Public Disclosure, Private Revelation or Silence: Whistleblowing Incentives and Managerial Policy

David Austen-Smith & Timothy J. Feddersen
MEDS Department, Kellogg School of Management
Northwestern University
Evanston, IL 60208

December 2008

Abstract

The public revelation of organizational wrongdoing by insiders, *whistleblowing*, is widely reported, economically significant and can be extremely costly to the whistleblowers. We develop a model of whistleblowing involving a manager and an employee. Each has a privately known type that specifies the relative weight placed on social rather than personal payoffs. The manager chooses a whistleblowing policy consisting of conditional penalties for various employee actions; the employee observes the policy and chooses between saying nothing, revealing a (privately observed) socially costly *violation* to the manager, or whistleblowing. Given common knowledge of manager types we characterize equilibrium whistleblowing policies and employee behavior. We show that there may be a nonmonotonic relationship between the severity of the violation and the likelihood of whistleblowing. When manager types are private information we provide sufficient conditions for a separating equilibrium. Managerial choice of whistleblowing policies thus serves a dual purpose: providing incentives for reporting violations and providing information to employees regarding the willingness of the manager to fix violations that are privately reported.

1 Introduction

Jeffrey Wigand worked for the Brown and Williamson Tobacco Company. In 1993, he went to his CEO with serious concerns about various public health issues, in particular the implications of the company's refusal to consider proposals to remove cancer-causing additives from tobacco. Wigand was summarily fired and became a *whistleblower* (Johnson, 2003). Largely as a result of his testimony, in 1997 the tobacco industry reached a court settlement: "The companies–Philip Morris Companies, RJR Nabisco Holdings Corp., B.A.T. Industries PLC's Brown & Williamson and Loews Corp.'s Lorillard–reached a resolution with the attorneys general of nearly 40 states in which the industry will pay out $368.5 billion over the next quarter-century in compensation, drastically alter their marketing programs and submit to the regulatory heel of the FDA" (*Time*, June 30, 1997). Wigand himself, however, was subjected to retaliatory lawsuits and professional villification from his erstwhile employers.

Wigand is but one example of a long list of individuals calling their institutions to account. The public revelation of organizational wrongdoing by insiders, *whistleblowing*, is widely reported, economically significant and can be extremely costly to the whistleblowers as management (and occasionally co-workers) retaliate: “Most [whistleblowers] limp away from the experience with their careers, reputations and finances in tatters” (Sandler, 2007; see also, for example, Devine 1997, Johnson 2003). These observations raise several questions: Why, despite these costs, do individuals blow the whistle at all? How should managers design incentives in response to the possible public revelation of organizational failures? What sorts of wrongdoing are most likely to be reported publicly rather than privately?

There is a very large literature on whistleblowing, most of which focuses on the first question listed above. This literature looks primarily for common personality characteristics among whistleblowers, with mixed results.[1] Save for recording the incidence and scale of retaliatory responses to whistleblowers, relatively little attention has so far been paid to the other side of the issue, the organization itself.[2] In this paper we develop a two-sided incomplete information model involving a manager and an employee. The manager chooses a ‘whistleblowing policy’ consisting of conditional penalties for various employee actions, and the (relatively informed) employee chooses between saying nothing or revealing an observed violation privately or publicly. In addition to the employee having private information about the extent of any violation, both the employee and the manager have private information about their respective types. Loosely speaking, an individual’s type in the model describes the importance the individual attaches to having violations fixed: higher types attach more weight to violations being fixed than lower types.

The whistleblower problem considered here (that is, whether to reveal an observed violation publicly, report it privately to the manager, or to remain silent) might usefully be understood as integrating three problems: (1) the employee must decide between blowing the whistle and staying silent when he knows that the manager has also observed the violation and has not fixed it; (2) the employee must decide between whistleblowing and staying silent when he has reported the violation but is unsure whether the manager will fix it; (3) the employee must decide between staying silent or privately reporting the violation to the manager. Our general model captures all three of these smaller problems within a single structure: the employee privately observes a violation and must

[1] See, e.g., Glazer (1983), Miceli, Near & Schwenk (1991), Near & Miceli (1996), Alford (2001) and Johnson (2003).

[2] But see King (1999) who informally suggests that different hierarchical structures differentially facilitate intra-organizational communication; and Alford (2001) who takes a post-modern political theory approach to understanding both the motives of whistleblowers and the power relations in organizatons. Dyck, Morse & Zingales (2007) and Bowen, Call & Rajgopal (2008) offer some early econometric evidence on the incidence and impact of whistleblowing. Ting (2008) studies a game-theoretic model of whistleblower protection (to our knowledge, the only extant formal modeling contribution to the literature) that focuses on the extent to which whistleblower protections compromise a managers’ abilities to monitor employee effort efficiently. In his model , however, whistleblowing does not involve reporting illegal or malign actions but only sharing information about the quality of managerial decisions in a government bureaucracy with a politician (oversight committee).

decide between revealing the violation privately, blowing the whistle, or staying silent.

Insights into how the composite problems are resolved follow easily from the results derived under the general model. Specifically, for problem (1) when the employee has informed the manager of the violation and observed that he has not fixed it then, when the manager cannot impose costs on whistleblowing, there is a threshold employee type such that higher types blow the whistle and lower types remain silent. However, when possible, the low type manager always imposes maximal costs on whistleblowers. In that case the threshold type for whistleblowing is strictly decreasing in the magnitude of the violation. As a result, higher violations are more likely to be revealed than lower violations.

In many cases, however, employees cannot be sure whether the manager will ultimately fix the violation or whether he even knows the violation has occurred. In problem (2) both the employee and the manager have observed the violation but the employee is unsure whether the manager will ultimately fix it. If the manager cannot impose costs on whistleblowing, the decision of the employee to blow the whistle depends upon the likelihood that the manager fixes the violation if it is not revealed publicly. If the manager is sufficiently likely to be a high type then employees are more likely to blow the whistle on smaller violations. Conversely, employees are more likely to blow the whistle on larger violations when the manager is more likely to be a low type. When managers can impose costs for whistleblowing the analysis becomes more subtle and introduces the possibility that managers' types are revealed by their choice of costs.

For problem (3), the issue is whether the employee privately reports a violation to the manager. As in (2) above when the manager cannot impose costs for reporting violations, and the manager is expected to be likely to fix the violation, larger violations are reported privately more frequently. However, managers who prefer to fix all violations also prefer to impose costs on employees who stay silent; there are then two possible consequences. The first is that all violations are reported privately and fixed; and the second is that all large and small violations are reported privately and fixed but moderate violations are left unreported. On the other hand, when the manager is not expected to fix all reported violations and there are no penalties imposed, lower violations are more likely to be reported than higher ones. If such a manager can impose costs on employees, however, private reporting of all but the most serious violations is maximally penalized. As a consequence, higher violations are more likely to be reported and fixed than lower ones.

After describing the model in the next section, we then analyse three salient benchmark cases. In the first, we identify the employee's decision calculus when the manager's type is unknown and there is no whistleblowing policy (that is, there are no penalties for any employee action). In the remaining two cases, the manager's type is assumed common knowledge and the focus is on the strategic interaction between managerial choice of whistleblowing policy and the employee's response conditional on any observed violation. Managerial types, however, are rarely known ex ante but their choice of whistleblowing policy might reveal at least some information to employees in this respect. A final substantive section of the paper considers some signaling properties of managerial choice of whistleblowing policy, paying particular attention to circumstances under which there exists a separating equilibrium in managerial types.

2 Model

Consider a manager and a generic employee in a firm subject to violations, either in its internal practices and procedures, or in its quality control over production processes or the delivery of various services. Any violation $v \in [0,1]$ is observed privately by the employee and we identify v with the social cost of the violation; so $v = 0$ means there is no violation and no external cost imposed on society. Let G be the known, exogenous distribution of violations with strictly positive density at every $v \in (0,1)$. Both the manager and the employee are characterized by their type, described below, assumed private information to the respective individuals; the manager's type is $s \in \{0,1\}$ and the employee's type is a number $t \in [0,1]$. Let β be the prior probability that the manager is type $s = 1$; and let η be the prior pdf of employee types t on $[0,1]$, assumed differentiable a.e. Both β and η are common knowledge and, in a large economy, can be interpreted as identifying the proportions of various types in the economy.

The sequence of events is that, first, nature draws the agents' types s and t, along with the violation v. Both t and v are privately revealed to the employee; s is privately revealed to the manager.[3] Next, the manager announces a *whistleblowing policy.* A whistleblowing policy is a schedule of penalties $C(h,s) \geq 0$ imposed on the employee, conditional on the obervable (to the manager) history of events h (described below) and the manager's type s. Imposition of these penalties is taken to be costless to the manager (and so credible) but we assume they are bounded above: for every (h,s), $\bar{c} \geq C(h,s) \geq 0$. Further, we also suppose (unless explicitly stated otherwise) that managers can commit to any announced whistleblowing policy. Following the manager's announcement, a type t employee chooses between saying nothing (ϕ), privately revealing (p) the

[3] Realized violations in the model, therefore, are not deliberate and illegal choices by the manager. They may, however, include consequences of (here unmodeled) inappropriate or illegal managerial decisions.

violation v to the manager, and whistleblowing (w) i.e., revealing the violation publicly. Let $a_e(v,t) \in \{\phi, p, w\}$ denote a type t employee's action conditional on observing a violation v.

If the employee says nothing, nature then decides whether to make the violation common knowledge among the manager, employee and society at large, and does so with probability $q_\phi v$, $q_\phi \in [0,1)$ and payoffs are distributed. Let $\Omega_\phi \in \{0,1\}$, with $\Omega_\phi = 1$ if and only if nature reveals the violation when the employee says nothing. Then the probability a violation is revealed, $\Pr[\Omega_\phi = 1|v]$, is both increasing in the social cost of the violation and depends upon the action of the employee.

If the employee privately reveals the violation to the manager, the manager chooses whether or not to fix it. Assume that, other things equal, fixing a violation v costs the firm (and only the firm) αv, $\alpha > 0$. (Below, we introduce an additional (social) cost in the event that a violation is only corrected if it becomes publicly exposed in one way or another.) Let $a_m(v,s) \in \{f, \sim f\}$ denote a type s manager's action conditional on having a violation v reported to her privately. If she does not fix the violation, nature chooses whether to make it common knowledge (denoted $\Omega_{\sim f} \in \{0,1\}$) and does so with probability $\Pr[\Omega_{\sim f} = 1|v] = q_p v$, $q_p \in (q_\phi, 1)$ and payoffs are distributed. If she does fix the violation, the violation never becomes public (that is $\Omega_f \equiv 0$) and the payoffs are distributed.[4]

Finally, if the employee reveals the violation v publicly (that is, *blows the whistle*), then v becomes common knowledge (denoted $\Omega_w \equiv 1$). Summarizing, the maintained assumption is that

$$\Pr[\Omega_w = 1|v] = 1 > \Pr[\Omega_{\sim f} = 1|v] = q_p v > \Pr[\Omega_\phi = 1|v] = q_\phi v > 0.$$

The assumption that the probabilities of a violation becoming common knowledge consequent on the violation not being fixed (either because the employee says nothing or reports privately) are linear in violations, is purely a convenience. The important substantive assumption is that the probability unreported or unfixed violations are revealed is increasing in the size of the violation. This assumption seems intuitively plausible and all of the principal qualititative results hold if we assume the relevant probabilities are strictly increasing in v with $\Pr[\Omega_\phi = 1|v] < \Pr[\Omega_{\sim f} = 1|v]$ for all $v \in (0,1)$.

A history h observed by the manager consists of: (1) the employee's action $a_e \in \{\phi, p, w\}$; (2)

[4] In general, an employee may decide to blow the whistle on a reported violation if the manager chooses not to fix it. Including such a second opportunity adds little insight. The employee's decision in this case is described informally in the Introduction - and discussed more formally later - as the first of the three identified component problems.

the manager's own action regarding whether to fix the violation, $a_m \in \{f, \sim f\}$, conditional on $a_e = p$; and (3) the violation v only if v is revealed either privately (because the employee chose $a_e = p$) or by nature. Specifically, given a realized violation v and $j \in \{\phi, f, \sim f, w\}$, a generic history observed by the manager can be written as

$$h = (a_e, a_m, \Omega_j, V) \in \{\phi, p, w\} \times \{f, \sim f\} \times \{0, 1\} \times [0, 1]$$

where $V = v$ if either $a_e = p$ or $\Omega_j = 1$, and $V = [0, 1]$ otherwise. Informally, V captures the manager's knowledge about the violation v: either she knows it surely, $V = v$, or she knows only that a violation might have occurred, $V = [0, 1]$. In what follows, only some components of the history turn out to be decision-relevant for the manager (in particular, the employee's action a_e); therefore, where there is no ambiguity, we leave the decision-irrelevant components of h as understood in the analysis and write the policy as $C(a_e, s)$, taking any dependence on v as understood when $a_e = p$ or $a_e = \phi$ and $\Omega_\phi = 1$.

Thus, any history observed by the employee differs from that observed by the manager (h) only in that the employee always knows the violation v.

Before we define payoffs to the employee and manager it is helpful to define payoffs for Society and the Firm. These payoffs then form the basis of the payoffs to the employee and manager. The payoffs to society are as follows:

$$\pi_S(h, v) = \begin{cases} 0 & \text{if } a_e = p, a_m = f \\ \\ -\delta v \quad \text{if} \begin{cases} a_e = \phi, \Omega_\phi = 1, \text{ or} \\ a_e = p, a_m = \sim f, \Omega_{\sim f} = 1, \text{ or} \\ a_e = w \end{cases} \\ \\ -v & \text{otherwise} \end{cases}$$

In words, the best outcome for society occurs when the employee privately reveals the violation to the manager and the manager fixes it. If the violation is not fixed then society suffers a loss $-v$. If the violation has not been privately fixed by the manager and is revealed by nature, then society forces the firm to fix it but the firm (and society) suffer a loss δv ($\delta > 0$) proportional to the social cost v. The additional loss δv might be thought of as a hit to the firm's reputation beyond the actual cost of the violation. Unlike the purely private cost to fix a violation, αv, reputational costs are associated with various negative externalities to the industryand society as a whole (for instance, a generalized loss of trust or increase in skepticism regarding market institutions). Furthermore, we imagine that at least some share of the social cost v is realized before the violation is exposed. As with the earlier assumption of linear probabilities, the assumption of linear costs for fixing violations here is a convenience; the substantive assumption is that any such losses increase in the size of the violation.

The payoffs to the Firm are as follows:

$$\pi_F(h,v) = \begin{cases} -\alpha v & \text{if } a_e = p, a_m = f \\ \\ -(\alpha+\delta)v & \text{if } \begin{cases} a_e = \phi, \Omega_\phi = 1, \text{ or} \\ a_e = p, a_m = \sim f, \Omega_{\sim f} = 1, \text{ or} \\ a_e = w \end{cases} \\ \\ 0 & \text{otherwise} \end{cases}$$

The firm's most preferred outcome is that violations are not publicly revealed or fixed. The firm incurs cost αv either if the employee reveals the violation privately and the manager chooses to fix it or if a violation becomes common knowledge. When the violation is revealed by nature the firm also incurs a reputational cost δv. The firm avoids reputational costs when the manager fixes a violation. Conditional on violations being fixed the firm prefers that violations are privately revealed to the manager.[5]

The manager's type s parameterizes the extent to which she cares about the social cost; alternatively, $s = 1$ is consistent with the idea that the long term interests of the firm are aligned with society and therefore the type one manager is simply a farsighted profit maximizer. Here we consider the polar case in which the manager either does not care about the social cost of a violation ($s = 0$) or cares exclusively about social costs ($s = 1$). The payoff to the manager facing

[5] In principle, the reputational loss δv might differ between society and the firm, or be conditioned on h, with the loss being strictly greater when $a_e = p$ than otherwise. To avoid introducing additional notation, however, we assume both society and the firm experience the same reputational cost and that society does not penalize the manager for knowingly failing to fix a privately reported violation. This does not affect the qualitative properties of the model in any significant way.

a violation v is therefore

$$\pi_m(h, v, s) = s\pi_S(h, v) + (1 - s)\pi_F(h, v)$$

Similarly, the employee's type t parameterizes the extent to which he or she cares about social and private costs. We assume that the employee's earnings from the firm are positively related to the firm's payoff, π_F, and, to save on notation, presume the employee simply cares directly about π_F net of any penalties incurred under the whistleblowing policy, $C(\cdot, s)$. Specifically, the payoff to a type t employee is

$$\pi_e(h, v, s, t) = t\pi_S(h, v) + (1 - t)\left(\pi_F(h, v) - C(h, s)\right).$$

It is important to notice that, just as with the type one manager, the highest employee type, $t = 1$, cares only about social costs and is wholly insensitive either to the firm's payoff. The type $t = 1$ employee is also insensitive to any penalties associated with the whistleblowing policy. A motivation for this assumption is the idea that the monetary payoff to an employee is tied to the payoff of the firm. Employees variously concerned about any social costs imposed through the firm's activities, in addition to their respective private rewards, therefore, balance these concerns in choosing their actions at work. In the extreme, an employee cares exclusively about social costs, irrespective of the monetary implications of his actions; such an extreme is the $t = 1$ employee, a type that cannot be deterred by any penalty imposed by management. At the other extreme, the lowest type, $t = 0$, essentially shares the preferences of the type zero manager and cares exclusively about private monetary rewards and, therefore, exclusively about π_F net of any incurred penalties. All types $t \in (0, 1)$, however, place some weight on every argument in the generic employee's payoff function, $\pi_e(\cdot)$.

Recall that the history h is (by definition) the history observed by the manager and, therefore, includes knowledge of the violation v only if $a_e = p$ or if $a_e \in \{\phi, w\}$ and nature makes v common knowledge. Let H be the set of possible histories.

Given the sequence of decisions described above, behavioural strategies for the manager and employee are defined in the usual way. Formally, strategies can be written as

$$\begin{aligned} C &: H \times \{0,1\} \to [0, \bar{c}]; \\ a_e &: [0,1] \times [0,1] \to \{\phi, p, w\}; \\ a_m &: [0,1] \times \{0,1\} \to \{f, \sim f\}. \end{aligned}$$

Finally, we focus on Perfect Bayesian equilibria in (weakly) undominated strategies.[6]

3 Benchmark results

We first develop three benchmark cases. The first prohibits managers from imposing any whistleblowing policy in that penalties for any employee action following an observed violation are inadmissible. In this benchmark, the manager's type is presumed unknown, with employees assigning common knowledge probabity $\beta \in (0,1)$ to the event that the manager is type one $(s = 1)$. The remaining two benchmark cases allow managers to choose a whistleblowing policy freely, but the manager's type is assumed in each instance to be common knowledge at the outset. It is useful to begin, however, by identifying two critical violations that prove central for managers' best response decisions, irrespective of the details of any whistleblowing policy. These thresholds are described in the following simple lemma.[7]

Lemma 1 *(a) Assume the employee reports a violation v to the type s manager, $a_e = p$. Then the manager strictly prefers to fix the violation, $a_m = f$, if and only if $v > \hat{v}(s)$, where $\hat{v}(1) = 0$ and*

$$\hat{v}(0) = \frac{\alpha}{q_p\,(\alpha + \delta)}. \tag{1}$$

(b) A type s manager strictly prefers the employee to report a violation v privately, $a_e = p$, rather than to stay silent, $a_e = \phi$, if and only if $v > \dot{v}(s)$, where $\dot{v}(1) = 0$ and

$$\dot{v}(0) = \frac{\alpha}{q_\phi(\alpha + \delta)}. \tag{2}$$

Furthermore, $\dot{v}(0) > \hat{v}(0)$.

[6] Mixed strategy equilibria turn out to exist only in knife-edge circumstances and so, unless stated explicitly otherwise, we consider only pure strategy equilibria below.

[7] All formal proofs for this and subsequent results are collected in the Appendix, below.

Thus, both manager types surely fix violations $v > \dot{v}(0)$ but, while a type one manager prefers all violations to be reported privately so that she can fix them surely, there are some violations that a type zero manager would fix if she learned of them but prefers that the employee stays silent, specifically violations $v \in (\hat{v}(0), \dot{v}(0))$. Since $\hat{v}(1) = \dot{v}(1) = 0$, we can take these thresholds as understood hereafter, and write $\hat{v} \equiv \hat{v}(0)$ and $\dot{v} \equiv \dot{v}(0)$.

It is important to emphasize here that $\hat{v}(0) < \dot{v}(0)$ depends critically on the substantive assumption that $q_p > q_\phi$. This assumption is motivated by the familiar intuition that keeping a violation secret becomes harder the larger the number of people who know about the violation. But such an assumption may not be compelling here: once a manager is apprised of the violation, she has an opportunity to implement a cover up, in which case the reverse inequality, $q_p < q_\phi$, is more plausible. And if $q_p < q_\phi$, we have $\dot{v}(0) = 0$. Nevertheless, for this version of the paper, we maintain the presumption that $q_p > q_\phi$, leaving the important comparative analysis of the more general case for later.

3.1 No admissible penalties

In the first benchmark case, managers cannot impose any penalties on employees, that is, $C(h, s) \equiv 0$, for all (h, s). Since whistleblowing policies are always penalties imposed after the employee has acted, such penalties can be thought of as being established by reputation. The benchmark case without penalties corresponds, then, to the case in which the employer cannot establish either formal announced costs or a reputation for imposing costs.

Fix a violation $v > 0$. Whether or not an employee reports v privately, stays silent or blows the whistle depends (inter alia) on the employee's type, t. Moreover, because the employee's payoff $\pi_e(h, v, s, t)$ is linear in t, for each pair of actions $\{p, \phi\}$, $\{p, w\}$ and $\{\phi, w\}$, either all types strictly prefer one of the two actions being considered, or there is unique type that is indifferent between each action of the pair. An employee's best response decision given the belief $\beta \in [0, 1]$ about the manager's type is then defined by comparing his type to the threshold types for each pair of actions and any violation v.[8] For the current setting with no penalties imposed for any employee action and any ordered pair of actions (x, y), let $T_{xy}(v, \beta)$ denote the unique employee type at violation v and belief β such that a type t employee strictly prefers x to y [respectively, y to x] if and only if $t < T_{xy}(v, \beta)$ [respectively, $t > T_{xy}(v, \beta)$].

[8] It is worth noting that an alternative approach is to fix the employee type and identify those violations on which the employee takes one or other of the three available actions. Mathematically, however, it is easiest to fix the violation and identify those types taking particular actions with respect to that violation.

By Lemma 1, there are, depending on the manager's type, at most three intervals of violation to consider: $[\dot{v}, 1]$, $[\hat{v}, \dot{v})$ and $[0, \hat{v})$. For expositional purposes, we refer to these intervals as describing high, moderate and low violations, respectively. However, it is worth emphasizing that if $q_p < \alpha / (\alpha + \delta)$, the set of 'low' violations includes every possible violation, $v \in [0, 1]$; similarly, the set of 'high' violations can be empty. Consider each of these intervals of violations in turn.

3.1.1 High violations, $v \in (\dot{v}, 1]$.

If $v > \dot{v}$ then Lemma 1 implies that both types of manager surely fix the violation if they know about it, in which case all types of employee strictly prefer to report v privately to the manager rather than blow the whistle. To check this, assume the employee's type is t and note that reporting privately yields a payoff $-(1-t)\alpha v$, whereas whistleblowing yields

$$-t\delta v - (1-t)(\alpha + \delta)v \tag{3}$$

which is clearly smaller. Similarly, since the preferences of the lowest type of employee are identical to those of the type zero manager ($t = s = 0$), all types of employee strictly prefer to report $v > \dot{v}$ privately rather than stay silent. Thus high violations are always reported privately and always fixed: if $v \in (\dot{v}, 1]$ then $T_{\phi p}(v, \beta) = 0$ and $T_{\phi w}(v, \beta) = T_{pw}(v, \beta) = 1$.

3.1.2 Moderate violations, $v \in [\hat{v}, \dot{v})$.

Suppose $v \in [\hat{v}, \dot{v})$. By Lemma 1, both types of manager fix the violation although the type zero manager would have preferred not to have heard about it at all. And since, by hypothesis, no penalties are incurred by any action, a type t employee's payoffs from staying silent and from reporting a violation $v \in [\hat{v}, \dot{v})$ are independent of the manager's type. Specifically, the employee's payoff from staying silent is

$$-t\left(q_\phi v \delta v + (1 - q_\phi v)v\right) - (1-t)q_\phi v(\alpha + \delta)v; \tag{4}$$

and that from reporting any $v \geq \hat{v}$ privately is simply $-(1-t)(\alpha + \delta)v$. Comparing these payoffs and solving for t yields that the employee strictly prefers to report the violation $v \in [\hat{v}, \dot{v})$ privately rather than stay silent if and only if

$$t > \frac{\alpha - (\alpha + \delta)\, q_\phi v}{(\alpha + 1)(1 - q_\phi v)}. \tag{5}$$

The right side of this inequality defines the threshold $T_{\phi p}(v, \beta)$ on the interval $[\hat{v}, \dot{v})$.[9] By assumption, $v < \dot{v}$ and so, by (2), $T_{\phi p}(v, \beta) > 0$ and strictly decreasing convex in v on $[\hat{v}, \dot{v})$, with $T_{\phi p}(\hat{v}, \beta) < 1$ and $\lim_{v \uparrow \dot{v}} T_{\phi p}(v, \beta) = 0$.

In the absence of any penalties for reporting privately, the employee's decision hinges only on the extent to which he prefers to have the violation fixed: low types ($t < T_{\phi p}(v, \beta)$) prefer not to incur any costs at all so stay silent, but higher types ($t > T_{\phi p}(v, \beta)$) are willing to avoid the reputational cost δv by reporting v privately to insure the violation is fixed. Likewise, because all violations $v > \hat{v}$ are fixed by both types of manager, thus surely avoiding any reputational cost, no type of employee prefers to blow the whistle for such violations rather than report privately: $v \in [\hat{v}, \dot{v})$ implies $T_{\phi w}(v, \beta) = T_{pw}(v, \beta) = 1 > T_{\phi p}(v, \beta)$ for all β.

3.1.3 Low violations, $v \in [0, \hat{v})$.

Suppose $v \in [0, \hat{v})$ and consider the employee's preferences between staying silent and reporting privately. The payoff for staying silent is given by (4). Unlike with moderate violations, different types of manager respond differently to learning of a low violation: from Lemma 1, type one managers fix every violation and type zero managers never choose to fix a low violation. Thus, the employee's payoff from reporting v privately is managerial type-dependent. Specifically, given the employee believes the manager to be type one with probability β, the expected payoff from reporting the violation privately is

$$\beta\left[-(1-t)\alpha v\right] + (1-\beta)\left[-t\left(q_p v \delta v + (1 - q_p v) v\right) - (1-t) q_p v (\alpha + \delta) v\right]. \tag{6}$$

Comparing (4) with (6) and solving for t yields that the employee strictly prefers to report the violation $v > 0$ privately if and only if

$$t > \frac{(q_p - q_\phi)\, v\, (\alpha + \delta) + \beta\left(\alpha - q_p v\, (\alpha + \delta)\right)}{(\alpha + 1)\left((q_p - q_\phi)\, v + \beta\,(1 - q_p v)\right)}. \tag{7}$$

[9] Clearly, $T_{\phi p}(v, \beta)$ is independent of β. It is nevertheless convenient to retain the argument explicitly since such independence does not hold for every interval of violations.

The right side of this inequality is $T_{\phi p}(v, \beta)$ on the interval $(0, \hat{v})$ (and note that $v < \hat{v}$ implies $\alpha > q_p v\,(\alpha + \delta)$). Then all types $t > T_{\phi p}(v, \beta)$ report v privately while the remaining set of employee types stay silent. Note that, for all $v \in (0, \hat{v})$,

$$\lim_{\beta \to 0} T_{\phi p}(v, \beta) = \frac{(\alpha + \delta)}{(\alpha + 1)}. \tag{8}$$

That is, if there are no penalties for any employee action then, whatever the employee's belief regarding the manager's type, all employee types $t > (\alpha + \delta)\,/\,(\alpha + 1)$ strictly prefer to report all violations privately to the manager rather than say nothing at all. And comparing (7) with (5) yields that, for all $\beta < 1$,

$$\lim_{v \uparrow \hat{v}} T_{\phi p}(v, \beta) > \lim_{v \downarrow \hat{v}} T_{\phi p}(v, \beta).$$

The discontinuity in the threshold between reporting privately and staying silent at $\hat{v}$ is due entirely to the fact that type zero managers surely fix known violations $v > \hat{v}$ and surely do not (voluntarily) fix known violations $v < \hat{v}$. In particular,

$$\lim_{\beta \to 1} \lim_{v \uparrow \hat{v}} T_{\phi p}(v, \beta) = \lim_{\beta \to 1} \lim_{v \downarrow \hat{v}} T_{\phi p}(v, \beta).$$

Doing the calculus and some algebra, we have $dT_{\phi p}(v, \beta)/d\beta < 0$ for $v \in (0, \hat{v})$ and so, as the likelihood that the manager is type one increases, relatively more violations are reported privately. On the other hand, although, for all $\beta \in (0, 1)$,

$$\lim_{v \to 0} T_{\phi p}(v, \beta) = \frac{\alpha}{(\alpha + 1)}, \tag{9}$$

we have

$$\left. \frac{dT_{\phi p}(v, \beta)}{dv} \right|_{v \in (0, \hat{v}), \beta > 0} \gtreqless 0 \Leftrightarrow \frac{q_p - q_\phi}{q_p} \gtreqless \beta. \tag{10}$$

Figures 1a and 1b, below, illustrate these facts.

Figure 1a

$\beta > (q_p - q_\Phi)/q_p,\ \beta > \beta'$

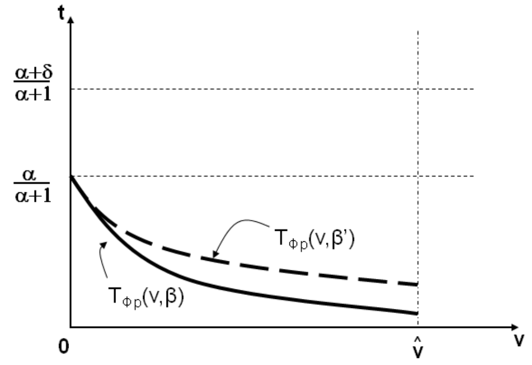
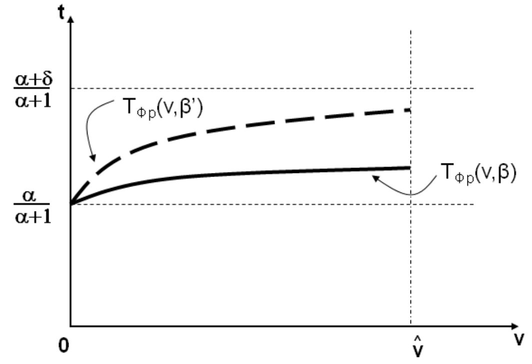


Figure 1b

$\beta < (q_p - q_\Phi)/q_p,\ \beta > \beta'$



If, as in Figure 1a, the manager is believed sufficiently likely to be type one (that is, $\beta > (q_p - q_\phi)/q_p$), higher violations are reported privately to the manager more frequently than lower violations. On the other hand, when, as illustrated in Figure 1b, the manager is believed suficiently likely to be type zero ($\beta < (q_p - q_\phi)/q_p$), (10) implies that it is lower violations that are reported privately more frequently than higher violations. But in both cases, the set of violations reported privately is increasing in the likelihood that the manager is type one. The key intuition here is that, given $\beta \in (0,1)$, the only type who can be indifferent between staying silent and privately reporting a violation v, is a type who strictly prefers to stay silent if he knows the manager will not fix the violation. For such a type, increasing the likelihood the manager is type zero (and so does not fix the violation) makes him strictly prefer to stay silent. Thus $T_{\phi p}(v, \beta)$ is strictly decreasing in β.

The intuition for why the threshold $T_{\phi p}(v, \beta)$ is decreasing or increasing in v on $(0, \hat{v})$, depending on whether β is less or greater than $(q_p - q_\phi)/q_p$, is a little more subtle. To appreciate this intuition, first note that in every case in which a violation v is fixed, the employee incurs a cost proportional to αv. On the other hand, the employee incurs the reputational cost δv only when the violation is publicly revealed, either by nature or through the employee blowing the whistle.

Suppose the manager is believed to be type one and recall that such a manager always fixes any violation of which she is aware. Therefore, given the manager is known to be type one, the employee could only incur the reputational cost δv by choosing to stay silent, in which case this cost is incurred with probability $q_\phi v$, rather than to report the violation v privately (with equivalent probability equal to $(1-\beta) q_p v = 0$ when $\beta = 1$). In turn, this implies that if an employee is indifferent between staying silent and reporting v privately, then he must strictly prefer to report any larger violation $v' > v$ privately: the difference between the two probabilities of incurring the reputational cost is $(q_\phi - 0)v$, which is clearly increasing in v. Hence, $T_{\phi p}(v, 1)$ is strictly decreasing in v on $(0, \hat{v})$.

Suppose the manager is type zero, so never willingly fixes a violation $v < \hat{v}$. In this case, there is a strictly positive probability that the employee incurs the reputational cost δv by staying silent (that is, $q_\phi v$) and by reporting privately (equal to $(1-\beta) q_p v = q_p v$ when $\beta = 0$). In contrast to when the manager is type one, $q_p v > q_\phi v$ implies that the employee is more likely to incur the reputational cost when reporting privately than he is by staying silent. Moreover, the difference in relative likelihoods, $(q_\phi - q_p)v$, is strictly decreasing in v. Therefore, given the manager is known to be type zero, if an employee is indifferent between staying silent and reporting v privately, then he must strictly prefer to remain silent regarding any larger violation $v' > v$; that is, $T_{\phi p}(v, 0)$ must be strictly increasing in v on $(0, \hat{v})$.

In general, the manager's type is not known surely. By continuity of $T_{\phi p}(v, \beta)$ in the belief β, however, the preceding discussion yields: (1) if β is sufficiently large, $T_{\phi p}(v, \beta)$ must be strictly decreasing in v on $(0, \hat{v})$; and (2) if β sufficiently small, $T_{\phi p}(v, \beta)$ must be strictly increasing in v

on $(0, \hat{v})$. And the boundary between what counts as 'sufficiently large' and 'sufficiently small' is identified formally by (10). In particular, the change in the relative likelihoods of a violation v being revealed if the employee is silent or reports privately, $(q_\phi - (1-\beta)q_p)\, v$, is independent of v when $\beta = (q_p - q_\phi)/q_p$.

Consider the employee's choice between staying silent and blowing the whistle on a low violation $v \in (0, \hat{v})$. By the no-penalties hypothesis, both actions are independent of the manager's type. The employee's payoff from staying silent is (4) and that from blowing the whistle is (3). Hence, an employee strictly prefers to blow the whistle on a violation $v \in (0, \hat{v})$ rather than stay silent if and only if,

$$-t\delta v - (1-t)(\alpha+\delta)v > -t\left(q_\phi v \delta v + (1-q_\phi v)v\right) - (1-t)q_\phi v(\alpha+\delta)v.$$

Doing the algebra yields that a type t employee blows the whistle on a violation $v \in (0, \hat{v})$ if and only if

$$\left[t(1+\alpha) - (\alpha+\delta)\right] v\left(1 - q_\phi v\right) > 0.$$

Let $T_{\phi w}(v, \beta)$ be the indifferent type between whistleblowing and staying silent for low violation v. Then,

$$T_{\phi w}(v, \beta) = \frac{(\alpha+\delta)}{(1+\alpha)}. \tag{11}$$

To see the intuition for why $T_{\phi w}(v, \beta)$ is constant in the violation v when there are no penalties for any employee action, consider the preceding inequality in t. Recalling that fixing and reputational costs are linear in v, the term $\left[t(1+\alpha) - (\alpha+\delta)\right] v$ describes the payoff difference between blowing the whistle and staying silent, *conditional* on the violation being exposed. Suppose $t = 1$. Then this difference is $(1-\delta)v$ because, while all types pay the reputational cost δv, higher types value having violations fixed with the highest type saving the cost v by insuring the violation is fixed by whistleblowing. On the other hand, suppose $t = 0$. Lower types put less value on having violations fixed but care more about the cost of fixing the violation αv, with the lowest type caring only about the total cost of a violation being fixed, $-(\alpha+\delta)v$. The remaining term of the inequality, $(1 - q_\phi v)$, is simply the difference in probabilities that the violation is fixed between whistleblowing and staying silent. Since this term is necessarily positive, the only decision relevant concern is the marginal payoff difference between the actions conditional on a violation being fixed; (11) follows.

A little algebra gives that for all $v \in [0, \hat{v})$ and all $\beta > 0$, $T_{\phi w}(v, \beta) > T_{\phi p}(v, \beta)$ and, therefore,

$\lim_{v\uparrow\hat{v}} T_{\phi w}(v,\beta) > \lim_{v\uparrow\hat{v}} T_{\phi p}(v,\beta)$ also. Therefore, whenever an employee prefers to blow the whistle on a violation $v \in (0,\hat{v})$ rather than stay silent, he prefers to report the violation privately rather than blow the whistle. It remains to check whether any type prefers whistleblowing to reporting a violation $v \in (0,\hat{v})$ privately.

Comparing (6) and (3) and doing the algebra, yields that the employee strictly prefers to blow the whistle rather than report $v \in (0,\hat{v})$ privately if and only if

$$t > \frac{(\alpha+\delta)(1-q_p v) - \beta(\alpha - (\alpha+\delta)q_p v)}{(1+\alpha)(1-\beta)(1-q_p v)}. \tag{12}$$

Since $v \in (0,\hat{v})$, $\alpha > q_p v(\alpha+\delta)$; nevertheless, as is easy to check, the right side of (12) is strictly positive and strictly increasing convex in v for all $\beta > 0$. It is not, however, necessarily bound above by $t = 1$. In particular, for all v, the right side of (12) is strictly increasing in β, has limit $T_{\phi w}(v,\beta)$ as $\beta \to 0$ and limit one as $\beta \to (1-q_p v)(1-\delta)/(1-q_p v(1-\delta))$. Let $T_{pw}(v,\beta)$ be the employee type such that whistleblowing is strictly preferred to reporting privately if and only if $t > T_{pw}(v,\beta)$; then,

$$T_{pw}(v,\beta) = \begin{cases} \text{Right side of (12) if } \beta < \frac{(1-q_p v)(1-\delta)}{(1-q_p v(1-\delta))} \\ 1 \text{ otherwise} \end{cases}.$$

The intuition for why $T_{pw}(v,\beta)$ is strictly increasing in v on $(0,\hat{v})$ whenever β is sufficiently small, is symmetric to that underlying why the threshold $T_{\phi p}(v,\beta)$ is increasing in $v < \hat{v}$ for sufficiently low β. Because the indifferent type $T_{\phi p}(v,\beta)$ is a low type, the main concern driving his decision about which of the two actions to take, silence or reporting privately, is with the relative likelihoods of incurring the reputational cost. On the other hand, a type $T_{pw}(v,\beta)$ is a high type in comparison and, here, the prime concern when choosing whether to blow the whistle or report privately is that the violation is fixed. Were this type sure the manager was type one, then reporting privately guarantees the violation is fixed without any reputational costs being incurred, the unequivocally best decision for him. However, since $\beta < 1$, the employee is not sure of the manager's response. In this case, he weighs having the violation fixed surely by blowing the whistle and incurring the reputational cost, against the uncertainty of having the violation fixed by reporting privately, which occurs with probability $(1-\beta)q_p v$, but possibly avoiding the reputational cost. Since the difference in the relevant likelihood, $(1-(1-\beta)q_p v)$, is decreasing in v, $T_{pw}(v,\beta)$ must increase in v.

3.1.4 Equilibrium behaviour in the no penalties model

Call the case whereby there are no admissible penalties for employee behaviour regarding the revelation or not of any observed violation, the *no penalties benchmark model.* Recall that for any two employee actions (x, y), $T_{xy}(v, \beta)$ denotes the threshold employee type such that, given (v, β), a type t employee strictly prefers x to y if and only if $t < T_{xy}(v, \beta)$, and conversely for $t > T_{xy}(v, \beta)$. Then the following proposition summarizes equilibrium behaviour for the no penalties benchmark model.

Proposition 1 *Assume the no penalties benchmark model and let the common belief that the manager is type one be $\beta \in (0, 1)$.*

(1) A type one manager fixes all known violations $v > 0$ and strictly prefers to have all violations reported privately.

(2) A type zero manager fixes all known violations $v > \hat{v}$ but strictly prefers to have only those violations $v > \dot{v}$ reported privately, where $\dot{v} > \hat{v}$.

(3) For all violations v and belief β, there exist unique threshold employee types $T_{\phi p}(v, \beta), T_{pw}(v, \beta)$ such that:

(a) If $v \in (0, \hat{v})$ then $0 < T_{\phi p}(v, \beta) < T_{pw}(v, \beta) \leq 1$, with strict inequality if and only if $\beta < (1 - \delta)$. $T_{\phi p}(v, \beta)$ is decreasing in β and decreasing [increasing] in v as $\beta > [<](q_p - q_\phi)/q_p$; and $\beta < (1 - \delta)$ implies $T_{pw}(v, \beta)$ is increasing in both v and β.

(b) If $v \in [\hat{v}, \dot{v})$ then $0 < T_{\phi p}(v, \beta) < T_{pw}(v, \beta) = 1$, where $T_{\phi p}(v, \beta)$ is independent of β and decreasing in v.

(c) If $v \in [\dot{v}, 1]$ then $T_{\phi p}(v, \beta) = 0 < T_{pw}(v, \beta) = 1$ and all types of employee surely report the violation privately.

These claims are illustrated in Figures 2a and 2b.

Before going on, recall from the Introduction that we identified three constituent problems within the whistleblowing framework: problem (1) involves the employee's decision after he has informed the manager of the violation and observed that she has not fixed it; problem (2) concerns identifying circumstances under which employees blow the whistle rather than remain silent when the manager knows the violation but the employee is unsure of the manager's decision; and problem (3) concerns circumstances under which employees choose to report violations privately rather than remain silent. Proposition 1 provides answers to all three problems when the manager cannot impose penalties for whistleblowing.

The first problem is captured by the threshold $T_{\phi w}(v,\beta)$, defining those employee types who are indifferent between blowing the whistle and staying silent. It is only necessary to reinterpret $q_\phi v$ as the probability a privately reported but unfixed violation is publicly revealed. Given the manager cannot impose any costs on the employee, it follows that for every violation smaller than $\hat{v}$ employee types higher than the threshold blow the whistle while lower types stay silent.

For the second problem, violations are assumed to be common knowledge between the employee and the manager. This case is exactly the problem that confronts the employee when deciding whether to report privately or blow the whistle in Proposition 1. When the employee privately reports the manager knows the violation; hence, reporting privately when the manager is assumed not to observe the violation, is equivalent to staying silent when the manager has observed the violation. The relevant threshold types for the second problem are therefore captured by the equations underlying $T_{pw}(v,\beta)$. In this case (as seen in Figure 2a below), if the manager is sufficiently likely to be a high type then no employee type blows the whistle. On the other hand, if the manager is likely to be a low type then (as seen in Figure 2b below), only high employee types blow the whistle and low violations are more likely to be publicly revealed than high violations.

The third problem focuses on whether or not to reveal a violation privately. In this case, the manager is assumed not to know the violation and the employee's choice is whether to stay silent or report privately. The relevant threshold conditions are defined by the equations underlying $T_{\phi p}(v,\beta)$. When the manager is sufficiently likely to be a high type (see Figure 2a below), the critical employee type that reports privately is decreasing in the violation and so higher violations are more likely to be privately reported than low ones. When the manager is more likely to be a low type (as seen in Figure 2b) low rather than high violations are more likely to be reported privately, at least until violations reach a level that even the low type of manager fixes them. From that point higher violations are more likely to be reported.

Returning to the general model in which all three actions, whistleblowing, reporting privately and remaining silent, are available to the employee, an important question concerns the likelihood that any violation is fixed. Under the benchmark model, Proposition 1 makes clear that all violations $v > \dot{v}$ are fixed, irrespective of manager type. This not true for smaller violations, however. Using Proposition 1, the probability that any violation $v \in (\hat{v}, \dot{v}]$ is fixed is simply

$$\Pr[v \text{ fixed } | \beta, v \in (\hat{v}, \dot{v}]] = q_\phi v \int_0^{T_{\phi p}(v,\beta)} \eta(t)dt + \int_{T_{\phi p}(v,\beta)}^1 \eta(t)dt$$

which is independent of any belief β regarding the manager's type. On the other hand, for any $v \in (0, \hat{v}]$, Proposition 1 yields

$$\Pr[v \text{ fixed } |\beta, v \in (0, \hat{v}]] = \left[\begin{array}{c} q_\phi v \int_0^{T_{\phi p}(v,\beta)} \eta(t)dt + (\beta + (1-\beta)q_p v) \int_{T_{\phi p}(v,\beta)}^{T_{pw}(v,\beta)} \eta(t)dt \\ + \int_{T_{pw}(v,\beta)}^{1} \eta(t)dt \end{array} \right].$$

Proposition 2 *If, in the benchmark model, $\beta \geq (1-\delta)$ then the probability that any violation $v < \hat{v}$ is fixed is increasing in β. For $\beta < (1-\delta)$, however, the probability that any violation $v < \hat{v}$ is fixed may be either increasing or decreasing in β.*

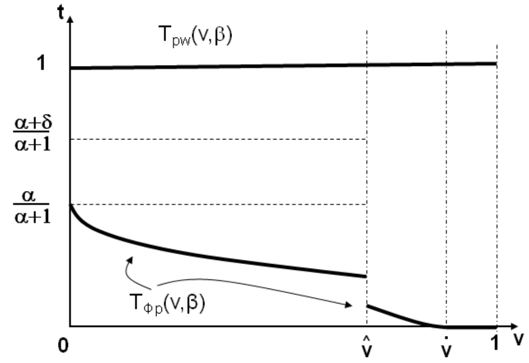
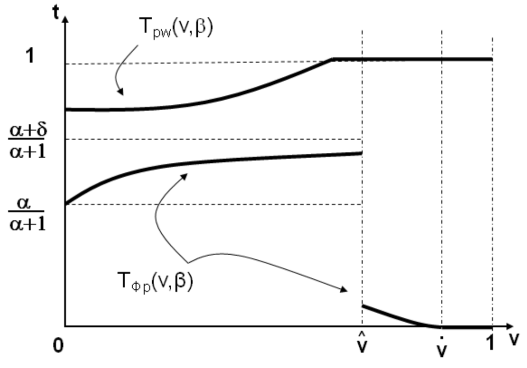


Figure 2a
$\beta > \max\{(1-\delta), (q_p - q_\Phi)/q_p\}$

Figure 2b

$\beta < \min\{(1-\delta), (q_p - q_\Phi)/q_p\}$



3.2 Whistleblowing under a type one manager

We now permit managers to implement a whistleblowing policy, that is, to impose penalties on employee behaviour should they so choose. Although we consider the existence and character of equilibria in which managers' types are revealed or not through their whistleblowing policy choice in a later section, it is convenient to begin by assuming that the manager's type, $s \in \{0, 1\}$, is common knowledge.

It is assumed that managers can costlessly impose penalties on employees up to a fixed limit. This allows any cost structure to be imposed in equilibrium by a manager. However, as indicated when describing the basic model earlier, we essentially refine the set of equilibria by restricting attention to whistleblowing policies that managers prefer to implement. Thus, the game is as if the manager first announces and commits to a whistleblowing policy that the employee observes and acts upon, following which the manager does in fact implement the relevant penalties under the announced policy. Obviously, because whistleblowing penalties are costless to implement for the manager, such a whistleblowing policy constitutes an equilibrium in an alternative game without commitment.

First suppose $\beta = 1$, so the manager is known to be type one. As for the no penalties benchmark model, the employee's best response behaviour given any whistleblowing policy $C(a_e, 1)$ is defined by a unique set of at most three threshold types. In the present case, with a known type one manager, these types are denoted generically by $t^*_{xy}(v,(c_x,c_y))$ for any pair of actions (x,y) and associated penalties under the whistleblowing policy, $(c_x,c_y)\equiv(C(x,1),C(y,1))$. Thus $t<t^*_{xy}(v,(c_x,c_y))$, for example, implies that given the violation v and the costs (c_x,c_y) of each action at v under the policy $C(\cdot,1)$, a type t employee strictly prefers x to y.

By Lemma 1, the type one manager fixes all reported violations and prefers to have all violations reported privately. Moreover, since it is assumed that the manager is surely type one, $\beta=1$ and claim (3) of Proposition 1 imply that, if there are no penalties for reporting a violation privately, all employee types strictly prefer to report any violation privately rather than blow the whistle: in current notation, that is, $t^*_{pw}(v,(0,c))=1$ for all $v>0$ and $c\geq 0$. Thus, conditional on her type being common knowledge as presumed here, a type one manager has no incentive to deter private reporting and no reason to penalize whistleblowing. Therefore, $C(a_e,1)=0$ for all $a_e\in\{p,w\}$. Furthermore, because the only incentive for blowing the whistle is to insure a violation is fixed, $C(p,1)=C(w,1)=0$ immediately implies that, if ever an employee prefers whistleblowing to remaining silent, that employee strictly prefers to report privately rather than blowing the whistle; thus, no type ever blows the whistle on any violation when the manager is known to be type one.

Whether the employee reports a violation v privately to a type one manager or stays silent, however, depends on the employee's type and the penalty $c\in[0,\bar{c}]$ imposed on any employee who says nothing. So consider a violation v and an employee type t; such a type weakly prefers to report the violation privately rather than stay silent when staying silent incurs a penalty $c\geq 0$ if and only if

$$-t\left(q_\phi v\delta v+(1-q_\phi v)v\right)-(1-t)\left(q_\phi v(\alpha+\delta)v+c\right)\leq -(1-t)\alpha v,$$

that is, if and only if

$$t\left(v(1+\alpha)(1-q_\phi v)-c\right)\geq v\left(\alpha-q_\phi v(\alpha+\delta)\right)-c,$$

where $\alpha\geq q_\phi v(\alpha+\delta)$ if and only $v\leq\dot{v}$, with strict inequality whenever $v<\dot{v}$.[10] Let $t^\circ(v,c)$ solve this last expression when it holds with equality. Then (assuming the left side of the equation is not

[10] In principle, the manager can condition the penalty on v, say $c(v)$, if the violation is revealed by nature, $\Omega_\phi=1$, but not otherwise, i.e. $\Omega_\phi=0$. In general, therefore, the term c in the preceding and subsequent expressions of this subsection is replaced by the sum, $(q_\phi vc(v)+c)$. However, as shown below, a type one manager invariably sets $(q_\phi vc(v)+c)\equiv\bar{c}$. So we save on notation by assuming at the outset that the penalty is independent of v, irrespective of Ω_ϕ.

zero),

$$t^{\circ}(v,c) = \frac{v\left(\alpha - q_{\phi}v(\alpha+\delta)\right) - c}{v(1+\alpha)(1-q_{\phi}v) - c}. \tag{13}$$

Given $C(\phi,1) = c \geq C(p,1)$, $t^{*}_{\phi p}(v,(c,0))$ is the threshold type such that a type t employee strictly prefers saying nothing to reporting v privately if and only if $t < t^{*}_{\phi p}(v,(c,0))$. When $c = 0$, (7) implies that $t^{\circ}(v,0)$ coincides with $T_{\phi p}(v,1)$; that is, $t^{*}_{\phi p}(v,(0,0)) \equiv T_{\phi p}(v,1)$. But this is not true for $c > 0$.

Lemma 2 *Suppose* $c \in (0,\bar{c})$*. Then there exist violations* $0 < v_1(c) < v^{*}(c) < v_2(c) < \dot{v}$ *such that*

(1) $v \notin (v_1(c),v_2(c)) \Rightarrow t^{*}_{\phi p}(v,(c,0)) = 0$*;*

(2) $v \in (v_1(c),v_2(c)) \Rightarrow t^{*}_{\phi p}(v,(c,0)) \in (0, t^{*}_{\phi p}(v^{*}(c),(c,0)))$ *with* $t^{*}_{\phi p}(v,(c,0))$ *strictly single-peaked on* $(v_1(c),v_2(c))$*.*

Furthermore, for all $v \in [0,1]$*,*

$\lim_{c\to 0} t^{*}_{\phi p}(v,(c,0)) = T_{\phi p}(v,1)$ *and* $\lim_{c\to \alpha\dot{v}/4} t^{*}_{\phi p}(v,(c,0)) = 0$.

When c is sufficiently large, that is, $c \geq \alpha\dot{v}/4$, all types always report all violations privately. The patterns of private reporting are more interesting when costs are smaller. For sufficiently small c, there always exist sufficiently low violations and low employee types who choose not to report such violations, even to a type one manager. In effect, sufficiently 'unethical' types prefer to stay silent, bear the penalty c and take the risk that low violations (relative to their type) are made public by Nature, rather than have the firm surely incur the costs of their correction due to reporting them directly. Further, when c is moderate, there exist types who report relatively low and high violations privately but stay silent for moderate violations. The intuition here is straightforward. For very low violations, the cost of staying silent is not worth the expected benefit of saving the expense of reporting the violation and having it surely fixed; on the other hand, for relatively large violations, the risk of the violation becoming common knowledge if it is not reported is sufficient to induce the employee to report it and eliminate the risk of having to pay the reputation cost. Finally, for moderate violations neither of the preceding two considerations for reporting privately are sufficient to induce private reporting and the employee prefers to say nothing and (in expectation) save the cost of having the violation fixed. All employee types prefer violations $v > \dot{v}$ to be fixed surely for any level of the cost c.[11] Figure 3 illustrates the cases.

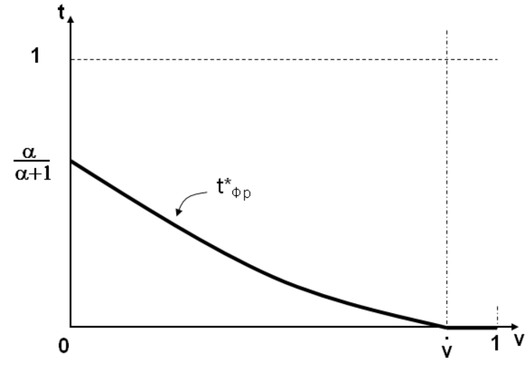


Figure 3a

$c = 0$

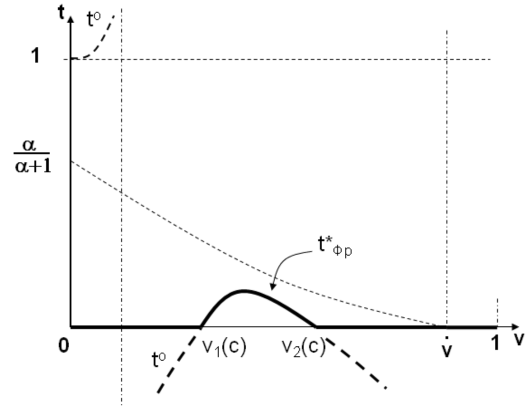
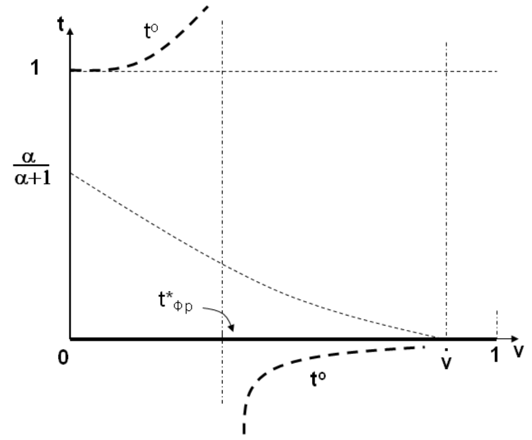


Figure 3b

$0 < c < \alpha\dot{v}/4$

[11] Note that $v \in (0, \hat{v}]$ implies $T_{\phi p}(v, 1) = t^*_{\phi p}(v, (0, 0))$, and $v \in (\hat{v}, \dot{v}]$ implies $T_{\phi p}(v, 1) = t^*_{\phi p}(v, (0, 0))$.

Figure 3c
c > αv̇/4



It follows from the above that the good manager's expected payoff (denoted hereon by $E\pi_m^1[\cdot]$) from implementing the whistleblowing policy $C(a_e, 1)$ can be written,

$$E\pi_m^1[C(a_e,1)] = -\int_0^{\dot{v}} \left(v\left(q_\phi v\delta + (1-q_\phi v)\right) \int_0^{t^*_{\phi p}(v,(c,0))} \eta(t)dt \right) dG(v),$$

which is surely maximized at $c = \bar{c}$. And assuming $\alpha\dot{v} \leq 4\bar{c}$, the manager's payoff from setting $c = \bar{c}$ is maximal and equal to zero.

Summarizing the discussion so far, we have equilibrium behaviour under a good manager as follows.

Proposition 3 *Suppose it is common knowledge that the manager is type one, $\beta = 1$. Then the manager fixes all violations of which she is aware, $a_m(v) = 1$ for all $v > 0$ and her whistleblowing policy is*

$$C(a_e,1) = \begin{cases} \bar{c} & \textit{if } a_e = \phi \\ 0 & \textit{otherwise} \end{cases}.$$

All violations $v \geq \dot{v}$ are reported to the manager. For any violation $v < \dot{v}$, an employee of type t reports v privately if $t \geq t^*_{\phi p}(v,(\bar{c},0))$ and otherwise stays silent. And if $\alpha\dot{v} \leq 4\bar{c}$, all employee types prefer to report all violations privately to the manager.

Note that there is an alternative equilibrium in which the type one manager imposes maximum costs on whistleblowers. Since employees never blow the whistle, the costs imposed on whistleblowers by type one managers are not behaviourally relevant.

It is immediate from the proposition that if $\alpha\dot{v} \leq 4\bar{c}$ then all violations are fixed with probability one under a type one manager. For smaller available penalties, some types keep silent regarding some violations and this loss is maximized when there is no penalty for staying silent, $C(a_e = \phi, 1) \equiv 0$. Let C_1^* denote the whistleblowing policy $C(a_e, 1)$ described in Proposition 3. By Proposition 3, therefore, the minimal probability that any violation $v < \dot{v}$ is fixed under a type one manager is

$$\Pr[v \text{ fixed } | \beta = 1, v \in (0, \dot{v}), C_1^*] = q_\phi v \int_0^{t^*_{\phi p}(v,(0,0))} \eta(t)dt + \int_{t^*_{\phi p}(v,(0,0))}^1 \eta(t)dt,$$

which is strictly increasing in v with $\lim_{v \to \dot{v}} \Pr[v \text{ fixed } | \cdot, v, C_1^*] = 1$.

3.3 Whistleblowing under a type zero manager

Suppose $\beta = 0$. If $v > \dot{v}$, all types of employee and manager strictly prefer that the violation is fixed. In particular, a type zero manager has no incentive to deter employees from reporting such violations privately and no reason to penalize whistleblowing. Thus, $C(p, 0) = C(w, 0) = 0$ for all $v \geq \dot{v}$ and, in directly analogous notation as in the previous subsection, $t_{pw}(v, (0, c)) = 1$ and $t_{\phi p}(v, (c, 0)) = 0$ for all $v \geq \dot{v}$ and $c \in [0, \bar{c}]$.

If $v < \dot{v}$ then, as is clear from Lemma 1, a type zero manager prefers not to be told privately of the violation v, but nevertheless fixes some such violations about which she is made aware. And if the manager imposes a cost for staying silent regarding a violation $v < \dot{v}$, she reinforces any prima facie incentive for the employee to report v privately or to blow the whistle. Hence, because (as observed above) there is no reason at all to impose a cost $c > 0$ on staying silent with respect to high violations $v \geq \dot{v}$, a type zero manager never has a strict incentive to impose a penalty for staying silent, so $C(\phi, 0) \equiv 0$.

Suppose $v \in [\hat{v}, \dot{v})$. Then the type zero manager fixes v if it is reported privately but prefers that the employee remain silent. As a result, the type zero manager has a strict incentive to impose the maximum penalty on an employee who either privately reveals or blows the whistle on a violation $v \in [\hat{v}, \dot{v})$. To confirm these observations, note that reporting such a violation v privately insures it is fixed and so, assuming a penalty $c \leq \bar{c}$ is imposed conditional on $a_e(v, \cdot) = p$, yields the employee a payoff $[-(1-t)(\alpha v + c)]$; and if the employee blows the whistle on v at some penalty $c' \geq 0$, he receives a payoff

$$-t\delta v - (1-t)\left((\alpha+\delta)v + c'\right), \tag{14}$$

which, conditional on $c' = \bar{c}$, is strictly lower for all types t and penalties c. In other words, if whistleblowing for $v \in [\hat{v}, \dot{v})$ is maximally penalized by the manager, private reporting strictly dominates whistleblowing for moderate violations and all employee types, even with maximal penalties imposed: $t_{pw}(v,(c,\bar{c})) = 1$ for all $v \in [\hat{v}, \dot{v})$ and all $c \in [0, \bar{c}]$. Therefore, by setting $C(p,0) = C(w,0) = \bar{c}$ for all $v \in [\hat{v}, \dot{v})$, the manager provides the maximal deterrent against reporting v privately rather than staying silent, while having no impact on the employee's decision between reporting privately and blowing the whistle. Thus, a type zero manager's whistleblowing policy involves imposing the maximal penalty $\bar{c}$ conditional on reporting moderate violations privately or publicly. And it is convenient here to note that, given the incentives to impose the maximal penalty for whistleblowing on moderate violations, the type zero manager likewise has a strict incentive to impose the maximal penalty for whistleblowing on low violations, $v < \hat{v}$. If such a penalty is warranted when, despite preferring not to hear of it, the manager fixes a violation $v \in [\hat{v}, \dot{v})$ that is reported privately, it is a fortiori warranted when she hears of a low violation that she is not going to fix unless it is subsequently exposed. Therefore, $C(w,0) = \bar{c}$ for all $v \in [0, \dot{v})$.

Despite penalties for reporting moderate violations privately, some employee types are willing to bear the cost of reporting some such violations privately to insure they are fixed. To identify these types for any $v \in [\hat{v}, \dot{v})$, note that a type t employee strictly prefers reporting v privately at cost $\bar{c}$ rather than staying silent with no penalty if and only if

$$-t\left(q_\phi v \delta v + (1-q_\phi v)v\right) - (1-t)q_\phi v(\alpha+\delta)v < -(1-t)\left(\alpha v + \bar{c}\right)$$

or, equivalently, if and only if

$$t > \frac{v\left(\alpha - (\alpha+\delta)q_\phi v\right) + \bar{c}}{v(1+\alpha)(1-q_\phi v) + \bar{c}}. \tag{15}$$

Since $v < \dot{v}$, the quotient here surely lies in the unit interval. Then the threshold type indifferent between staying silent and reporting $v \in [\hat{v}, \dot{v})$ privately, $t_{\phi p}(v,(0,\bar{c}))$, is exactly the right side of the inequality (15) and is strictly decreasing convex in v on $[\hat{v}, \dot{v})$.[12] All types $t > t_{\phi p}(v,(0,\bar{c}))$ report the violation $v \in [\hat{v}, \dot{v})$ privately to the manager and elicit a penalty $\bar{c}$ for doing so; all types

[12] Note that if there are no penalties for reporting privately then $t_{\phi p}(v,(0,0))$, $T_{\phi p}(v,0)$ and $t^*_{\phi p}(v,(0,0))$ coincide on $(\hat{v}, \dot{v})$.

$t < t_{\phi p}(v, (0, \bar{c}))$ remain silent without penalty.[13]

It remains to identify the penalty, if any, incurred by the employee conditional on reporting a violation $v < \hat{v}$ privately. To do this, we first assume the type zero manager imposes a cost $c(v) \leq \bar{c}$ conditional on a violation $v \in (0, \hat{v})$ being reported privately, and identify the threshold employee types, $t_{pw}(v, \cdot)$, $t_{\phi p}(v, \cdot)$ and $t_{\phi w}(v, \cdot)$.

Suppose the employee observes a violation $v \in (0, \hat{v})$. The employee's payoff from reporting privately is

$$-t\left(q_p v \delta v + (1 - q_p v) v\right) - (1 - t)\left(q_p v(\alpha + \delta) v + c(v)\right). \tag{16}$$

Comparing (16) with (14) and doing the algebra yields the threshold type between reporting v privately and blowing the whistle, $t_{pw}(v, (c(v), \bar{c}))$:

$$t_{pw}(v, (c(v), \bar{c})) = \frac{\bar{c} - c(v) + v(1 - q_p v)(\alpha + \delta)}{\bar{c} - c(v) + v(1 - q_p v)(\alpha + 1)}. \tag{17}$$

Thus, a type t strictly prefers to blow the whistle rather than report a violation $v \in (0, \hat{v})$ to a type zero manager, if and only if $t > t_{pw}(v, (c(v), \bar{c}))$. Conditional on $c(v) < \bar{c}$, $t_{pw}(v, (c(v), \bar{c}))$ is strictly decreasing in v on $(0, \hat{v})$ whenever $c'(v) \geq 0$.

Similarly, comparing (4) with (16) identifies the threshold type between staying silent and reporting v privately, $t_{\phi p}(v, (0, c(v)))$:

$$t_{\phi p}(v, (0, c(v))) = \frac{(q_p - q_\phi) v^2 (\alpha + \delta) + c(v)}{(q_p - q_\phi) v^2 (1 + \alpha) + c(v)}. \tag{18}$$

Given $v > 0$, the threshold is strictly increasing in $c(v)$ and (on doing the calculus and collecting terms), if the violation elasticity of $c(v)$ is strictly less than two, decreasing in v. While we simply assume this condition holds, it is worth noting that it surely obtains if $c'(v) \equiv 0$, which is shown below (Lemma 4) to be consistent with best response behaviour.

Finally, comparing (4) and (14) yields the threshold type between staying silent and blowing

[13] Intuitions for the comparative statics on $t_{\phi p}(v, (0, \bar{c}))$ and the analogous thresholds to be derived below, essentially mirror those for the thresholds $T_{xy}(v, \beta)$ discussed earlier in the no penalties benchmark model with $\beta = 0$. The only difference here is that the employee's basic incentives are qualified by the possibility of incurring various managerial penalties under the whistleblowing policy. But the particular implications of these costs for the structure of the thresholds are, for the most part, straightforward in each instance, so we typically leave them unstated in what follows.

the whistle on the violation v, $t_{\phi w}(v,(0,\bar{c}))$:

$$t_{\phi w}(v,(0,\bar{c})) = \frac{(\bar{c} + v(1-q_\phi v)\,(\alpha+\delta))}{(\bar{c} + v(1-q_\phi v)\,(\alpha+1))}. \tag{19}$$

This threshold is easily checked to be strictly decreasing in v on $(0,\hat{v})$.[14]

Now suppose $c(v) = c > 0$.

Lemma 3 *There exists $\tilde{c}(v) \in (0,\bar{c})$ such that $t_{pw}(v,(c,\bar{c})) \gtreqless t_{\phi p}(v,(0,c))$ and $t_{\phi w}(v,(0,\bar{c})) \gtreqless t_{\phi p}(v,(0,c))$ as $c \lesseqgtr \tilde{c}(v)$. Moreover,*

$$\tilde{c}(v) \equiv \frac{(q_p - q_\phi)v}{(1-q_\phi v)}\bar{c} \tag{20}$$

is zero at $v=0$ and strictly increasing convex in v on $(0,\hat{v}]$.

An immediate implication of Lemma 3 is the following useful observation.

Corollary 1 *Assume $v \in (0,\hat{v})$. Then,*

(1) For all $c < \tilde{c}(v)$, $t_{pw}(v,(c,\bar{c})) > t_{\phi w}(v,(0,\bar{c})) > t_{\phi p}(v,(0,c))$;

(2) For all $c \in [\tilde{c}(v),\bar{c}]$, $t_{pw}(v,(c,\bar{c})) \leq t_{\phi w}(v,(0,\bar{c})) \leq t_{\phi p}(v,(0,c))$

with strict inequalities for any $c > \tilde{c}(v)$.

In particular, the threshold $t_{\phi w}$ is never binding.

The type zero manager faces a tradeoff when choosing $c(v)$ for $v \in (0,\hat{v}]$: the best response penalty $c(v)$ must balance the gain from inducing the employee to stay silent rather than report privately, against the loss of encouraging him to blow the whistle. Setting $c(v) = c$ and differentiating appropriately at $v \in (0,\hat{v}]$ yields,

$$\frac{dt_{\phi p}(v,(0,c))}{dc} = \frac{(q_p - q_\phi)v^2\,(1-\delta)}{\left[(q_p - q_\phi)v^2(\alpha+1) + c\right]^2} > 0; \tag{21}$$

$$\frac{dt_{pw}(v,(c,\bar{c}))}{dc} = -\frac{(1-q_p v)v\,(1-\delta)}{\left[(1-q_p v)v(\alpha+1) + \bar{c} - c\right]^2} < 0. \tag{22}$$

But since the manager does not know the employee's type, her choice of penalty $c(v)$ depends on her beliefs over t. To save on notation, write $t_{\phi p}(v,(0,c)) \equiv t_{\phi p}(v,c)$ and $t_{pw}(v,(c,\bar{c})) \equiv t_{pw}(v,c)$ for $v \leq \hat{v}$, where c is the penalty imposed for reporting a violation $v \leq \hat{v}$ privately. Then, recalling

[14] Note that if there are no whistleblowing penalties, then $t_{pw}(v,(0,0)) = T_{pw}(v,0)$ and $t_{\phi p}(v,(0,0)) = T_{\phi p}(v,0)$. Also, tedious algebra shows $\lim_{v\uparrow\hat{v}} t_{\phi p}(v,(0,\tilde{c})) > \lim_{v\downarrow\hat{v}} t_{\phi p}(\hat{v},(0,\bar{c}))$.

that η is the pdf of employee types on $[0,1]$, the type zero manager's expected payoff conditional on a violation $v<\hat{v}$ and penalty c for reporting v privately is:

$$\begin{aligned}
E\pi_m^0[c,\cdot|v] &= \begin{bmatrix} \Pr[t<t_{\phi p}(v,c)]\left[-q_\phi v^2(\alpha+\delta)\right]+ \\ \Pr[t_{\phi p}(v,c)<t<t_{pw}(v,c)][-q_p v^2(\alpha+\delta)]+ \\ \Pr[t>t_{pw}(v,c)]\left[-v(\alpha+\delta)\right] \end{bmatrix} \\
&= -v(\alpha+\delta)\begin{pmatrix} q_\phi v\Pr[t<t_{\phi p}(v,c)]+ \\ q_p v\Pr[t_{\phi p}(v,c)<t<t_{pw}(v,c)]+ \\ \Pr[t>t_{pw}(v,c)] \end{pmatrix} \\
&= -v(\alpha+\delta)\begin{pmatrix} q_\phi v\int_0^{t_{\phi p}(v,c)}\eta(t)dt+ \\ q_p v\int_{t_{\phi p}(v,c)}^{t_{pw}(v,c)}\eta(t)dt+\int_{t_{pw}(v,c)}^1\eta(t)dt \end{pmatrix}. \qquad (23)
\end{aligned}$$

We can now prove

Lemma 4 *Suppose $\eta'(t)\leq 0$ for all $t\in[0,1]$. Then for any violation $v\in(0,\hat{v})$, setting $c(v)\in[\tilde{c}(v),\bar{c}]$ is a best response penalty for reporting the violation privately to the type zero manager.*

Whatever best response penalty $c(v)\in[\tilde{c}(v),\bar{c}]$ a type zero manager implements in her whistleblowing policy, employees either stay silent or blow the whistle in regard to a violation $v<\hat{v}$. Having said this, however, it is important to note that this extreme prediction is a consequence of assuming $\eta'(t)\leq 0$ on $[0,1]$. Although it is plausible that higher types are less likely than lower types, particularly under the natural interpretation of type as the extent to which an employee is ethical, it is not necessary. Moreover, the economics of the tradeoff facing the manager are independent of the distributional restriction. The logic suggesting the type zero manager chooses a penalty $c(v)$ to balance the gain from encouraging relatively low types to remain silent rather than report privately, against the loss from inducing relatively high types from whistleblowing rather than reporting privately, is quite general and typically can be expected to lead to at least some private reporting in equilibrium.

Before going on, it is instructive briefly to consider the manager's choice of penalty $c(v)$ for $v<\hat{v}$ when the employee's type t is known surely. Because the thresholds $t_{\phi p}(v,\cdot)$ and $t_{pw}(v,\cdot)$ are invertible, the manager in this case can identify the critical violations (if any) at which the type-t employee is, respectively, indifferent between staying silent and reporting privately (say, $v_{\phi p}(t,\cdot)$), and between reporting privately and blowing the whistle (say, $v_{pw}(t,\cdot)$). It then follows from earlier discussion, that the type zero manager's best choice is to impose penalties for reporting any violation $v < \hat{v}$ privately such that

$$\begin{aligned} v &\leq v_{\phi p}(t, (0, \tilde{c}(v))) \Rightarrow c(v) = \tilde{c}(v); \\ v &\in (v_{\phi p}(t, (0, \tilde{c}(v))), \hat{v}) \Rightarrow c(v) = 0. \end{aligned}$$

In effect, this policy maximizes the subset of violations within $(0, \hat{v})$ for which the employee stays silent without simultaneously broadening the subset of violations in this interval for which the employee blows the whistle. When the employee's type is unknown, however, the manager can condition her policy only on the revealed violation itself, inducing the tradeoff supporting setting $c(v) = \tilde{c}(v)$ over the interval $(0, \hat{v})$.

The following result summarizes equilibrium behaviour when the manager is known to be type zero.

Proposition 4 *Suppose it is common knowledge that the manager is type zero, $\beta = 0$ and assume the pdf of employee types is weakly decreasing on $[0, 1]$. Then the manager fixes all violations $v \geq \hat{v}$ of which she is aware, but strictly prefers only to have violations $v \geq \dot{v} > \hat{v}$ reported privately to her. Moreover, her whistleblowing policy is*

$$C(a_e, 0) = \begin{cases} \bar{c} & \textit{if } a_e = p \textit{ and } v \in [\hat{v}, \dot{v}) \\ c(v) \in [\tilde{c}(v), \bar{c}] & \textit{if } a_e = p \textit{ and } v < \hat{v} \\ \bar{c} & \textit{if } a_e = w \textit{ and } v < \dot{v} \\ 0 & \textit{otherwise} \end{cases} .$$

All violations $v \geq \dot{v}$ are reported privately to the manager. For any violation $v \in [\hat{v}, \dot{v})$, the employee reports v privately if $t > t_{\phi p}(v, (0, \bar{c}))$ and otherwise stays silent. For any violation $v \in (0, \hat{v})$ and $c \geq \tilde{c}(v)$ for reporting v privately, the employee stays silent if $t \leq t_{\phi p}(v, (0, c))$ and otherwise blows the whistle.

Figure 4a illustrates equilibrium employee behaviour given the whistleblowing policy identified in Proposition 4; Figure 4b illustrates the type zero manager's whistleblowing policy in response to any violation $v \in (0, \dot{v})$ being reported privately.

Figure 4a
Employee behaviour with a type 0 manager

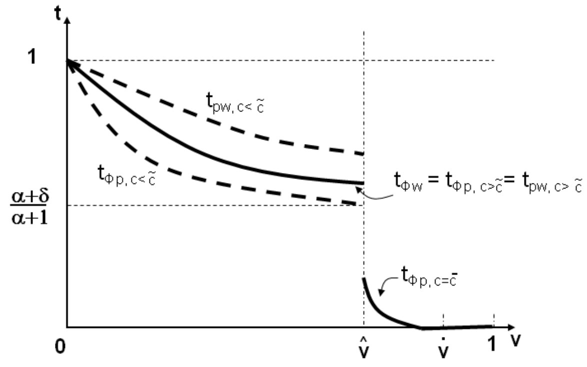
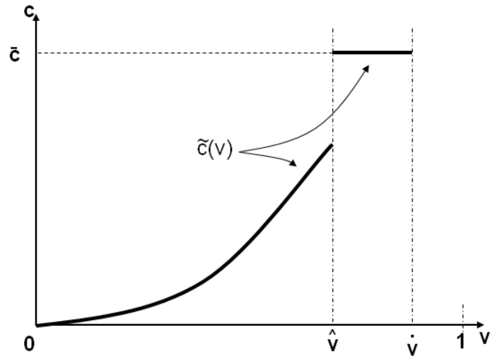


Figure 4b
Costs imposed for reporting privately



Denote the whistleblowing policy $C(a_e, 0)$ described in Proposition 4 by C_0^*. Assuming $\eta'(t) \leq 0$ on $[0, 1]$, Corollary 1 and Proposition 4 imply that the probability any violation $v < \hat{v}$ is fixed under a type zero manager is

$$\Pr[v \text{ fixed } |\beta = 0, v \in (0, \hat{v}), C_0^*] = q_\phi v \int_0^{t_{\phi w}(v,(0,\bar{c}))} \eta(t)dt + \int_{t_{\phi w}(v,(0,\bar{c}))}^1 \eta(t)dt$$

which is strictly increasing in v with $\lim_{v \to \hat{v}} \Pr[v \text{ fixed } |\cdot, v] < 1$. Similarly, for $v \in [\hat{v}, \dot{v})$ we obtain

$$\Pr[v \text{ fixed } |\beta = 0, v \in [\hat{v}, \dot{v}), C_0^*] = q_\phi v \int_0^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt + \int_{t_{\phi p}(v,(0,\bar{c}))}^1 \eta(t)dt$$

which is also strictly increasing in v with $\lim_{v \to \dot{v}} \Pr[v \text{ fixed } |\cdot, v, C_0^*] = 1$. Comparing these values with those for a type one manager yields that, for all $v \in (0, \dot{v})$, the probability a violation v is fixed under a type zero manager is strictly less than the probability v is fixed under a type one manager.

4 Signaling equilibria

To this point, we have identified three important benchmark settings: in the no penalties case, the manager's type is subject to uncertainty and the whistleblowing policy is trivial; in the remaining two cases, the manager's type is known surely and the focus is on the best response whistleblowing policy for each type, along with the associated behaviour of the employee. It is clear from these three cases that an employee's belief about a manager's type plays a key role in determining how that employee responds to an observed violation and, therefore, has an impact on the manager's payoff. Consequently, in a more general setting where a manager can choose the whistleblowing policy, but the employee is unsure of the manager's type, managers might use such policies to signal their type and thus influence subsequent employee behaviour. In this section, therefore, we consider some implications of the signaling incentive for observed whistleblowing policies and employee behaviour.

Given that a whistleblowing policy involves a schedule of costless penalties, any policy announcement by a manager is cheap talk. And because implementing whistleblowing policies is costless but employee beliefs about how exactly a manager might react affects their decisions, there exist a very large set of supportable pooling equilibria.[15] Moreover, pooling equilibria in pure strategies depend in part on more or less arbitrary specifications of players' out-of-equilibrium beliefs. On

[15] For example, if whistleblowing policies are cheap talk, then the no penalties benchmark equilibrium (Proposition 1) can be supported as a babbling equilibrium.

balance, therefore, there seems little insight to be had from exploring this class of equilibrium in any detail. Instead, we initially maintain the assumption that managers can commit to implementing any feasible announced policy and look for conditions under which the manager's announced policy surely signals their type, that is, separating equilibria. We then argue that at least some equilibria identified under the commitment assumption are also equilibria when the assumption is relaxed.[16]

Without loss of generality, let the set of available messages for managers be the set of possible whistleblowing policies, say $\mathcal{C}$. Suppose that the type one manager announces the whistleblowing policy $C_1^* \in \mathcal{C}$, identified in Proposition 3, and the type zero manager announces the policy $C_0^* \in \mathcal{C}$ specified in Proposition 4. Then the employee updates his belief such that $\beta(C_1^*) = 1$ and $\beta(C_0^*) = 0$. Then, by Proposition 3, the type one manager's expected payoff is

$$E\pi_m^1\left[C_1^*\right] = -\int_0^{\dot{v}} \left(v\left(q_\phi v\delta + (1 - q_\phi v)\right) \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt \right) dG(v). \tag{24}$$

And, by Proposition 4 and Lemma 1, we obtain the type zero manager's expected payoff,

$$\begin{aligned} E\pi_m^0\left[C_0^*\right] &= -\int_0^{\hat{v}} v(\alpha+\delta) \left(\begin{array}{c} q_\phi v \int_0^{t_{pw}(v,(\tilde{c},\bar{c}))} \eta(t)dt + \\ \int_{t_{pw}(v,(\tilde{c},\bar{c}))}^1 \eta(t)dt \end{array} \right) dG(v) \\ &\quad - \int_{\hat{v}}^{\dot{v}} \left(\begin{array}{c} v(\alpha+\delta) q_\phi v \int_0^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt + \\ v\alpha \int_{t_{\phi p}(v,(0,\bar{c}))}^1 \eta(t)dt \end{array} \right) dG(v) \\ &\quad - \int_{\dot{v}}^1 v\alpha dG(v). \end{aligned} \tag{25}$$

Although not, unfortunately, a very sharp result, the following identifies a sufficient condition for a separating equilibrium in which the type one manager announces C_1^* and the type zero manager announces C_0^*; employees thereby identify their manager's type and respond accordingly (that is, as described in Propositions 3 and 4, respectively).

Proposition 5 *There exists $n > 0$ such that, if $\int_{t_{pw}(\hat{v},(\bar{c},\bar{c}))}^1 \eta(t)dt \leq n$, then there exists a separating equilibrium in which the type one manager adopts the whistleblowing policy C_1^* and the type zero manager adopts the whistleblowing policy C_0^*.*

In effect, the proposition claims that the less likely it is that the employee is a high ethical type (here, a type $t > t_{pw}(v, (\tilde{c}, \bar{c}))$, the less likely it is that the type zero manager profits by deviating

[16] A more general analysis of pooling equilibria with commitment and the cheap talk model is deferred to subsequent work.

from a separating announcement to mimic the type one manager. The gain from mimicking the type one manager is entirely due to those ethical types who, had they known the manager was type zero, would have blown the whistle on a violation v close to $\hat{v}$ but, believing the manager to be type one, report privately instead. Similarly, it is clear from the proof of the result that separation can also be supported when the distribution of violations places most weight on violations $v > \hat{v}$.

Although the argument above was predicated on the assumption that managers commit to their announced policies, the role of any imposed penalties is purely through affecting the employee's beliefs. In particular, given that employee behaviour is conditioned on their beliefs about the penalties exacted ex post as a result of their actions, implementing any penalties other than those anticipated by the employee has no impact on the manager's payoff. It follows that the separating equilibrium above is also an equilibrium under cheap talk policy announcements. Of course, whether or not there is any ability to commit to an announced policy, there can be other separating equilibria. The analysis of the two benchmark cases in which managerial type is common knowledge, however, makes clear that no other separating equilibrium can improve either type of manager's payoff.

5 Conclusion

In the Introduction, we pointed to three important questions about whistleblowing. First, why, despite the evident personal costs, do individuals blow the whistle at all? Second, how should managers design incentives in response to the possible public revelation of organizational failures? And third, what sort of violations are most likely to be reported publicly rather than privately? The analysis in the paper focuses on the latter two concerns, finessing the first question by simply positing that individuals, either employees or managers, vary with respect to the relative weight they place on social and private payoffs, that is, their types: the higher the type, the more weight the individual places on the social consequences of firm behaviour. While an employee's type may take any real value between zero and one, managers (with no loss of substantive insight) are assumed to be extreme, either low type (profit maximizing) or high type (socially responsible). Then, assuming an agent's type is private information, the main results with respect to the latter two questions can be summarised as follows.

Given the proportion of very high type employees in the population is not too large, there exists a signaling equilibrium in which manager types separate by adopting their best response whistleblowing policy conditional on their type being full information. In this equilibrium:

(1) violations can be partitioned into three sets, of which only the "lowest" need be non-empty;

(2) the "highest" violations are invariably reported privately by all employee types to either type of manager who surely fixes the violation and does not penalize the employee;

(3) "moderate" violations are fixed by both types of manager conditional on being privately informed, but only the high type manager prefers to be told of such violations whereas the low type manager penalizes any private reporting of "moderate" or "low" violations;[17]

(4) only the high type manager fixes "low" violations when informed but not all employee types are willing to report all such violations; the low type manager prefers not to be informed and does not fix the violation if the employee nevertheless reports it privately; and sufficiently high employee types blow the whistle on "low" violations;

(5) only the low type manager penalizes whistleblowing or private reporting; the high type manager only penalizes remaining silent. Moreover, not all low or moderate violations are reported to either type of manager, even when there are no penalties for any action.

Many organizations provide mechanisms that allow anonymous reporting of observed violations. A motivation for such policies may be to encourage more reporting by decreasing the fear of retribution.[18] Our model gives insight into the implications of anonymous reporting. For example, if the impact of an anonymous reporting system is to remove all threat of penalties, then a simple comparison of the cases with and without costs is revealing. Perhaps suprisingly, when managers are high types, anonymous whistleblowing policies may be counterproductive. High types punish silence and thereby encourage private reporting of violations that are then fixed; the removal of costs reduces incentives for private reporting. The result is that anonymity reduces the number of violations that are fixed. On the other hand, if the manager is known to be a low type then the effect of eliminating penalities increases the number of violations that are fixed. The removal of penalties encourages more private reporting and whistleblowing.

When employees are uncertain about the manager's type the impact of anonymity may be more subtle. As before, removing costs has direct effects on the incentives for reporting. There is also, however, an indirect effect on the ability of managers to signal their type through type-specific whistleblowing policies. The inability of the manager types to seperate has ambiguous welfare

[17] It is important to emphasize that this last observation depends on the assumption that a privately reported violation is more likely to be exposed than an unreported violation (i.e., $q_p > q_\phi$). Without this assumption, a low type manager always wants to be informed but does not necessarily fix the violation.

[18] See, for example, Boatright (2007: 109); Alford (2001: 36); Velasquez (2002: 474); King (1999).

implications. While we have some preliminary results on this problem a comprehensive analysis is an appropriate topic for future research.

6 Appendix

Proof of Lemma 1. (a) Fix v and consider the subgame in which $a_e = p$ and the manager must choose $a_m \in \{f, \sim f\}$. That is, the employee has informed the manager of the violation and the manager must choose whether or not to fix a violation v. A type s manager prefers to fix the violation if

$$-(1-s)\alpha v \geq -s\left(q_p v\delta + (1-q_p v)\right)v - (1-s)vq_p(\alpha+\delta)v.$$

Let $\hat{v}(s) \in [0,1]$ be the critical violation for which the manager of type s is just indifferent between fixing and not fixing the violation. Substituting for s and collecting terms yields the result.

(b) Fix v and consider the manager's preferences over whether the employee reports this violation privately, $a_e = p$, or stays silent, $a_e = \phi$. Let $\dot{v}(s) \in [0,1]$ be the critical violation for which the manager of type s is just indifferent between $a_e = p$ and $a_e = \phi$. A type s manager's payoff when $a_e = \phi$ is

$$-s\left(q_\phi v\delta + (1-q_\phi v)\right)v - (1-s)q_\phi v(\alpha+\delta)v. \tag{26}$$

From part (a) above, $\hat{v}(1) = 0$, so a type one manager fixes all violations and enjoys zero payoff. Substituting for $s = 1$ in (26), therefore, gives that a type one manager strictly prefers any violation $v > 0$ to be reported privately; that is, $\dot{v}(1) = 0$. Now suppose the manager is type zero, $s = 0$; then the payoff (26) becomes $-q_\phi v(\alpha+\delta)v$. By part (a), there are two cases.

(1) If $v < \hat{v}(0)$ and $a_e = p$, a type zero manager does not fix the violation, giving the payoff, $-vq_p(\alpha+\delta)v$. Since $q_p > q_\phi$, therefore, we have that the type zero manager prefers the employee to stay silent rather than report any violation $v < \hat{v}(0)$ privately.

(2) If $v \geq \hat{v}(0)$ and $a_e = p$, a type zero manager fixes the violation, yielding a payoff, $-\alpha v$. Hence, the type zero manager prefers the employee to report v privately rather than stay silent only if $v \geq \alpha/\left[q_\phi(\alpha+\delta)\right]$. Combining (1) and (2), therefore, yields

$$\dot{v}(0) = \frac{\alpha}{q_\phi(\alpha+\delta)}.$$

Finally, since $q_p > q_\phi$, $\hat{v}(0) < \dot{v}(0)$. □

Proof of Proposition 2. Fix a violation $v \in (0, \hat{v})$ and differentiate $\Pr[v \text{ fixed } |\beta, v \in (0, \hat{v})]$ with respect to β to yield

$$\begin{aligned}
& \left.\frac{d\Pr[v \text{ fixed } |\beta, \cdot]}{d\beta}\right|_{v \in (0,\hat{v})} \\
= \;& q_\phi v \eta(T_{\phi p}(v,\beta)) \frac{dT_{\phi p}}{d\beta} + (1 - q_p v) \int_{T_{\phi p}(v,\beta)}^{T_{pw}(v,\beta)} \eta(t)dt \\
& +(\beta + (1-\beta) q_p v) \left[\eta(T_{pw}(v,\beta)) \frac{dT_{pw}}{d\beta} - \eta(T_{\phi p}(v,\beta)) \frac{dT_{\phi p}}{d\beta} \right] \\
& -\eta(T_{pw}(v,\beta)) \frac{dT_{pw}}{d\beta}.
\end{aligned}$$

Now, $dT_{\phi p}/d\beta < 0$ and $dT_{pw}/d\beta \geq 0$. Hence, the first and last terms of the derivative are negative, while the second and third terms are both positive, leaving the sign of the derivative ambiguous in general. However, from Proposition 1, for all $\beta \geq (1 - \delta)$ and $v \in (0, \hat{v})$, $T_{pw}(v, \beta) = 1$. Therefore, $\beta \geq (1 - \delta)$ implies $dT_{pw}/d\beta \equiv 0$ and we obtain

$$\begin{aligned}
& \left.\frac{d\Pr[v \text{ fixed } |\beta, \cdot]}{d\beta}\right|_{v \in (0,\hat{v}), \beta \geq (1-\delta)} \\
= \;& \left[\begin{array}{c} q_\phi v \eta(T_{\phi p}(v,\beta)) \frac{dT_{\phi p}}{d\beta} + (1 - q_p v) \int_{T_{\phi p}(v,\beta)}^{1} \eta(t)dt \\ +(\beta + (1-\beta) q_p v) \left[-\eta(T_{\phi p}(v,\beta)) \frac{dT_{\phi p}}{d\beta} \right] \end{array} \right]
\end{aligned}$$

$$= (1 - q_p v) \int_{T_{\phi p}(v,\beta)}^{1} \eta(t)dt - \left[\beta(1 - q_p v) + v\left(q_p - q_\phi\right)\right] \eta(T_{\phi p}(v,\beta)) \frac{dT_{\phi p}}{d\beta} > 0,$$

which completes the proof. □

Proof of Lemma 2. By definition of $t^*_{\phi p}(v, (c, 0))$, $t^*_{\phi p}(v, (c, 0)) = t^\circ(v, c)$ *if* (1) $v(1+\alpha)(1-q_\phi v) > c$ and (2) $t^\circ(v, c) \in [0, 1]$. But neither (1) nor (2) can hold for all violations $v \in (0, \dot{v})$ when $c > 0$. In particular, because $v(1 + \alpha)(1 - q_\phi v) > v\left(\alpha - q_\phi(\alpha + \delta)v\right)$ for all $v > 0$,

$$[v(1 + \alpha)(1 - q_\phi v) < c] \Rightarrow [v\left(\alpha - q_\phi(\alpha + \delta)v\right) < c] \Rightarrow t^\circ(v, c) > 1,$$

in which case all types t strictly prefer to report violations v for which $[v(1 + \alpha)(1 - q_\phi v) < c]$ to the manager; that is, $t^*_{\phi p}(v, (c, 0)) = 0$. Furthermore, since $v < \dot{v}$ and $v\left(\alpha - q_\phi(\alpha + \delta)v\right) \geq 0$ is maximized at $v = \alpha/2q_\phi(\alpha + \delta) = \dot{v}/2$, if

$$\frac{\dot{v}}{2}(\alpha - q_\phi \frac{\dot{v}}{2}(\alpha + \delta)) = \frac{\dot{v}}{4}\alpha < c \tag{27}$$

then, for all $v \in (0, \dot{v})$,

$$[v(1+\alpha)(1-q_\phi v) > \bar{c}] \Rightarrow t^\circ(v,c) < 0$$

and again all types prefer to report the violations privately, $t^*_{\phi p}(v,(c,0)) = 0$. Furthermore, $t^\circ(v,c)$ is clearly continuous single-peaked in v conditional on $0 < t^\circ(v,c) < 1$. Therefore, there exist violations $v_1(c) \leq v_2(c)$ such that $t^*_{\phi p}(v,(c,0)) = t^\circ(v,c)$ for all $v \in (v_1(c), v_2(c))$, from which the lemma follows easily from (13). □

Proof of Lemma 3. For any $v < \hat{v}$, $t_{pw}(v,(c,\bar{c})) \gtreqless t_{\phi p}(v,(0,c))$ as

$$\begin{aligned} \frac{v(1-q_p v)(\alpha+\delta)+\bar{c}-c}{v(1-q_p v)(\alpha+1)+\bar{c}-c} &\gtreqless \frac{(q_p-q_\phi)v^2(\alpha+\delta)+c}{(q_p-q_\phi)v^2(1+\alpha)+c} \Leftrightarrow \\ \frac{(q_p-q_\phi)v}{(1-q_\phi v)}\bar{c} &\gtreqless c. \end{aligned}$$

Therefore, since $(q_p - q_\phi)v < (1-q_\phi v)$, there exists $\tilde{c}(v) \in (0,\bar{c})$, such that $t_{pw}(v,(c,\bar{c})) \gtreqless t_{\phi p}(v,(0,c))$ as $c \lesseqgtr \tilde{c}(v)$; specifically,

$$\tilde{c}(v) \equiv \frac{(q_p-q_\phi)v}{(1-q_\phi v)}\bar{c}.$$

A similar calculation shows $t_{\phi w}(v,(0,\bar{c})) \gtreqless t_{\phi p}(v,(0,c))$ as $c \lesseqgtr \tilde{c}(v)$. The lemma now follows easily. □

Proof of Lemma 4. Differentiating (23), we find,

$$\begin{aligned} \frac{dE\pi^0_m[c,\cdot|v]}{dc} &\propto -\left[\begin{array}{l} q_\phi v\eta(t_{\phi p}(v,c))\frac{dt_{\phi p}(v,c)}{dc} \\ +q_p v\left(\eta(t_{pw}(v,c))\frac{dt_{pw}(v,c)}{dc} - \eta(t_{\phi p}(v,c))\frac{dt_{\phi p}(v,c)}{dc}\right) \\ -\eta(t_{pw}(v,c))\frac{dt_{pw}(v,c)}{dc} \end{array}\right] \\ &= -\left[\begin{array}{c} (q_\phi - q_p)\, v\eta(t_{\phi p}(v,c))\frac{dt_{\phi p}(v,c)}{dc} + \\ (q_p v - 1)\,\eta(t_{pw}(v,c))\frac{dt_{pw}(v,c)}{dc} \end{array}\right] \\ &= \left[\begin{array}{c} (q_p - q_\phi)\, v\eta(t_{\phi p}(v,c))\frac{dt_{\phi p}(v,c)}{dc} + \\ (1 - q_p v)\,\eta(t_{pw}(v,c))\frac{dt_{pw}(v,c)}{dc} \end{array}\right] \end{aligned}$$

Substituting for the relevant derivatives, (21) and (22), $dE\pi_m^0[c,\cdot|v]/dc \gtreqless 0$ as

$$\begin{array}{rl} (q_p - q_\phi)v\eta(t_{\phi p}(v,c))\frac{(q_p-q_\phi)v^2(1-\delta)}{\left[(q_p-q_\phi)v^2(\alpha+1)+c\right]^2} \gtreqless & \Leftrightarrow \\ (1-q_pv)\,\eta(t_{pw}(v,c))\frac{(1-q_pv)v(1-\delta)}{[(1-q_pv)v(\alpha+1)+\bar{c}-c]^2} & \\ \eta(t_{\phi p}(v,c))\frac{(q_p-q_\phi)^2v^2}{\left[(q_p-q_\phi)v^2(\alpha+1)+c\right]^2} \gtreqless & \Leftrightarrow \\ \eta(t_{pw}(v,c))\frac{(1-q_pv)^2}{[(1-q_pv)v(\alpha+1)+\bar{c}-c]^2} & \end{array}$$

$$\frac{\eta(t_{\phi p}(v,c))}{\eta(t_{pw}(v,c))}\frac{(q_p-q_\phi)^2v^2}{\left[(q_p-q_\phi)v^2(\alpha+1)+c\right]^2} \gtreqless \frac{(1-q_pv)^2}{\left[(1-q_pv)v(\alpha+1)+\bar{c}-c\right]^2}$$

Write $N(v,c) \equiv \frac{\eta(t_{\phi p}(v,c))}{\eta(t_{pw}(v,c))}$. Then the last expression is equivalent to

$$\sqrt{N(v,c)}\frac{(q_p-q_\phi)}{[(q_p-q_\phi)v^2(1+\alpha)+c]} \gtreqless \frac{(1-q_pv)}{[\bar{c}-c+v(1-q_pv)(\alpha+1)]}$$

and, $dE\pi_m^0[c,\cdot|v]/dc \gtreqless 0$ as

$$\begin{array}{rl} & \sqrt{N(v,c)}v(q_p-q_\phi)\left[(\bar{c}-c)+v(1-q_pv)(\alpha+1)\right] \\ \gtreqless & (1-q_pv)\left[(q_p-q_\phi)v^2(1+\alpha)+c\right]. \end{array}$$

That is, $dE\pi_m^0[c,\cdot|v]/dc \gtreqless 0$ as

$$\frac{v(q_p-q_\phi)\left(\bar{c}\sqrt{N(v,c)}+v(1-q_pv)(\alpha+1)\left(\sqrt{N(v,c)}-1\right)\right)}{\left[(1-q_pv)+v(q_p-q_\phi)\sqrt{N(v,c)}\right]} \gtreqless c \qquad (28)$$

By assumption, $\eta'(t) \leq 0$. Therefore, $N(v,c) \geq 1$ in which case the left side of this inequality is strictly positive. Hence, at $c = 0$, we have that for all $v < \hat{v}$,

$$\frac{dE\pi_m^0[0,\cdot|v]}{dc} > 0.$$

By definition of $N(v,c)$, at $c = \tilde{c}(v)$, $N(v,\tilde{c}(v)) = 1$. Substituting into the left side of (28) yields,

$$\begin{aligned} & \frac{v(q_p - q_\phi)\left(\bar{c}\sqrt{N(v,c)} + v(1-q_pv)(\alpha+1)\left(\sqrt{N(v,c)}-1\right)\right)}{\left[(1-q_pv) + v(q_p - q_\phi)\sqrt{N(v,c)}\right]} \\ = \; & \frac{v(q_p - q_\phi)\bar{c}}{[(1-q_pv) + v(q_p - q_\phi]} \\ = \; & \frac{v(q_p - q_\phi)\bar{c}}{(1-q_\phi v)} \equiv \tilde{c}(v). \end{aligned}$$

Thus, for all $v \in (0, \hat{v}]$, the first-order condition, $dE\pi_m^0[c, \cdot|v]/dc = 0$, is satisfied exactly at $c = \tilde{c}(v)$. Consider the second order condition. For all $c > 0$,

$$sgn\left[\frac{d^2E\pi_m^0[c,\cdot|v]}{dc^2}\right] =$$

$$sgn\left[\begin{array}{c} v\left(q_p - q_\phi\right)\left(\eta'(t_{\phi p}(v,c))\left(\frac{dt_{\phi p}(v,c)}{dc}\right)^2 + \eta(t_{\phi p}(v,c))\frac{d^2t_{\phi p}(v,c)}{dc^2}\right) + \\ (1-q_pv)\left(\eta'(t_{pw}(v,c))\left(\frac{dt_{pw}(v,c)}{dc}\right)^2 + \eta(t_{pw}(v,c))\frac{d^2t_{pw}(v,c)}{dc^2}\right) \end{array}\right]$$

and from (21) and (22),

$$\frac{d^2t_{\phi p}(v,c)}{dc^2} < 0; \; \frac{d^2t_{pw}(v,c)}{dc^2} < 0.$$

Hence, $\eta'(t) \leq 0$ implies $d^2E\pi_m^0[c,\cdot|v]/dc^2 < 0$ and choosing $c(v) = \tilde{c}(v)$ is a best response.

By definition of the penalty $\tilde{c}(v)$ and the constancy of the threshold $t_{\phi w}(v,(c,\bar{c}))$ in c, it now follows from Corollary 1 that, for any $c \in [\tilde{c}(v), \bar{c}]$, all employees of type $t < t_{\phi w}(v,(c,\bar{c}))$ strictly prefer staying silent either to reporting privately or to blowing the whistle, and all employees of type $t > t_{\phi w}(v,(c,\bar{c}))$ strictly prefer blowing the whistle either to reporting privately or staying silent. Therefore, the type zero manager is indifferent over imposing any penalty $c \in [\tilde{c}(v), \bar{c}]$ for reporting a violation $v < \hat{v}$ privately. □

Proof of Proposition 5. Suppose each managerial type $s \in \{0,1\}$ announces a different whistleblowing policy C_s^*. Let $\beta^* = 1$ (respectively, $\beta^* = 0$) describe the employee's belief that the manager is type one conditional on observing the message C_1^* (respectively, C_0^*). Then payoffs to the type one and type zero types of manager are given by (24) and (25), respectively.

Suppose the type one manager deviates and proposes the type zero manager's whistleblowing policy during the signaling stage. Then the employees believe she is a type zero manager for sure and respond as described in Proposition 4. But then there exists whistleblowing and many violations are not reported at all. Hence, such a deviation cannot prove profitable for the type one type of manager (in particular, by Lemma 2, if $\alpha\dot{v} \leq 4\bar{c}$ the type one manager obtains maximal feasible payoff by separating with C_1^*).

Suppose the type zero manager deviates and announces the same policy, $C' = C_1^*$ as the type one manager and employees react accordingly, believing she is a type one manager. Then the type zero manager's expected payoff from the deviation is

$$\begin{aligned} E\pi_m^0\left[C' = C_1^* | \beta^* = 1\right] \quad = \quad & -\int_0^{\dot{v}} \left(v\left(\alpha + \delta\right) q_\phi v \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt \right) dG(v) \\ & -\int_0^{\hat{v}} \left(v\left(\alpha + \delta\right) q_p v \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right) dG(v) \\ & -\int_{\hat{v}}^{\dot{v}} \left(v\alpha \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right) dG(v) \\ & -\int_{\dot{v}}^1 v\alpha dG(v) \end{aligned}$$

Then $\Delta_{sep} E\pi_m^0 = \left[\left(E\pi_m^0\left[C' = C_1^* | \beta^* = 1\right] - E\pi_m^0\left[C_0^* | \beta^* = 0\right]\right)\right]$ is

$$\begin{aligned} & -\int_0^{\dot{v}} \left(v\left(\alpha + \delta\right) q_\phi v \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt \right) dG(v) \\ & +\int_0^{\hat{v}} v(\alpha + \delta) \left(q_\phi v \int_0^{t_{pw}(v,(\tilde{c},\bar{c}))} \eta(t)dt + \int_{t_{pw}(v,(\tilde{c},\bar{c}))}^1 \eta(t)dt \right) dG(v) \\ & -\int_0^{\hat{v}} \left(v\left(\alpha + \delta\right) q_p v \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right) dG(v) \\ & +\int_{\hat{v}}^{\dot{v}} \left(v(\alpha + \delta) q_\phi v \int_0^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt + v\alpha \int_{t_{\phi p}(v,(0,\bar{c}))}^1 \eta(t)dt \right) dG(v) \\ & -\int_{\hat{v}}^{\dot{v}} \left(v\alpha \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right) dG(v) \\ & -\int_{\dot{v}}^1 v\alpha dG(v) + \int_{\dot{v}}^1 v\alpha dG(v) \end{aligned}$$

Collecting terms

$$\begin{aligned}
\Delta_{sep}E\pi_m^0 &= \int_0^{\hat{v}} \left(v(\alpha+\delta)q_\phi v \left[\int_0^{t_{pw}(v,(\tilde{c},\bar{c}))} \eta(t)dt - \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt \right] \right) dG(v) \\
&+ \int_{\hat{v}}^{\dot{v}} \left(v(\alpha+\delta)q_\phi v \left[\int_0^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt - \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt \right] \right) dG(v) \\
&+ \int_0^{\hat{v}} \left(v(\alpha+\delta) \left[\int_{t_{pw}(v,(\tilde{c},\bar{c}))}^1 \eta(t)dt - q_p v \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right] \right) dG(v) \\
&+ \int_{\hat{v}}^{\dot{v}} \left(v\alpha \left[\int_{t_{\phi p}(v,(0,\bar{c}))}^1 \eta(t)dt - \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt \right] \right) dG(v)
\end{aligned}$$

Recall

$$\begin{aligned}
t^*_{\phi p}(v,(\bar{c},0)) &= \frac{v\left(\alpha - q_\phi(\alpha+\delta)v\right) - \bar{c}}{v(\alpha+1)(1-q_\phi v) - \bar{c}}; \\
t_{\phi p}(v,(0,\bar{c})) &= \frac{v\left(\alpha - (\alpha+\delta)q_\phi v\right) + \bar{c}}{v(\alpha+1)(1-q_\phi v) + \bar{c}}; \\
t_{pw}(v,(\tilde{c},\bar{c})) &= \frac{v\left(1-q_\phi v\right)(\alpha+\delta) + \bar{c}}{v\left(1-q_\phi v\right)(\alpha+1) + \bar{c}}.
\end{aligned}$$

Where we (again) abuse notation obviously and we have substituted for $\tilde{c}$ in the expression for $t_{pw}(v,c)$ and collected terms. So, doing the tedious algebra, we find that for all $v \in (0,\hat{v})$, $t_{pw}(v,(\tilde{c},\bar{c})) = t_{\phi p}(v,(0,\tilde{c})) > t^*_{\phi p}(v,(\bar{c},0))$, and $t_{\phi p}(v,(0,\bar{c})) > t^*_{\phi p}(v,(\bar{c},0))$ on $(\hat{v},\dot{v})$ also. Therefore,

$$\begin{aligned}
\int_0^{t_{pw}(v,(\tilde{c},\bar{c}))} \eta(t)dt - \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt &= \int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{pw}(v,(\tilde{c},\bar{c}))} \eta(t)dt > 0; \\
\int_0^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt - \int_0^{t^*_{\phi p}(v,(\bar{c},0))} \eta(t)dt &= \int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt > 0; \\
\int_{t_{pw}(v,(\tilde{c},\bar{c}))}^1 \eta(t)dt - q_p v \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt &\gtreqqless 0; \\
\int_{t_{\phi p}(v,(0,\bar{c}))}^1 \eta(t)dt - \int_{t^*_{\phi p}(v,(\bar{c},0))}^1 \eta(t)dt &= -\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))} \eta(t)dt < 0.
\end{aligned}$$

Substituting, we can write $\Delta_{sep}E\pi_m^0$ as

$$\begin{aligned}\Delta_{sep}E\pi_m^0 \;=\;& \int_0^{\hat{v}}\left(v(\alpha+\delta)q_\phi v\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{pw}(v,(\tilde{c},\bar{c}))}\eta(t)dt\right)dG(v)\\ &+\int_{\hat{v}}^{\dot{v}}\left(v(\alpha+\delta)q_\phi v\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))}\eta(t)dt\right)dG(v)\\ &+\int_0^{\hat{v}}\left(v(\alpha+\delta)\left[\int_{t_{pw}(v,(\tilde{c},\bar{c}))}^{1}\eta(t)dt-q_pv\int_{t^*_{\phi p}(v,(\bar{c},0))}^{1}\eta(t)dt\right]\right)dG(v)\\ &-\int_{\hat{v}}^{\dot{v}}\left(v\alpha\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))}\eta(t)dt\right)dG(v);\end{aligned}$$

collecting terms,

$$\begin{aligned}\Delta_{sep}E\pi_m^0 \;=\;& (\alpha+\delta)\int_0^{\hat{v}}\left(v(1-q_pv)\int_{t_{pw}(v,(\tilde{c},\bar{c}))}^{1}\eta(t)dt\right)dG(v)\\ &-(\alpha+\delta)\int_0^{\hat{v}}\left(v^2\left(q_p-q_\phi\right)\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{pw}(v,(\tilde{c},\bar{c}))}\eta(t)dt\right)dG(v)\\ &-\int_{\hat{v}}^{\dot{v}}\left(v\left(\alpha-q_\phi v\left(\alpha+\delta\right)\right)\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))}\eta(t)dt\right)dG(v).\end{aligned}$$

There exists a separating equilibrium if and only if $\Delta_{sep}E\pi_m^0 \leq 0$. Dividing through this inequality by $(\alpha+\delta)$ and collecting terms one last time, therefore, there exists a separating equilibrium here if and only if

$$\begin{aligned}&\int_0^{\hat{v}}\left(v(1-q_pv)\int_{t_{pw}(v,(\tilde{c},\bar{c}))}^{1}\eta(t)dt\right)dG(v)\\ &-\int_0^{\hat{v}}\left(v\left(q_p-q_\phi\right)v\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{pw}(v,(\tilde{c},\bar{c}))}\eta(t)dt\right)dG(v)\\ &-\int_{\hat{v}}^{\dot{v}}v\left(q_\phi\left(\dot{v}-v\right)\int_{t^*_{\phi p}(v,(\bar{c},0))}^{t_{\phi p}(v,(0,\bar{c}))}\eta(t)dt\right)dG(v)\\ \leq\;& 0.\end{aligned}$$

Since $q_p > q_\phi$ and $\dot{v} > v$, the sum of the second and third terms is strictly negative. However, the first integral in v is strictly positive, leaving the sign of $\Delta_{sep}E\pi_m^0$ in general equivocal. But since, by Lemma 3, $\tilde{c}'(v) \geq 0$ and $\tilde{c}(v) < \bar{c}$, $t_{pw}(v,(\tilde{c}(v),\bar{c}))$ is strictly decreasing in v, and $\int_{t_{pw}(\hat{v},(\tilde{c},\bar{c}))}^{1}\eta(t)dt = 0$ implies $\Delta_{sep}E\pi_m^0 < 0$, there must exist a number $n > 0$ such that

$\Delta_{sep}E\pi_m^0 \leq 0$ for all distributions for which $\int_{t_{pw}(\hat{v},(\bar{c},\bar{c}))}^{1} \eta(t)dt \leq n$, as claimed (where again we use $t_{pw}(v,(\tilde{c}(v),\bar{c})) = t_{pw}(v,(\bar{c},\bar{c}))$, given $\tilde{c}(v)$ is the best response). □

References

Alford, C. F. (2001) *Whistleblowers: Broken Lives and Organizational Power,* Ithaca, NY: Cornell University Press.

Boatright, J. R. (2007) *Ethics and the Conduct of Business* (5e), Upper Sadle River, NJ: Pearson Prentice Hall.

Bowen, R., A. Call and S. Rajgopal (2008) "Whistle-Blowing: Target Firm Characteristics and Economic Consequences," Working Paper, University of Washington Business School.

Devine, T. (1997) *The Whistleblower's Survival Guide: Courage without Martyrdom,* Washington DC: Fund for Constitutional Government Accountability Project.

Dyck, A., A. Morse and L. Zingales (2007) "Who Blows The Whistle on Corporate Fraud?" Working Paper, University of Toronto.

Glazer, M. (1983) "Ten Whistleblowers and How They Fared," *The Hastings Center Report*, 13: 33-41.

Johnson, R. A. (2003) *Whistleblowing: When It Works and Why,* Boulder, CO: Lynne Reiner Publishers.

King, G. (1999) "The Implications of an Organization's Structure on Whistleblowing," *Journal of Business Ethics*, 20: 315-26.

Miceli, M. P., J. P. Near and C. R. Schwenk (1991) "Who Blows the Whistle and Why?" *Industrial and Labor Relations Review*, 45: 113-30.

Near J. P. and M. P. Miceli (1996) "Whistle-Blowing: Myth and Reality," *Journal of Management*, 22: 507-26

Sandler, J. (2007) "The War on Whistleblowers," Center for Investigative Reporting, November: http://centerforinvestigativereporting.org/articles/thewaronwhistleblowers.

Ting, M. (2008) "Whistleblowing," *American Political Science Review*, 102: 249-267.

Velasquez, M. G. (2002) *Business Ethics: Concepts and Cases* (5e), Upper Sadle River, NJ: Prentice Hall.