Received: 3 November 2015 | Revised: 2 October 2017 | Accepted: 3 January 2018

DOI: 10.1111/jems.12241

ORIGINAL ARTICLE

WILEY Journal of Economics & Management Strategy

The role of performance appraisals in motivating employees

Jurjen J.A. Kamphorst | Otto H. Swank

Erasmus School of Economics and Tinbergen Institute, Rotterdam, The Netherlands (Email: kamphorst@ese.eur.nl; swank@ese.eur.nl)

Abstract

Workers' rewards and career perspectives often depend on how their supervisors perceive their performance. However, evaluating a worker's performance is often difficult. We develop a model in which a worker is uncertain about his own performance and about his supervisor's ability to assess him. The supervisor gives the worker a performance appraisal aiming to affect both the worker's self-perception and his own credibility in assessing the performance. We examine how performance appraisals affect the worker's future performance. Our model's predictions are consistent with empirical findings. Supervisors give, on average, "too" positive appraisals, and both positive and negative feedback can (de)motivate workers.

1 | INTRODUCTION

In 2003, a large financial service provider (FSP) in the Netherlands introduced a new incentive system for its workers (Bol, 2008). An important feature of the new incentive system was that compensation of workers depended on various subjective performance measures. Managers rated subordinates on a five-point scale on issues like cooperative behavior, business development skills, and organization skills. Bol (2008) examined how managers at the FSP rated their workers. She found that these managers were subject to two kinds of biases. First, they provided their subordinates higher ratings than was warranted by their performances. This tendency is known as the leniency bias.[1] Second, managers tended to discriminate too little. This phenomenon is known as the centrality bias.[2] Bol also examined the determinants of the biases. She found that managers who had less information about their workers' performances were more subject to both biases. Finally, Bol examined the consequences of biased ratings for workers' future performances.[3] She found that the leniency bias on average positively influenced future performances of workers, whereas the centrality bias on average negatively affected future performances.

Measuring performance is an essential part of any compensation system. Objective indicators of all aspects of a complex job are rarely available. For those jobs, firms often use subjective performance evaluation. Subjective performance evaluation, by its very nature, requires that supervisors form perceptions. Supervisors have been found to vary in their skills to appraise workers (see, apart from Bol, 2008, Napier & Latham, 1986; Tziner, Murphy, & Cleveland, 2001). This finding has potentially important implications when workers' rewards or career perspectives depend on how their supervisors perceive their performance. Doubts about a supervisor's ability to assess performance accurately weaken a worker'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.[4] In this paper, we develop a model in which supervisors differ in their ability to appraise workers' performances. We use this model to better understand how supervisors appraise their workers, and how workers respond to these performance appraisals.

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, Bastian Westbrock, two anonymous referees and coeditor, 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.

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 workers can contain information when the supervisor is concerned about her reputation for being able to assess performances correctly.[5]

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 workers should affect compensation schemes.[6] Important for the results of the present paper is that, *if recognized by his supervisor*, a better performance benefits a worker.

The model we develop has four key characteristics. First, at the beginning of the game, both the supervisor and the worker form a perception of the worker's ability. We model this formation of perceptions by assuming that the supervisor and the worker receive private signals.[7] Second, we assume that supervisors differ in their abilities to assess the worker's performance correctly. The motivation for this assumption is that supervisors have been found to vary in their beliefs about their skills to appraise their subordinates. In our model, it is important that a supervisor who is better at observing the worker's ability is also better in judging the worker's performance. Third, we assume an environment where the worker's utility depends on his supervisor's assessment of his performance.[8] For instance, the worker may desire recognition,[9] or the assessment may affect the assignment of tasks, the worker's bonus or his (internal) career opportunities. At the FSP studied by Bol (2008), a worker's bonus depends on his supervisor's rating. Key is that a better performance benefits the worker *only if it is recognized by the supervisor*. Finally, the worker's ability and his effort are complements. The implication of this last characteristic is that the more confident the worker is about his ability, the more effort he exerts.

We derive several results. Our first set of results pertains to a situation where the worker knows his own ability. In this extreme situation, performance appraisals only provide information about the supervisor's ability to assess the worker's ability correctly. A supervisor who gives an incorrect assessment of a worker's performance loses credibility. A direct implication is that a good supervisor, who observes a worker's performance, has no incentive to rate it incorrectly. This would only damage his credibility. The worker would doubt whether his future performance would be correctly assessed. For a bad supervisor, who does not observe a worker'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 worker's perception. Second, as the worker's ability and his effort are complements, it is more important for the supervisor that her evaluation is correct in case the worker is more able. This force leads to a positive bias in performance appraisals. Finally, a bad supervisor wants to come across as good. 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 is derived from the version of the model in which we have relaxed the assumption that the worker 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 worker'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 workers' 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 a worker signals confidence and thereby motivates. New is that simple cheap-talk messages may motivate workers.[10,11]

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 Baker, Gibbons, & Murphy, 1994; Bull, 1987; and Gibbs et al., 2004; see Prendergast, 1999; and Bol, 2008, for reviews of the literature, and Ederhof, Rajan & 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 a worker'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.

Zábojník (2011) models the interaction between an objective and a subjective performance measure. Like us, he assumes that the supervisor has superior information about the worker's ability. Unlike us, the supervisor is the residual claimant. As a result, the supervisor has an incentive to underreport performance. Zábojník (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 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 a worker's ability. In both Zábojník (2011) and Suvorov and van de Ven (2009), having to pay a bonus for good performance make positive feedback credible. In our model, positive feedback is credible because the supervisor wants to show that he is able to assess the performance.

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. Our paper deviates from MacLeod (2003) in that we also examine the effects of performance appraisals on the worker's future performance.

This paper is also related to Prendergast (1993) who shows that when firms use subjective performance evaluations a worker may have an incentive to conform to the opinion of his supervisor. In our model, it is the supervisor rather than the worker 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 worker. 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

2.1 | Background

In an appraisal interview, a worker and a supervisor exchange their views about the worker's performance. An appraisal interview is usually meant to improve the worker's future performance. We model the link between past performance and future performance through the worker's ability. To keep things simple, we model a performance appraisal as an *ex ante* statement about the worker's ability.[12] Implicit is an earlier period in which the worker's performance contains information about the worker's ability, and in turn about the worker's future potential. Under the assumption that the worker is uncertain about his ability, in this setting a performance appraisal may affect a worker's future work effort.

Another important feature of our model is that the worker's payoff depends on the supervisor's view about his performance. We do not explicitly model why a better perceived performance benefits the worker. We simply assume that the worker's payoff depends linearly on his performance as perceived by the supervisor. Essential is that the worker must have confidence in the supervisor's ability to assess his performance accurately. We assume that some supervisors can perfectly assess workers' performances, whereas other supervisors are completely unable to assess workers' performances. In Section 5, we relax this assumption.

2.2 | The formal model

Our model describes the interaction between two risk neutral players: a worker (he) and his supervisor, the supervisor (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 α : $\Pr(a = h) = \alpha$ 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, $\zeta > 0$, the worker's signal is accurate: $s = a$. With probability $(1 - \zeta)$, s does not contain any information about a. In that case, $\Pr(s = h|a = k) = \Pr(a = h) = \alpha \; \forall k \in \{l, h\}$.

There are two types of supervisors: $t = \{b, g\}$, with $\Pr(t = g) = \rho$. A good supervisor, $t = g$, observes both a and y. A bad supervisor, $t = b$, observes neither a nor y. The supervisor knows her type, but the worker does not know the supervisor's type. The supervisor takes two actions. First, before the agent chooses effort, the supervisor sends a message, $m \in \{l, h\}$, about her perception of the worker's ability.[13,14] Second, after the worker has chosen effort, the supervisor assesses the output that the worker has produced. The key feature of our model is that the supervisor'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 supervisor observes his performance. It equals $y - \frac{1}{2}\gamma e^2$ if $t = g$, and equals $c - \frac{1}{2}\gamma e^2$ if $t = b$, where $\gamma > 0$ and c is a constant. The crucial assumption is that the worker's effort yields benefits only if the supervisor is good.[15]

The timing of the game is as follows:

- Nature draws a and t. The supervisor observes t. A good supervisor also observes a.
- The worker receives a signal, s, about a, which is informative with probability ζ and uninformative otherwise.
- The supervisor 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 supervisor observes y. A bad supervisor does not observe y. The payoff to the worker depends on the supervisor's type.

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

Our model is a dynamic game with incomplete information. Based on s and m, the worker updates his beliefs about his own ability and the competence of his supervisor. Let $\mathrm{E}(a \mid s, m)$ be his posterior expected ability and $\hat{\rho}(s, m)$ the posterior probability that the supervisor is good. The effort strategy of the worker maps his signal about his ability and the message he received from the supervisor into an effort level $e(s, m) \in [0, \infty)$. A good supervisor'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 supervisor sends message $m = h$, conditional on a. A bad supervisor's feedback strategy denotes the likelihood $\mu_b \in [0, 1]$ with which a bad supervisor chooses message $m = h$. We focus on equilibria with monotonic beliefs in the sense that (i) by sending $m = l$ the supervisor does not decrease the probability that the worker believes that $a = l$, and by sending $m = h$ the supervisor does not decrease the probability that the worker believes that $a = h$; and (ii) the supervisor 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) $\mathrm{E}(a \mid s, h) \geq \mathrm{E}(a \mid s) \geq \mathrm{E}(a \mid s, l)$ 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 $\mathrm{E}(a \mid s, m)$ as the worker's self-confidence. We identify Perfect Bayesian Equilibria in which (i) given the posterior beliefs [$\mathrm{E}(a \mid s, m)$ and $\hat{\rho}(s, m)$] and feedback strategies of the two types of supervisors, 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 supervisor 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 is on equilibria in which feedback can affect effort.

3 | EQUILIBRIA

We split the analysis into three parts. First, we determine the worker's effort decision, and discuss how this effort decision affects the supervisor's incentives to provide positive or negative feedback. On the basis of the effort decision, we identify four possible effects of feedback. Next, in Section 3.1, we solve the model under the assumption that the worker receives a fully informative signal about his ability. Finally, in Section 3.2, we solve the model for the case that the worker receives a noisy signal about his ability.

Consider the worker's effort decision. The worker's effort results from maximizing $\hat{\rho}(s, m)\mathrm{E}(a|s, m)e - \frac{1}{2}\gamma e^2$ with respect to e, yielding

$$e^*(s, m) = \frac{\hat{\rho}(s, m)\,\mathrm{E}(a \mid s, m)}{\gamma}. \tag{1}$$

The worker's effort depends positively on the probability that the supervisor is good because only good supervisors reward effort. Moreover, the worker's effort depends positively on his perception of his ability, as in our model effort and ability are complements. The supervisor aims at maximizing the worker's expected output, which depends on $e^*(s, m)$, the possible effects of effort on output, l and h, and the probability that $a = h$, α. We now distinguish four possible effects of feedback on the worker's output.

The first effect runs through the effect of the worker's self-confidence, $\mathrm{E}(a \mid s, m)$, on his effort, $e^*(s, m)$. The assumption of monotonic beliefs implies that $\mathrm{E}(a \mid s, h) \geq \mathrm{E}(a \mid s, l)$. This provides the supervisor an incentive to give a positive assessment ($m = h$). We call this the *self-confidence effect*. The other three effects of feedback run through the worker's perception of the supervisor's type, $\hat{\rho}(s, m)$. In equilibrium, the reputation of the supervisor is better when the worker's own signal is consistent with the feedback that he receives. From this perspective in isolation, sending $m = l$ motivates a low-ability worker more than

$m = h$, and a high-ability worker less. To increase the probability of consistent feedback, the supervisor has an incentive to base her feedback on the prior probability that the worker is of the high-ability type, α: send $m = h$ if $\alpha > \frac{1}{2}$ and $m = l$ if $\alpha < \frac{1}{2}$. We call this the *playing the odds effect*.

In equilibrium, the probability that a good superviser gives a particular feedback, say $m = h$, usually differ from the probability that a bad supervisor gives this feedback. This gives bad supervisors an incentive to give feedback that good supervisors give relatively frequently. We call this the *supervisor credibility effect*. We show in the next section that this effect is responsible for the existence of equilibria in mixed strategies. The last effect is the *productivity effect*. To understand this effect, consider the expression of the equilibrium performance of a worker.

$$y(s, m) = \frac{\hat{\rho}(s, m)\mathrm{E}(a|s, m)}{\gamma} a.$$

The expression shows that the credibility of the supervisor matters more for the worker's performance if (i) the worker is more productive and (ii) the worker believes he is more productive. Note that more able workers are also more likely to be self-confident. Therefore, the supervisor has an incentive to protect her credibility with highly productive workers.

The self-confidence effect and the productivity effect explain the leniency bias of supervisors. Both effects hinge on the assumption that effort and ability are complements (see Section 5). The playing the odds effect can strengthen or weaken the leniency bias, depending on whether α is higher or lower than $\frac{1}{2}$. The supervisor credibility effect weakens the dominant inclination of bad supervisors.

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

In this section, we assume that the worker knows his own ability. A direct implication of this assumption is that for the worker the supervisor's message does not contain information about his ability. It only contains information about the supervisor's type. The supervisor chooses a message with an eye on making the worker believe that he can recognize the worker's contribution.

Because the worker knows his ability, his optimal effort (1) reduces to

$$e^*(s, m) = \begin{cases} \frac{1}{\gamma}\hat{\rho}(h, m)h & \text{if } a = h, \text{ and} \\ \frac{1}{\gamma}\hat{\rho}(l, m)l & \text{if } a = l. \end{cases} \tag{2}$$

The assumption of monotonic beliefs implies that it is a dominant strategy for a good supervisor to send $m = s$, $\mu_g^*(h) = 1$ and $\mu_g^*(l) = 0$. Given the equilibrium strategy of a good supervisor, Bayes' rule implies the following posterior beliefs [16]:

$$\begin{aligned} \hat{\rho}(h, l) &= 0, \\ \hat{\rho}(l, h) &= 0, \\ \hat{\rho}(l, l) &= \frac{\rho}{\rho + (1 - \rho)\left(1 - \mu_b^{ant}\right)}, \\ \hat{\rho}(h, h) &= \frac{\rho}{\rho + (1 - \rho)\mu_b^{ant}}, \end{aligned} \tag{3}$$

where μ_b^{ant} is the bad supervisor's feedback strategy as anticipated by the workers.

The posteriors show that $m \neq s$ ruins a supervisor's reputation, whereas $m = s$ improves it. The extent to which a correct message improves the supervisor's reputation depends on the probability with which a bad supervisor sends that message. For instance, if a bad supervisor rarely sends $m = l$ (μ_b close to 1), then $\hat{\rho}(l, l)$ is close to 1. In particular, the higher is μ_b, the lower is $\hat{\rho}(h, h)$, and the higher is $\hat{\rho}(l, l)$.

We have established how much effort the worker exerts in equilibrium, the dominant strategy of a good supervisor, and the posteriors. Now we analyze the optimal response of a bad supervisor given these strategies and beliefs. Using (2) and the posteriors, the bad supervisor's expected payoff is

$$y(m) = \begin{cases} \frac{1}{\gamma}\alpha\frac{\rho}{\rho + (1 - \rho)\mu_b^{ant}}h^2 & \text{if } m = h, \\ \frac{1}{\gamma}(1 - \alpha)\frac{\rho}{\rho + (1 - \rho)\left(1 - \mu_b^{ant}\right)}l^2 & \text{if } m = l. \end{cases} \tag{4}$$

It is easy to verify that always choosing $m = l$ ($\mu_b = 0$) is an equilibrium response of a bad supervisor[17] if

$$(1-\alpha)\rho l^2 \geq \alpha h^2. \tag{5}$$

The productivity effect makes it less likely that this condition holds ($h > l$). The same is true for the supervisor credibility effect ($\rho < 1$). Hence, $\mu_b = 0$ can only be an equilibrium if the playing the odds effect is sufficiently strong toward sending $m = l$, $\alpha < \frac{1}{2}$. As the worker knows his own ability, there is no self-confidence effect.

Similarly, (2) implies that always sending $m = h$ ($\mu_b = 1$) is an equilibrium response of a bad supervisor if

$$\alpha\rho h^2 \geq (1-\alpha) l^2. \tag{6}$$

This condition is more likely to hold due to the productivity effect ($h > l$). Again the supervisor credibility effect works against this condition in favor of mixing ($\rho < 1$). The playing the odds effect makes this type of equilibrium more likely if h type workers are the more common type ($\alpha > \frac{1}{2}$). However, due to the productivity effect, always sending $m = h$ ($\mu_b = 1$) can still be an equilibrium response of a bad supervisor if α is slightly below $\frac{1}{2}$ provided that $\rho h^2 > l^2$.

The supervisor credibility effect is the reason why the bad supervisor may employ a mixed strategy in equilibrium. Such an equilibrium exists if both (5) and (6) are violated. A bad supervisor is indifferent between sending $m = l$ and sending $m = h$ if

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

$$\mu_b = \frac{\alpha h^2 - (1-\alpha)\rho l^2}{(1-\rho)\left(\alpha h^2 + (1-\alpha) l^2\right)}. \tag{7}$$

Due to the productivity effect, μ_b is increasing in h and decreasing in l. Moreover, due to the playing the odds effect μ_b is increasing in α. Finally, we find that 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 tied to the supervisor credibility effect. The supervisor credibility effect favors neither feedback message if $\mu_b = \frac{1}{2}$. This is the case if $\alpha = \frac{l^2}{l^2+h^2}$. For other α, we have that $\mu_b \neq \frac{1}{2}$. For $\mu_b \neq \frac{1}{2}$, the supervisor credibility effect pushes μ_b toward $\frac{1}{2}$. An increase in ρ dampens the supervisor credibility 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 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 the worker knows his ability* ($\zeta = 1$) *and the supervisor is of unknown ability* ($0 < \rho < 1$). *Then on the basis of the bad supervisor's strategy three equilibria can be distinguished:*

- **(i)** *an equilibrium in pure strategies exists, in which a bad supervisor always says that the worker has low ability* ($\mu_b = 0$)*, if and only if* $(1-\alpha)\rho l^2 \geq \alpha h^2$*;*
- **(ii)** *an equilibrium in pure strategies exists, in which a bad supervisor always says that the worker has high ability* ($\mu_b = 1$)*, if and only if* $\alpha h^2 \geq \frac{(1-\alpha) l^2}{\rho}$*;*
- **(iii)** *an equilibrium exists, in which the bad supervisor plays a mixed strategy [with* μ_b *given by equation* (7)*], if and only if* $\frac{(1-\alpha) l^2}{\rho} > a h^2 > (1-\alpha)\rho l^2$.

In equilibria (i–iii), the worker's effort is given by (2)*, the good supervisor's strategy is to be honest in his feedback* ($\mu_g^*(h) = 1$ *and* $\mu_g^*(l) = 0$)*, and the posteriors are given by* (3)*. The equilibrium probability with which a bad supervisor gives positive feedback* ($m = h$) *is nonincreasing in the productivity of a low-ability worker* (l)*, and nondecreasing in both the productivity of a high-ability worker* (h) *and the probability that he is a high-productivity worker* (α).

3.2 | The worker is unsure about his ability, $\zeta < 1$

In the previous section, the self-confidence of the worker was not at stake. The worker knew his own ability. Now we allow him to be uncertain about his ability. This gives the supervisor a reason to lie to him. If she can convince him that he is good, then he exerts more effort (the self-confidence effect). In spite of this, the equilibrium can be informative because a good supervisor

has an incentive to show her competence. In this section, we characterize such equilibria. But first we show that informative equilibria exist only if the worker is unsure about the supervisor's type.

Proposition 2. *Suppose that in the feedback model the worker is not certain about his ability* $(0 < \zeta < 1)$*, but that he knows whether his supervisor can assess him* $(\rho \in \{0, 1\})$*. Then no informative equilibrium exists, that is, in any equilibrium:*

- **(i)** *The feedback of a good supervisor does not depend on the ability of the worker:* $\mu_g^*(h) = \mu_g^*(l)$*;*
- **(ii)** *The worker's effort is given by* (2)*;*
- **(iii)** *The posteriors are equal to their priors.*

The intuition behind Proposition 2 is straightforward. If the credibility of the supervisor is not at stake, the only effect of feedback is the self-confidence effect. As the supervisor always wants the worker to have a positive perception of himself, she has a strict preference for giving positive feedback. As a result, the supervisor's message does not contain any information about the worker's type. Bol (2011) presents evidence that supervisors are more inclined to provide biased, positive feedback to workers when their relationships are more long-lasting. It seems plausible that when time elapses uncertainty about a supervisor's ability decreases. Against this background, Bol's finding is consistent with Proposition 2.

Having established equilibrium behavior for $\rho \in \{0, 1\}$, Proposition 3 describes equilibrium behavior for $0 < \rho < 1$.

Proposition 3. *Consider the feedback model where the worker is uncertain whether his supervisor can assess him* $(0 < \rho < 1)$*. Then, for any equilibrium in which effort depends on the feedback for some private information of the worker* (*so* $e^*(s, m = l) \neq e^*(s, m = h)$ *for some* $s \in \{l, h\}$)*, we have:*

- **(I)** *A bad supervisor gives positive feedback with positive probability* $(\mu_b > 0)$*. A good supervisor who observes a high-ability worker always gives positive feedback* $(\mu_g^*(h) = 1)$*. A good supervisor who observes a low-ability worker does not always give positive feedback* $(\mu_g(l) < 1)$*, and always gives negative feedback if a bad supervisor mixes* $(\mu_g(l) = 0$ *if* $\mu_b < 1)$*.*
- **(II)** *The worker's effort is given by* (2)*;*
- **(III)** *Positive feedback boosts a worker's self-confidence:* $E(a \mid s, h) > E(a \mid s, l)\ \forall s \in \{h, l\}$*;*
- **(IV)** $\hat{\rho}(l, l) \geq \hat{\rho}(h, h)$ *if and only if* $\mu_b^* \geq \frac{(\zeta + (1-\zeta)\alpha)}{(1+\zeta)}$ *and* $\hat{\rho}(h, l) \geq \hat{\rho}(l, h)$ *if and only if* $\mu_b^* \geq \alpha$*.*

Proof. See Appendix A. □

Proposition 3 presents a wide variety of results. First, consider Part I. It says that a good supervisor, who believes to face 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 supervisor, who believes to deal with 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 supervisor. Finally, Part I of Proposition 3 shows that a bad supervisor has weaker incentives to provide positive feedback than a good supervisor who faces a high-ability worker, but stronger incentives than a good supervisor 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 supervisor knows that the worker is of the high-ability type, and minimal if the supervisor knows that the worker is of the low-ability type.

Parts III and IV of Proposition 3 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 supervisor. As in Section 3.1, the sign of the supervisor credibility effect depends on the probability with which a bad supervisor gives positive feedback. If a bad supervisor predominantly provides positive (negative) feedback, providing negative (positive) feedback signals being a good supervisor. Together with equation (2), Part II of Proposition 3 shows how the worker's self-confidence and his confidence in the supervisor determine effort.

To gain deeper insight into the variety of effects of feedback, suppose that $\alpha = \frac{1}{2}$. In that case, the playing the odds effect is canceled out. The productivity effect and the self-confidence effect give incentives to a bad supervisor to provide positive feedback. The supervisor credibility 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 supervisor, facing a high-ability worker, always provides positive feedback, our model is able to explain the widely observed leniency bias: in general, supervisors tend to provide positive feedback. For $\alpha > \frac{1}{2}$, bad supervisors 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 supervisors may lean to negative feedback.

Our model highlights the importance of the interplay of the worker's self-perception and his perception of the supervisor's ability to assess his performance correctly. Negative feedback discourages a worker who thinks highly of himself. Such a worker

would dismiss a supervisor who provides negative feedback as being incompetent. Feedback that is consistent with the worker's self-perception enhances the worker's confidence in the supervisor'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 credibility in assessing the worker's performance correctly.

4 | THE CENTRALITY BIAS

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

In this section, we extend the basic feedback model of Section 3.1, where the worker knows his ability, 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 show later, in the extended model another kind of the playing-the-odds effect arises. In line with having three ability types, we allow the supervisor to send three messages, $m \in \{l, n, h\}$. The second extension is that we relax the assumption that a good supervisor always observes the worker's ability correctly. In this section, we assume that a good supervisor receives a partially informative signal about the worker's ability, $s_g = \{l, n, h\}$. A good supervisor may make small mistakes, but never makes big mistakes. Specifically, we assume that

$$\begin{cases} \Pr(s_g = l|a = l) = \Pr(s_g = h|a = h) = \pi; \\ \Pr(s_g = n|a = l) = \Pr(s_g = n|a = h) = 1 - \pi; \\ \Pr(s_g = n|a = n) = 2\pi - 1; \\ \Pr(s_g = h|a = n) = \Pr(s_g = l|a = n) = 1 - \pi; \end{cases}$$

where $\pi \in (\frac{2}{3}, 1)$. The lower bound of π ensures that, given s_g, the most likely ability level of the worker corresponds to s_g. As in the basic model, we assume that a bad supervisor does not receive an informative 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 supervisor, but not by a bad supervisor.

The main objective of this section is twofold. First, we want to show that equilibria of the extended feedback model exist in which bad supervisors 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 do not present a full analysis of the extended model, but focus on a few interesting outcomes.

As before, a Perfect Bayesian Equilibrium consists of a set of beliefs , a strategy of the worker, a strategy of the bad supervisor, and a strategy of the good supervisor. 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 supervisor's ability. To minimize on notation, we assume that $\gamma = 1$. Concerning the good supervisor, we focus on equilibria in which the good supervisor honestly reveals her signal. The focus of the analysis is therefore on the strategy of the bad supervisor, $\theta_b^k = \Pr(m = k|t = b)$. In English, what are the probabilities that in equilibrium a bad supervisor gives positive, neutral, and negative feedback?

When a good supervisor always honestly reveals her signal, the posterior probabilities that a supervisor is good are[18]

$$\hat{\rho}(l, l) = \frac{\rho\pi}{\rho\pi + (1 - \rho)\theta_b^l},$$

$$\hat{\rho}(n, l) = \frac{\rho(1 - \pi)}{\rho(1 - \pi) + (1 - \rho)\theta_b^l},$$

$$\hat{\rho}(h, l) = 0 = \hat{\rho}(l, h),$$

$$\hat{\rho}(l,n) = \frac{\rho(1-\pi)}{\rho(1-\pi)+(1-\rho)\theta_b^n} = \hat{\rho}(h,n),$$

$$\hat{\rho}(n,n) = \frac{\rho(2\pi-1)}{\rho(2\pi-1)+(1-\rho)\theta_b^n},$$

$$\hat{\rho}(n,h) = \frac{\rho(1-\pi)}{\rho(1-\pi)+(1-\rho)\theta_b^h},$$

$$\hat{\rho}(h,h) = \frac{\rho\pi}{\rho\pi+(1-\rho)\theta_b^h}.$$

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 supervisor an incentive to try to provide feedback that is consistent with the worker's self-perception. Second, $\hat{\rho}(a,m)$ is decreasing in θ_b^m. This feature is responsible for the supervisor credibility effect. As a result of this effect, in equilibrium bad supervisors may follow mixed strategies.

Now consider the problem a bad supervisor faces. For the three alternative messages, her expected payoff equals

$$\begin{cases} \frac{1}{3}\left[\hat{\rho}(l,l)l^2 + \hat{\rho}(n,l)n^2\right] & \text{if } m = l, \\ \frac{1}{3}\left[\hat{\rho}(l,n)l^2 + \hat{\rho}\,(n,n)\,n^2 + \hat{\rho}(h,n)h^2\right] & \text{if } m = n, \\ \frac{1}{3}\left[\hat{\rho}(n,h)n^2 + \hat{\rho}(h,h)h^2\right] & \text{if } m = h. \end{cases} \tag{8}$$

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

Proposition 4. *Consider an equilibrium of the extended feedback model in which a good supervisor is honest* ($m = s^g$) *and a bad supervisor follows a pure strategy. Then, the bad supervisor sends always* $m = n$ ($\theta_b^n = 1$). *An equilibrium in pure strategies exists for a range of parameters.*

Proof. See Appendix B. □

Proposition 4 is a direct implication of the feature of our model that good supervisors may make small errors but never make large errors. By giving neutral feedback, a bad supervisor never ruins her reputation. Feedback $m = n$ is a safe haven. Proposition 4 can also be interpreted as a variation of the playing-the-odds effect. In Section 3, the playing the odds effect reflects a bad supervisor's inclination to send a message that is likely correct. In the present model, a bad supervisor has an incentive to send $m = n$ to avoid that his message is completely wrong.

Proposition 4 illustrates that bad supervisors may have an incentive to avoid extreme ratings if better supervisors are relatively less likely to make large mistakes. In the model, there are two forces that may drive bad supervisors away from sending $m = n$. These forces emerge from the productivity and supervisor credibility effect. First, with $\theta_b^n = 1$, $m \neq n$ boosts the worker's confidence in the supervisor unless m is the opposite of a. This gives a bad supervisor an incentive to deviate from $m = n$. The lower is the probability that a supervisor is good, the stronger is the supervisor credibility effect. For this reason, $\theta_b^n = 1$ can only be part of an equilibrium if ρ is high. Second, the productivity effect reflects that it is especially important that higher ability workers expend effort. This gives bad supervisors an incentive to send $m = h$ rather than $m = l$. Note that if in equilibrium $\theta_b^h > 0$, the supervisor credibility effect favors sending $m = l$. For (i) low values of ρ, and (ii) l, n, and h close to each other, an equilibrium exists in which a bad supervisor sends all messages with positive probability. Due to the productivity effect, we must have that $\theta_b^h > \theta_b^l$ in those equilibria.

5 | EXTENSIONS

In the analysis above, we made several assumptions in order to keep the analysis as simple as possible, while still getting the main results across. In this section, we discuss several variations of our model to examine the robustness of our results and to study how the nature of feedback and the responses to feedback depend on the environment.

5.1 | Moderately able supervisor types

We first relax our assumption that a good supervisor perfectly observes the worker's performance, whereas a bad supervisor has no clue. Consider our basic model, but with two other supervisor types: Good (G) and Bad (B), $t \in \{G, B\}$.[19] Suppose supervisor of type t receives signal ω on a, $\omega \in \{l, h\}$, which is correct with probability κ_t and uninformative with probability $(1 - \kappa_t)$, where $1 > \kappa_G > \kappa_B > 0$. Like in the basic model, κ_t is also the probability that a supervisor of type t recognizes the worker's contribution to the firm. Let $\mu_t(\omega)$ be the probability that the supervisor of type t gives feedback $m = h$ if she observes signal ω. We assume that the supervisor knows her type, but does not know whether her signal is correct. With this modification, the optimal effort of the worker becomes

$$e^*(s, m) = \frac{(\kappa_B + \hat{\rho}(s, m)(\kappa_G - \kappa_B))\mathrm{E}(a|s, m)}{\gamma}. \tag{9}$$

Note that for $\kappa_B = 0$ and $\kappa_G = 1$, (9) simplifies to (1). The expression above shows that as in the basic model, effort increases in $\mathrm{E}(a \mid s, m)$ and $\hat{\rho}(s, m)$. The role of $\mathrm{E}(a|s, m)$ is almost identical as in the basic model. It leads to the self-confidence effect. To better focus on the effects of feedback through $\hat{\rho}(s, m)$ we eliminate the effect of feedback through $\mathrm{E}(a|s, m)$ by assuming that the worker knows his type.

Relaxing the assumption that the supervisor's signal is either fully informative or fully uninformative has three consequences. First, the effects of feedback are affected by the difference of the quality of the two types of supervisors, $(\kappa_G - \kappa_B)$. This mitigates the effects of feedback through $\hat{\rho}(s, m)$. Second, the bad supervisor may now condition her feedback on her signal. Third, inconsistent feedback, $s \neq m$, does not completely ruin a supervisor's reputation. The reason is that in the present model also good supervisors err. As a result of the last two consequences, the feedback itself rather than whether or not it is consistent becomes important. To see this, suppose an equilibrium where a bad supervisor, irrespective of her type, always sends $m = h$. In such an equilibrium, $\rho(s, l) = 1$. This gives a strong incentive to each kind of supervisor, irrespective of her type, to send $m = l$. One can show that as a result of these incentives in any equilibrium, where at least some information is conveyed from the supervisor to the worker, the good supervisor always follows her signal, whereas the feedback of the bad supervisor always contains at least some information. Against this background, it is hardly surprising that for ranges of parameter values equilibria exist in which both good and bad supervisors always reveal their signals.[20]

All in all, relaxing the assumption that good supervisors receive fully informative signals, whereas bad supervisor receive fully uninformative signals leads to smaller effects of feedback, but strengthen supervisors' incentives to follow their signals.

5.2 | Additive production function

We now again turn to the case where $\kappa_G = 1$ and $\kappa_B = 0$, but assume an additive production function $y(a, e) = a + e$. In the resulting model, the worker chooses an effort level equal to $e^*(s, m) = \frac{1}{\gamma}\hat{\rho}(s, m)$. The main difference with the basic model is that $\mathrm{E}(a|s, m)$ does not affect the worker's effort choice. A direct implication is that the self-confidence effect disappears.

Another implication of the additive production function is that the productivity effect is no longer present. Guessing correctly is not more important for more productive workers in the present variation. The supervisor's feedback, therefore, solely focuses on building a reputation for being a good supervisor. In equilibrium, a bad supervisor is mainly driven by the playing-the-odds effect. If $\alpha = \frac{1}{2}$, a bad supervisor chooses $m = l$ and $m = h$ with equal probability.[21] If $\alpha > \frac{1}{2}$ ($\alpha < \frac{1}{2}$), a bad supervisor is biased toward $m = h$ ($m = l$). The supervisor credibility effect reduces the bias, but not entirely.

5.3 | Performance targets

In this section, we change the reward structure. Instead of a reward which is strictly increasing in the (reported) performance, now the worker receives some fixed reward if and only if the supervisor reports that he has met his target, ψ. In this setting, ability and effort are strategic substitutes. A worker who has low ability will need to put in more effort to reach the threshold than a worker who has high ability. In the main model above, ability and effort were strategic complements. Therefore, this reward structure allows us to see whether our previous results are specific to settings with a strategic complementarity of effort and ability.

Formally, consider the setting of Section 3.2 with the following changes. Let the production function be $y(a, e)$, where y increases both in a and e. The worker gets bonus b if his supervisor reports that $y > \psi$. The supervisor wants to maximize expected output y. The worker maximizes $\mathrm{E}U(e) = \Pr(y(a, e) > \psi)b - c(e)$, where $c(e)$ represents the cost of effort, $c' > 0$.

Define e_l and e_h as the minimum effort levels which a worker with, respectively, l and h needs to meet threshold ψ. Let c_l and c_h be the respective cost of effort associated with e_l and e_h. Note that $e_l > e_h$ and thus $c_l > c_h$. We consider the case where $b > c_l$. This implies that any worker is willing to exert effort if he believes that the supervisor is competent.

Because the worker knows that he has either ability l or ability h, his optimal effort is 0, e_h, or e_l. His expected utility is given by

$$\mathrm{E}U(e|s,m) = \begin{cases} 0 & \text{if } e = 0, \\ \hat{\rho}(s,m)\Pr(a = h|s,m)b - c_h & \text{if } e = e_h, \\ \hat{\rho}(s,m)b - c_l & \text{if } e = e_l. \end{cases} \tag{10}$$

The question which effort level the worker chooses yields some surprising answers. Intuitively, we expect that in case of strategic substitutes the self-confidence effect favors $m = l$. The reason is that workers with lower self-confidence tend to work harder ($e_l > e_h$), and workers are more likely to choose $e = e_l$ if $\Pr(a = h|s,m)$ is lower. Thus, the self-confidence effect depends crucially on whether ability and effort are complements or substitutes. However, (10) shows two caveats to this story. First, the direction of the self-confidence effect depends on which effort is likely to be chosen if not e_h. When c_l is high, $e = 0$ is optimal if $e \neq e_h$. In that case, a supervisor would like to boost the worker's self-confidence. The alternative would be getting no effort. Second, the self-confidence effect may be irrelevant. $\Pr(a = h|s,m)$ affects only the worker's expected utility if the worker chooses e_h. If for each combination of s and m the relevant choice is between $e = 0$ and $e = e_l$, then $\Pr(a = h|s,m)$ is irrelevant.

Reputation concerns remain important, however. Equation (10) shows that the better the reputation of the supervisor, the more willing a worker is to exert effort. As before, this concern allows for informative equilibria.[22] The playing the odds effect and supervisor credibility effect are similar to our basic case. Feedback $m = h$ becomes more attractive the more common highly able workers are and the less $m = h$ is associated with bad supervisors. The productivity effect, however, is weaker in the current case with strategic substitutes. In the basic model, it is more important to motivate the high-ability worker for two reasons. First, he is more productive. Second, his private signal tends to reinforce his self-confidence, causing him to work harder. The latter effect is reversed in case of strategic substitutes. When the worker assigns a higher probability to the event that he is able, he works less hard. Therefore, the productivity effect will tend to favor $m = h$ less if effort and ability are substitutes.

Summarizing, the reputational constraints also allow for informative equilibria in case of strategic substitutes. The differences are in the self-confidence and productivity effects. Both tend to favor $m = h$ less strongly, or even favor $m = l$.

5.4 | A richer worker type and message space

We are aware that in practice firms usually do not rate workers on two-point scales. Three- and five-point scales are much more common. We now show that the results from the basic model carry over to the case where there are three types of workers. Consider the same setting as Section 3.1 with the following changes. There are three types of workers, $a \in \{l, n, h\}$ and three types of messages: $m = \{l, n, h\}$, with $0 < l < n < h$, and $\alpha_k = \Pr(a = k)$ for $k \in \{l, n, h\}$. The worker knows his ability. In the modified model, as in the basic model, there is an equilibrium in which the worker believes that his supervisor is bad, unless the supervisor gives accurate feedback. In such an equilibrium, it is in the interest of a good supervisor to report her signal honestly: $m = a$. Now consider the bad supervisor. She is in a worse situation than in the model with two types. Because there are more worker types, it is harder for her to guess the worker's type correctly. This is especially problematic when both the good supervisor and the worker know the worker's type. The good supervisor, by contrast, benefits from the additional worker type, as her feedback tends to enhance her reputation relatively more.

In equilibrium, a bad supervisor adopts the pure strategy to always send $m = l$ if and only if

$$\alpha_l \rho l^2 \geq \alpha_n n^2 \quad \text{and} \quad \alpha_l \rho l^2 \geq \alpha_h h^2.$$

Similarly, she adopts the pure strategy to always send $m = n$ ($m = h$), if and only if

$$\alpha_n \rho n^2 \geq \alpha_l l^2 \quad \text{and} \quad \alpha_n \rho n^2 \geq \alpha_h h^2$$
$$\left(\alpha_h \rho h^2 \geq \alpha_l l^2 \quad \text{and} \quad \alpha_h \rho h^2 \geq \alpha_n n^2\right).$$

The range of parameters for which in equilibrium $m = h$ is wider than that range for which $m = n$, which in turn is higher than that range for which $m = l$. If none of the above conditions hold, the bad supervisor follows a mixed strategy in equilibrium.

6 | CONCLUSIONS

In this paper, we have investigated how a supervisor's performance appraisals affect a worker's future performance. A key feature of our model is that both the supervisor and the worker form a perception of the worker's past performance. We have derived several results. First, we have shown that cheap-talk statements about the worker's performance may contain information that is relevant for the worker. Second, for a wide range of parameters the supervisor tends to give positive appraisals. Third, on average, a positive appraisal motivates a worker more than a negative appraisal. Fourth, the effect of appraisals on a worker's future performance depends on how it affects the supervisor's reputation for being able to assess a worker's performance. Finally, our analysis suggests an explanation for the centrality bias.

In our model, there are two main forces at work. First, workers want the fruits of their efforts to be observed. When a worker's supervisor is bad, that is, she is not able to distinguish good work from bad work, the worker's incentives to expend effort are totally absent. This force has been the main focus of Section 3.1. A bad supervisor demotivates both high ability and low-ability workers, as their efforts are to no avail. It is this force that leads to the prediction that negative feedback to a low-ability worker may increase this worker's incentives. More generally, our model predicts that feedback that is inconsistent with the worker's own perception damages the worker's confidence in the supervisor's ability and thereby weakens his incentive to expend effort. The second force is that feedback may contain information about the worker's ability. This force has been introduced in Section 3.2. This force gives incentives to supervisors to give positive feedback. It explains the leniency bias. An alternative explanation for the leniency bias is that supervisors find it unpleasant to give negative feedback. We tend to regard such supervisors as special cases of bad supervisors. Once workers have discovered that their supervisors do not condition feedback on their performance, workers' incentives to expend effort weaken. Good supervisors have incentives to distinguish themselves from supervisors who shy away from negative feedback. Bol (2011) finds that strong supervisor–worker relationships positively affect the leniency bias. In case of strong supervisor–worker relationships, supervisors may find it particularly unpleasant to give negative feedback. Our model predicts that positive feedback has smaller effects on subsequent performance the stronger the supervisor–worker relationship is.

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 example, we have assumed two types of supervisors, one receiving a fully informative signal, and one receiving a fully uninformative signal. In the previous section, we have shown that these assumptions are innocuous. Relaxing them only qualitatively affects our main results. In the previous section, we also showed that the assumption that effort and ability are complements is far less innocuous. Relaxing that assumption may change the direction of some of the feedback effects.

Another crucial assumption underlying our analysis is that the supervisor is not the residual claimant. When the supervisor is a residual claimant, the worker may wonder whether good performance ever leads to a reward. Once effort has been exerted, the supervisor has no incentive to reward the worker. It is well-known that a reputation for keeping promises may help the supervisor to solve this kind of moral hazard problems. The resulting analysis would require a dynamic setting. This would also allow for analyzing long working relationships between supervisors and workers. We have shown that in a static model performance appraisals only matter when the worker is uncertain about the supervisor's type. As a dynamic setting allows for learning, we expect that the effects of performance appraisals diminish over time.

NOTES

[1] The leniency bias is a well-known. Medoff and Abraham (1980) report that of 7,000 performance ratings 95% were in just two categories: Good and Outstanding (see also Jawahar & Williams, 1997; Prendergast, 1999).

[2] In terms of a five-point scale, the centrality bias means that the ratings 1 and 5 are rarely used. In the 1980s, at Merck, 97% of workers were offered ratings of 3 or 4 (Prendergast, 1999). Moers (2005) finds that performance ratings on subjective dimensions are closer to the median rating than performance ratings on objective dimensions.

[3] Bol (2008) contributes to the business literature showing how performance evaluations affect employees' future performances (see, for example, Alvero, Bucklin, & Austin, 2001; Balcazar, Hopkins, & Suarez, 1986; Kluger & DeNisi, 1996). This literature finds that feedback can both 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.

[4] Many scholars emphasize the importance of the supervisor's credibility (see, e.g., Early, 1986; Lawler, 1971, Longenecker, 1997; Meyer, 1975). Gibbs, Merchant, van der Stede, and Vargus (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." [p. 415; emphasis added].

[5] This paper also contributes to the principal–agent literature in general. Our paper shows that uncertainty about the principal's ability to assess the agent can affect a principal's communication to the agent and the agent's subsequent effort.

[6] This version is available at the authors upon request.

[7] In our model, assuming that the employee receives an inconclusive 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, Ackerman, Beier, & Bowen, 2002; Baumeister, 1998; Klar, Medding, & Sarel, 1996; Kruger, 1999; Sedikides & Strube, 1995). Of course, employees are more likely to know their ability the longer they have been doing their task, and the better they can compare their own performance and effort to those of others with similar tasks. By assuming a conclusive signal, as we do in Section 3.1, we also study the complementary case where the employee knows his ability perfectly.

[8] In our model, effort is not or cannot be contracted upon.

[9] There is increasing evidence that workers care for recognition. See, for example, Bradler, Dur, Neckermann, and Non (2016), Grant and Gino (2010), and Stajkovic and Luthans (2003).

[10] In Bénabou and Tirole (2003), talk to motivate employees does not work. It is a dominant strategy for the manager (supervisor) to send the message that boosts the employee's self-confidence. In our model, the supervisor'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, e.g., Suurmond, Swank, & 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.

[11] 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.

[12] In this paper, the supervisor gives feedback during appraisal interviews. Also most of the empirical literature on feedback focuses on these formal meetings. In practice, supervisors often talk to their employees during the whole year. Also, in these talks, the credibility of supervisors is at stake.

[13] In our model, no supervisor wants to reveal himself as a bad supervisor. Therefore, allowing for a richer message space (with more cheap-talk messages than worker types) does not affect the results.

[14] To keep things simple, we assume that the initial appraisal is not tied to a reward. Allowing for a reward at this stage does not change the results. The reason is that all other decisions in our model are made after the initial appraisal, while the supervisor does not pay the reward herself. In the conclusions, we discuss why our main findings do not carry over to a setting where the supervisor is a residual claimant.

[15] In case of a reward scheme which is linear in performance, it is innocent to focus on a deterministic production function. To see this, consider a stochastic performance function: $y(a,e) = ae + \varepsilon$, where ε is a stochastic term independent of e. This stochastic term does not affect the optimal effort of the worker, because the benefits to the worker are linear in y, and e does not affect ε. Consequently, the term ε does not affect the optimal feedback message of the supervisor.

For reward schemes which are nonlinear in performance, the inclusion of a stochastic term can affect the optimal effort and, therefore, the feedback strategy.

[16] 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 supervisor rather than to a good supervisor.

[17] Note that in such an equilibrium $\mu_b^{ant} = \mu_b = 0$. Thus, (4) becomes $(1-\alpha)\rho\frac{l^2}{\gamma}$. By sending $m = h$ if $\mu_b^{ant} = 0$, the bad supervisor 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 $\frac{h^2}{\gamma}$. Consequently, a bad supervisor would prefer $m = h$ unless $(1-\alpha)\rho\frac{l^2}{\gamma} > \alpha\frac{h^2}{\gamma}$.

[18] In case, $\theta_b^k = 1$, we assume $\hat{\rho}(h,l) = 0 = \hat{\rho}(l,h)$ as out-of-equilibrium beliefs.

[19] Let "good" and "bad" (without capitals) refer to the supervisor types in the basic model, whereas "Good" and "Bad" refer to the types in this section.

[20] For example, an equilibrium in which both supervisor types send $m = h$ if and only if $\omega = h$ exists for $\alpha = \rho = 0.5$, $\kappa_B = 0.4$, $\kappa_G = 0.45$, $l = 2$, $h = 3$, and $\gamma = 1$.

[21] Let $\gamma = 1$ and $\alpha = \zeta = \rho = \frac{1}{2}$. Then the following is an equilibrium. The good supervisor sends honest feedback ($m = a$) and the bad supervisor sends positive feedback ($m = h$) with probability $\frac{1}{2}$. The worker then chooses effort equal to $\hat{\rho}(s,m)$, where $\hat{\rho}(s, m = s) = 0.6$ and $\hat{\rho}(s, m \neq s) = \frac{1}{3}$.

[22] Let $\zeta = \alpha = \rho = \frac{1}{2}$, so $\Pr(s = h|a = h) = \frac{3}{4} = \Pr(s = l|a = l)$. Let $y(a,e) = ae$, $h = 2$, and $l = 1$. Let $\psi = 8$, so $e_l = 4$ and $e_h = 2$. Let $b = 4$, $c_l = c(4) = 3$, and $c_h = c(2) = 1$. Then the following is an equilibrium. The good supervisor sends $m = a$, and the bad supervisor sends $m = h$ with probability $\frac{3}{4}$. The worker chooses $e = e_h$ if $s = m = h$, and $e = 0$ if $s \neq m$. If $s = m = l$, the worker chooses $e = e_l$ with probability 0.7 and $e = 0$ with the remaining probability.

ORCID

Jurjen J.A. Kamphorst http://orcid.org/0000-0001-8079-1591

REFERENCES

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

Alvero, A. M., Bucklin, B. R., & Austin, J. (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.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics*, *109*, 1125–1156.

Balcazar, F., Hopkins, B. L., & Suarez, Y. (1986). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, *7*(3/4), 65–89.

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 680–740). New York: McGraw-Hill.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, *70*, 489–520.

Bol, J. C. (2008). Subjectivity in compensation contracting. *Journal of Accounting Literature*, *27*, 1–32.

Bol, J. C. (2011). The determinants and performance effects of manager's performance evaluation biases. *Accounting Review*, *85*(5), 1549–1575.

Bradler, C., Dur, R., Neckermann, S., & Non, A. (2016). Employee recognition and performance: A field experiment. *Management Science*, *62*(11), 3085–3099.

Bull, C. (1987). The existence of self-enforcing wage contracts. *Quarterly Journal of Economics*, *102*, 147–160.

Crutzen, B. S. Y., Swank, O. H., & Visser, B. (2007). *Confidence management: On interpersonal comparisons in teams*. Tinbergen Institute Discussion Papers 07-040/1, Tinbergen Institute.

Daley, B., & Green, B. (2012). Market signalling with grades. Mimeo.

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.

Ederhof, M., Rajan, M. V., & Reichelstein, S. J. (2011). Discretion in managerial bonus pools. *Foundations and Trends in Accounting*, *5*(4), 243–316.

Gibbs, M., Merchant, K. A., van der Stede, W. A., & Vargus, M. E. (2004). Determinants and effects of subjectivity in incentives. *Accounting Review*, *79*(2), 409–436.

Grant, A. M., & Gino, F. (2010). A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology*, *98*(6), 946–955.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50*, 905–926.

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

Kluger, A. N., & DeNisi, A. (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.

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.

Lawler, E. E. (1971). *Pay and organizational effectiveness: A psychological view*. New York: McGraw-Hill.

Longenecker, C. O. (1997). Why managerial performance appraisals are ineffective: Causes and lessons. *Career Development International*, *2*(5), 212–218.

MacLeod, W. B. (2003). Optimal contracting with subjective evaluation. *American Economic Review*, *93*(1), 216–240.

Medoff, J., & Abraham, K. (1980). Experience, performance, and earnings. *Quarterly Journal of Economics*, *95*(4), 703–736.

Meyer, H. H. (1975). The pay-for-performance dilemma. *Organizational Dynamics*, *3*(3), 39–50.

Moers, F. (2005). Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society*, *30*, 67–80.

Napier, N. K., & Latham, G. P. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology*, *39*, 827–837.

Prendergast, C. (1993). A theory of yes men. *American Economic Review*, *83*(4), 757–770.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7–63.

Prendergast, C., & Topel, R. H. (1993). Discretion and bias in performance evaluation. *European Economic Review*, *37*(2–3), 355–365.

Prendergast, C., & Topel, R. H. (1996). Favoritism in organizations. *Journal of Political Economy*, *104*(5), 958–978.

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

Stajkovic, A. D., & Luthans, F. (2003). Behavioral management and task performance in organizations: Conceptual background, meta-analysis, and test of alternative models. *Personnel Psychology*, *56*(1), 155–194.

Steelman, L. A., & Rutkoswki, K. A. (2004). Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology*, *19*(1), 6–18.

Suurmond, G., Swank, O. H., & Visser, B. (2004). On the bad reputation of reputational concerns. *Journal of Public Economics*, *88*, 2817–2838.

Suvorov, A., & van de Ven, J. (2009). Discretionary rewards as a feedback mechanism. *Games and Economic Behavior*, *67*, 665–681.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2001). Relationships between attitudes towards organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment*, *9*(3), 226–239.

Zábojník, J. (2011). *Subjective evaluations with performance feedback*. Queen's Economics Department Working Paper, 1283, Queen's University.

How to cite this article: Kamphorst JJA, Swank OH. The role of performance appraisals in motivating employees. *J Econ Manage Strat*. 2018;1–19. https://doi.org/10.1111/jems.12241

APPENDIX A: PROOF OF PROPOSITION 3

We prove each part of Proposition 3 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 monotonic beliefs—which we assume throughout the paper—implies $\mu_g^*(l) \leq \mu_g^*(h)$. Then, we show that it is better for the supervisor 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 supervisor 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. Let $\hat{\alpha}(s,m) = \Pr(a = h|s,m)$ be the posterior belief of the worker on the probability that he has high ability. By our assumption of monotonic beliefs, we have $\hat{\alpha}(s,h) \geq \hat{\alpha}(s,l)$ for all $s \in \{h,l\}$. We now show that $\hat{\alpha}(s,h) \geq \hat{\alpha}(s,l)$ implies $\mu_g^*(h) \geq \mu_g^*(l)$.

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

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

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

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

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

$$\left(\rho\mu_g^*(h) + (1-\rho)\,\mu_b^*\right)\left(\rho\left(1-\mu_g^*(l)\right) + (1-\rho)\left(1-\mu_b^*\right)\right) \geq \left(\rho\left(1-\mu_g^*(h)\right) + (1-\rho)\left(1-\mu_b^*\right)\right)\left(\rho\mu_g^*(l) + (1-\rho)\mu_b^*\right),$$

$$\rho\left(\mu_g^*(h) - \mu_g^*(l)\right)\left((1-\rho)\left(1-\mu_b^*\right) + (1-\rho)\mu_b^* + \rho\right) \geq 0,$$

$$\left(\mu_g^*(h) - \mu_g^*(l)\right) \geq 0,$$

and the result follows for $s = h$. The same steps 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 supervisor wants to match the private signal of the worker.

Lemma 2. *Consider a nonbabbling 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 $m = k$, whereas $m = k'$ can never lead to a worse result. Thus any supervisor 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 supervisors, 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 supervisor is willing to send $m = l$.

Let η denote the probability that the worker assigns to $a = h$ after he has received signal $s = h$, so $\eta = \Pr(a = h|s = h) = \zeta + (1-\zeta)\alpha$ 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 $s = l$, so $\lambda = \Pr(a = l|s = l) = \zeta + (1-\zeta)(1-\alpha)$ and $1 - \lambda = \Pr(a = h|s = l)$. Note that $\alpha = \Pr(s = h)$, $\eta = \Pr(a = h|s = h) = \Pr(s = h|a = h)$ and $\lambda = \Pr(a = l|s = l) = \Pr(s = l|a = l)$. The bad supervisor is willing to send $m = l$ only if

$$\left\{ \begin{array}{c} \Pr(s=l)e^*(s=l,m=l)\mathrm{E}(a|s=l) + \\ \Pr(s=h)e^*(s=h,m=l)\mathrm{E}(a|s=h) \end{array} \right\} \geq \left\{ \begin{array}{c} \Pr(s=l)e^*(s=l,m=h)\mathrm{E}(a|s=l) + \\ \Pr(s=h)e^*(s=h,m=h)\mathrm{E}(a|s=h) \end{array} \right\}$$

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

$$(1-\alpha)\mathrm{E}(a|s=l)(e^*(l,l) - e^*(l,h)) \geq \alpha\mathrm{E}(a|s=h)(e^*(h,h) - e^*(h,l)). \tag{11}$$

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

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

$$(1-\eta)(e^*(l,l) - e^*(l,h)) \geq \eta(e^*(h,h) - e^*(h,l)). \tag{12}$$

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

$$\lambda(e^*(l,l) - e^*(l,h)) \geq (1-\lambda)(e^*(h,h) - e^*(h,l)). \tag{13}$$

The bad supervisor is willing to adopt a mixed strategy if and only if (11) holds with equality. If (11) is violated, the bad supervisor strictly prefers to send $m = h$. The same applies for a good supervisor with respect to (12) if $a = h$ and with respect to (13) if $a = l$.

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

Lemma 3. *Consider a nonbabbling 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 supervisor is bad, whenever he observes a message which cannot be observed from a good supervisor. This would constitute a babbling equilibrium: a contradiction. □

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

Lemma 4. *Consider a nonbabbling 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 supervisor facing a high-ability worker strictly prefers $m = h$ if the bad supervisor is willing to send message $m = h$. Suppose not. Then, (11) either holds with equality or is violated while (12) holds. By Lemma 3, this implies that

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

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

Combining yields

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

Note that

$$\mathrm{E}(a|s=l) = l + (1-\zeta)\alpha(h-l),$$

$$\mathrm{E}(a|s=h) = l + (\zeta + (1-\zeta)\alpha)(h-l),$$

$$\eta = \zeta + (1-\zeta)\alpha,$$

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 $\mu_b^* > 0$, then $\mu_g^*(h) = 1$. □

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

Lemma 5. *Consider a nonbabbling 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, (11) holds while (13) is either violated or satisfied with equality. Thus

$$\frac{\lambda}{1-\lambda} \leq \frac{(e^*(h,h)-e^*(h,l))}{(e^*(l,l)-e^*(l,h))} \leq \frac{(1-\alpha)\mathrm{E}(a|s=l)}{\alpha\mathrm{E}(a|s=h)},$$

$$\alpha\mathrm{E}(a|s=h)\lambda \leq (1-\alpha)\mathrm{E}(a|s=l)(1-\lambda),$$

$$\alpha(l + (\zeta + (1-\zeta)\alpha)(h-l))(\zeta + (1-\zeta)(1-\alpha)) \leq (1-\alpha)(l + (1-\zeta)\alpha(h-l))(1-\zeta)\alpha,$$

$$\alpha h\zeta \leq 0.$$

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

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

Lemma 6. *Consider a nonbabbling 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 supervisor could get the best possible result by sending feedback message h. She would convince the worker that she is a competent supervisor 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 if $\mu_b^* < 1$, then $\mu_g^*(l) = 0$.

Proof of Part (II): This follows from the derivation of (2).

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

Proof of Part (IV): There are two cases. The first case is $\mu_g^*(l) > 0$. 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 supervisor. Thus, $\mu_b^* \geq \max\{\frac{(\zeta+(1-\zeta)\alpha)}{(1+\zeta)}, \alpha\}$ and $\hat{\rho}(s,l) \geq \hat{\rho}(s,h)$.

The second case is that $\mu_g^*(l) = 0$. 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:

$$\Pr(s = l \wedge m = l \wedge t = g) = (1-\alpha)\rho(\beta + (1-\beta)(1-\alpha)),$$

$$\Pr(s = l \wedge m = l \wedge t = b) = (1-\rho)(1-\alpha)\left(1-\mu_b^*\right),$$

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

and

$$\Pr(s = h \wedge m = h \wedge t = g) = \alpha\rho(\beta + (1-\beta)\alpha),$$

$$\Pr(s = h \wedge m = h \wedge t = b) = (1-\rho)\alpha\mu_b^*,$$

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

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

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

$$\frac{\mu_b^*(1+\beta) - (\beta + (1-\beta)\alpha)}{\left(1 - (1-\rho)\mu_b^* - \rho\alpha(1-\beta)\right)\left(-(1-\rho)\mu_b^* - \rho(\beta + (1-\beta)\alpha)\right)} \leq 0.$$

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

$$\Pr(s = h \wedge m = l \wedge t = g) = (1-\alpha)\rho(1-\beta)\alpha,$$

$$\Pr(s = h \wedge m = l \wedge t = b) = (1-\rho)\alpha\left(1-\mu_b^*\right),$$

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

and

$$\Pr(s = l \wedge m = h \wedge t = g) = \alpha\rho(1-\beta)(1-\alpha),$$

$$\Pr(s = l \wedge m = h \wedge t = b) = (1-\rho)(1-\alpha)\mu_b^*,$$

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

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

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

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

Note that the denominator is negative as $(1-(1-\rho)\mu_b^* - \rho(\beta+(1-\beta)\alpha))$ 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.

APPENDIX B: PROOF OF PROPOSITION 4

Proof. We first prove that in any equilibrium such that a good supervisor sends $m = s$ we have $\theta_b^n > 0$. Suppose not. Then, by $\theta_b^n = 0$ we have that $\theta_b^k \geq 0$ 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 $\pi = \frac{2}{3}$. This implies that if a good supervisor observes $m = n$ each ability level is equally likely as $(1-\pi) = (2\pi - 1) = \frac{1}{3}$. 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 $\theta_b^n = 1$ if $l = 9$, $n = 10$ and $h = 11$. Let us first consider the bad supervisor. In this equilibrium, the bad supervisor expects to earn $\frac{\hat{\rho}(a\neq n,n)9^2+\hat{\rho}(n,n)10^2+\hat{\rho}(a\neq n,n)11^2}{3}\frac{3}{4} = \frac{302}{4} > 75$. If the supervisor deviates by sending message $m \in \{l, h\}$ then the supervisor earns $\frac{m^2+10^2}{3} \leq \frac{h^2+10^2}{3} = \frac{221}{3} < 75$. Thus the bad supervisor strictly prefers $m = n$. Now consider the good supervisor. If he receives signal $s = n$, then each ability is equally likely. In other words, he faces the same choice as the bad supervisor 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 $\pi\hat{\rho}(s,n)s^2 + (1-\pi)\hat{\rho}(n,n)n^2 = \frac{3}{4}(\pi s^2 + (1-\pi)n^2)$. Sending $m = \{l, h\}\backslash\{s\}$ yields $(1-\pi)n^2$ which is also strictly less than $(\pi s^2 + (1-\pi)n^2)$. Therefore, the good supervisor strictly prefers $m(s) = s$. As both players strictly prefer their proposed strategy to any strategy, it follows that this is an equilibrium. By continuity of the payoffs and strategies in the parameters such an equilibrium exists for a range of parameters. □