by showing that \( \hat{\rho} (l, l) \geq \hat{\rho} (h, h) \) 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:

\[
\Pr(s = l \land m = l \land t = g) = (1 - \alpha) \rho (\beta + (1 - \beta) (1 - \alpha))
\]

\[
\Pr(s = l \land m = l \land t = b) = (1 - \rho) (1 - \alpha) (1 - \mu_b^*)
\]

and thus $\hat{\rho}(l, l) = \frac{(1 - \alpha) \rho (\beta + (1 - \beta) (1 - \alpha))}{(1 - \alpha) \rho (\beta + (1 - \beta) (1 - \alpha)) + (1 - \rho) (1 - \alpha) (1 - \mu_b^*)}$;

and

\[
\Pr(s = h \land m = l \land t = g) = \alpha \rho (\beta + (1 - \beta) \alpha)
\]

\[
\Pr(s = h \land m = l \land t = b) = (1 - \rho) \alpha \mu_b^*
\]

and thus $\hat{\rho}(h, h) = \frac{\alpha \rho (\beta + (1 - \beta) \alpha)}{\alpha \rho (\beta + (1 - \beta) \alpha) + (1 - \rho) \alpha \mu_b^*}$.

Now we can rewrite $\hat{\rho}(l, l) - \hat{\rho}(h, h) \geq 0$ as:

\[
\frac{(1 - \alpha) \rho (\beta + (1 - \beta) (1 - \alpha))}{(1 - \alpha) \rho (\beta + (1 - \beta) (1 - \alpha)) + (1 - \alpha) (1 - \rho) (1 - \mu_b^*)} - \frac{\alpha \rho (\beta + (1 - \beta) \alpha)}{\alpha \rho (\beta + (1 - \beta) \alpha) + (1 - \rho) \alpha \mu_b^*} \geq 0
\]

\[
\frac{\mu_b^* (1 + \beta) - (\beta + (1 - \beta) \alpha)}{(1 - \alpha) (1 - \rho) (1 - \mu_b^*) - (1 - \rho) (1 - \mu_b^*) - \rho (\beta + (1 - \beta) \alpha)} \leq 0
\]

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

\[
\mu_b^* \geq \frac{(\beta + (1 - \beta) \alpha)}{1 + \beta}.
\]

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

\[
\Pr(s = h \land m = l \land t = g) = (1 - \alpha) \rho (1 - \beta) \alpha
\]

\[
\Pr(s = h \land m = l \land t = b) = (1 - \rho) \alpha (1 - \mu_b^*)
\]

and thus $\hat{\rho}(h, l) = \frac{(1 - \alpha) \rho (1 - \beta) \alpha}{(1 - \alpha) \rho (1 - \beta) \alpha + (1 - \rho) \alpha (1 - \mu_b^*)}$;

and

\[
\Pr(s = l \land m = h \land t = g) = \alpha \rho (1 - \beta) (1 - \alpha)
\]
\[
\Pr(s = l \land m = h \land t = b) = (1 - \rho) (1 - \alpha) \mu_b^*
\]
and thus \( \hat{\rho} (l, h) = \frac{\alpha \rho (1 - \beta) (1 - \alpha)}{\alpha \rho (1 - \beta) (1 - \alpha) + (1 - \rho) (1 - \alpha) \mu_b^*}. \)

Thus \( \hat{\rho} (h, l) \geq \hat{\rho} (l, h) \) if

\[
\frac{(1 - \alpha) \rho (1 - \beta) \alpha}{(1 - \alpha) \rho (1 - \beta) \alpha + (1 - \rho) \alpha (1 - \mu_b^*)} - \frac{\alpha \rho (1 - \beta) (1 - \alpha)}{\alpha \rho (1 - \beta) (1 - \alpha) + (1 - \rho) (1 - \alpha) \mu_b^*} \geq 0
\]

\[
\frac{\alpha - \mu_b^*}{(1 - (1 - \rho) \mu_b^* - \rho (1 - \beta) \alpha)/(1 - (1 - \rho) \mu_b^* - \rho (1 - \beta) \alpha)} \geq 0
\]

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

\[
\mu_b^* \geq \alpha
\]

This concludes the proof.

## B Proof of Proposition 5

**Proof.** We first prove that in any equilibrium such that a good manager sends \( m = s \) we have \( \theta_b^a > 0 \). Suppose not. Then by \( \theta_b^a = 0 \) we have that \( \theta_b^a \geq 0 \) for some \( k \in \{l, h\} \). Then it must be the case that

\[
\frac{1}{3} \hat{\rho} (h, k) h^2 + \frac{1}{3} \hat{\rho} (n, k) n^2 + \frac{1}{3} \hat{\rho} (l, k) l^2 \geq \frac{1}{3} h^2 + \frac{1}{3} n^2 + \frac{1}{3} l^2.
\]

This inequality does not hold as either \( \hat{\rho} (h, k) = 0 \) or \( \hat{\rho} (l, k) = 0 \).

We now prove by example that such an equilibrium exists for a specific set of parameters. By continuity and strictness of preferences, this equilibrium exists for a range of parameters. Let \( \pi = \frac{3}{4} \). This implies that if a good manager observes \( m = n \) each ability level is equally likely as \( (1 - \pi) = (2\pi - 1) = \frac{1}{5} \). Moreover let \( \rho = \frac{9}{10} \). Consequently \( \hat{\rho} (a \neq n, n) = \frac{\rho (1 - \pi)}{\rho (1 - \pi) + (1 - \rho)} = \frac{3}{4} \) and \( \hat{\rho} (n, n) = \frac{\rho (2\pi - 1)}{\rho (2\pi - 1) + (1 - \rho)} = \frac{3}{4} \). Then there is an equilibrium in which \( \theta_g (s) = m \) and \( \theta_b^a = 1 \) if \( l = 9, n = 10 \) and \( h = 11 \). Let us first consider the bad manager. In this equilibrium the bad manager expects to earn \( \hat{\rho} (a \neq n, n) 9^2 + \hat{\rho} (n, n) 10^2 + \hat{\rho} (a \neq n, n) 11^2 \frac{3}{4} = \frac{302}{4} > 75 \). If the manager deviates by sending message \( m \in \{l, h\} \) then the manager earns \( \frac{m^2 + 10^2}{3} \leq \frac{h^2 + 10^2}{3} = \frac{221}{3} < 75 \). Thus the bad manager strictly prefers \( m = n \). Now consider the good manager. If he receives signal \( s = n \), then
each ability is equally likely. In other words, he faces the same choice as the bad manager and prefers \( m = n \). If \( s \neq n \), then by sending \( m = s \), he receives \( \pi s^2 + (1 - \pi) n^2 \). By sending \( m = n \), he receives only \( \pi \hat{p}(s,n)s^2 + (1 - \pi) \hat{p}(n,n)n^2 = \frac{3}{4} \left( \pi s^2 + (1 - \pi) n^2 \right) \). Sending \( m = \{l,h\} \setminus \{s\} \) yields \( (1 - \pi) n^2 \) which is also strictly less than \( \left( \pi s^2 + (1 - \pi) n^2 \right) \).

Therefore the good manager strictly prefers \( m(s) = s \). As both players strictly prefer their proposed strategy to any strategy, it follows that this is an equilibrium. \( \blacksquare \)