Wages, Performance and Sexual Harassment

Mehmet Bac*

April 24, 2016

Abstract

I study wage contracts that induce performance and contain co-worker sexual harassment in a hierarchy with heterogeneous agents. In contrast with the theory of compensating differentials, higher wages reduce harassment by raising the cost for harassers and attracting the agent types who file complaint if harassed. A tension arises between effort and anti-harassment objectives only at high levels of harassment where wages are low and collusion may occur. Better internal compliance structures (lower cost of filing complaint, accurate and speedy investigations) reduce the wage bill and/or the frequency of harassment but narrow down the range of feasible anti-harassment targets. In an organization with a performing workforce and an excellent internal compliance structure, wages must be high and harassment, extremely low.

JEL Classification Numbers: D86, K31, M52, J33, J79.

Key Words: contracts; employer liability; moral hazard; incomplete information.

*Sabanci University, Faculty of Arts and Social Sciences, Tuzla, Istanbul 34956, Turkey. E-mail: bac@sabanciuniv.edu.

Annual research grant from Turkish Academy of Sciences is gratefully acknowledged.
1 Introduction

Estimates of women who have experienced some form of sexual harassment in their workplace range between 40 and 75 percent (Aggarwal and Gupta, 2000.) Responding to the facts and public concerns, the law on sexual harassment in the U.S. has evolved with Supreme Court decisions, the Equal Employment Opportunity Commission’s (EEOC) guidelines in 1980 and the Organizational Sentencing Guidelines in 1991. Two broad categories of sexual harassment are defined: in a quid pro quo type a supervisor sexually harasses a subordinate under the threat of an adverse employment action. In hostile environment sexual harassment, there is no tangible employment action; the act typically involves co-workers, is unwelcome, severe and creates an abusive work environment. The majority of sexual harassment litigation falls into the second category, where the liability regime is negligence-based. The law basically sets the prevention and enforcement standards and leaves it to the employer to institute its own compliance structures and contractual measures to contain harassment.¹

There is a large literature outside economics on the causes of sexual harassment and its consequences on victims and the workplace environment, as well as a large legal literature on liability issues and effectiveness of sexual harassment laws. We know that sexual harassment reduces productivity by creating a hostile, unpleasant workplace environment. We have little understanding, however, as to whether and how contracts can cope with the threat of sexual harassment in conjunction with the internal compliance structures imposed by the law. Can employers induce specific performance and anti-harassment targets at the same time, even if harassment does not directly affect productivity? What modification in the wages can reduce the probability of co-worker harassment and encourage victims to file complaint? What role do internal compliance structures play in this mechanism?

I address these questions in a three-layer hierarchy model where a principal hires a supervisor and two agents, X and Y. In the first stage, the agents choose independently their effort levels and the supervisor employs a stochastic inspection

¹The employer is liable if it knew or should have known the incidence and was negligent in promptly taking the remedial steps. In practice negligence could be evidenced by the quality of internal anti-harassment compliance structures such as employee training, dissemination of the organization’s policies designed to prevent harassment and formal grievance procedures. The extremely high potential costs of a workplace harassment action have lead large employers to comply with the EEOC standards.
technology which may or may not produce hard evidence about the effort levels. Following the effort-inspection stage, Y may harass X. If harassed, X may stay silent or file complaint internally. In line with the EEOC guidelines, confirmation of harassment has a contractual consequence for Y, which in the model is equivalent to termination. The quality of the internal compliance structure is captured by three parameters, the cost of filing complaint, the probability with which the victim prevails, and the remedial effects of procedural factors such as the speed and accuracy of investigations on the victim. As such, the analysis applies as well to other, non-sexualized forms of co-worker mistreatment.

The victim’s harm and the harasser’s benefit are private information. In this context, contracts can in part serve to attract the high-harm X types who resist and file complaint if harassed. Although the unsavory high-benefit harassers cannot be screened out, if the contractual consequence of confirmed harassment is termination or an alternative financial sanction, wages serve as costs for, therefore deter, harassment. The principal has two contracts to design, one for the supervisor position and the other for the agent position, specifying a wage under each inspection report. The principal has two objectives, ensure that the agents exert high effort and induce a pair of harassment probabilities \((p_H, p_\emptyset)\), the first in the presence of high effort evidence and the second in the absence of it. I briefly explain the results and relate the paper to the literature below.

The two anti-harassment targets determine the agents’ wages under a high effort report and an empty report, \(w^H_A\) and \(w^\emptyset_A\). The higher the targets \(p_H\) and \(p_\emptyset\), the lower are \(w^H_A\) and \(w^\emptyset_A\). To induce a lower pair of harassment probabilities the principal raises these wages which, first, increases the potential cost for the harasser and second, raises potential victims’ contractual utility and attracts the high-harm whistleblower types into the workplace, thereby, raises the probability of a complaint upon harassment. This mechanism operates only in the presence of an effective grievance procedure—without the threat of an internal complaint, wages have no impact on sexual harassment.

Some anti-harassment targets are not compatible with the effort objective. Roughly put, \(p_H\) should not be set too high relative to \(p_\emptyset\), for otherwise the agent wage under a high effort report falls below the level that prevents shirking, primarily by the Y types whose harassment benefits are large. The tension between effort and anti-harassment objectives is stronger at lower levels of \(p_\emptyset\). That is, for low \(p_\emptyset\) targets, the maximal \(p_H\) that is compatible with inducing high effort is also low.
The intuition is that the probability of a complaint upon harassment is high when at least one of the two harassment probabilities is low, whereas at high $p_H$ targets the wage under a high effort report is small. A high probability of complaint coupled with a relatively low wage for high effort reduce the incentives to exert high effort and it becomes impossible to prevent shirking by the harasser types of Y. Thus, if $p_0$ is low, $p_H$ must also be low and the wages paid for a report of high effort, large enough.

Introducing the possibility of collusion does not change this feasibility result substantially; the qualitative effect is similar. Collusion cannot happen when inspection fails because hard evidence cannot be forged. But hard evidence can be concealed to submit an inconclusive report, in two occasions. A joint deviation to low effort to collude when the supervisor obtains low effort evidence can be prevented by appropriately raising the supervisor’s wage for reporting the effort evidence. This comes at no cost to the principal because low effort lies off the path induced by the contracts. In the other occasion, when the supervisor obtains evidence of high effort, the surplus from collusion is positive if $p_H$ is sufficiently higher than $p_0$ because the agent wage for a high effort report $w_H^A$ is relatively low. Though this type of collusion can be prevented by increasing the supervisor’s wage from reporting the high effort evidence, raising the wage bill for a high harassment probability $p_H$ can hardly be justified in the presence of a lower collusion-proof $p_H$ which can be induced at lower cost. Combining thus incentive compatibility with collusion-proofness considerations rules out high $p_H$ targets, in particular when $p_0$ is low.

These results shed light on the contractual mechanisms which contain sexual harassment in the workplace without sacrificing from effort objectives. The last set of results highlight the impact of an improvement in the internal compliance structures such as a fall in the cost of filing complaint, an increase in the probability that harassment victims prevail or a betterment in the investigation process. I show that these improvements raise the probability of a complaint, first by making the workplace more attractive for the high-harm whistleblower types, second by switching inframarginal victims from silence to filing complaint. As victims are more likely to file complaint, it is possible to implement the same anti-harassment objective by reducing the agents’ wages, or to reduce harassment given the wages. However, these same improvements narrow down the range of anti-harassment objectives that are compatible with high effort. If the harassment level is not low to
begin with, the principal may have to raise the wage under a high effort report and reduce the target $p_H$ to ensure that the high-benefit types of $Y$ do not shirk. In a sense, a better institutional infrastructure to combat sexual harassment confines the organization to a smaller, lower range of harassment targets.

The empirical literature on sexual harassment in organizations, though very large and multidisciplinary, is silent about the role of contractual incentives. The theory of incentives on the other hand has a variety of sophisticated models incorporating informational asymmetries and collusion, but not co-worker sexual harassment. The three-layer hierarchy model in this paper follows the approach initiated by Tirole (1986), Kofmann and Lawarrée (1993), with an extension to include a harassment stage played under two-sided incomplete information.

Various types of workplace offensive conduct have been studied in formal hierarchy or principal-agent models. The few contributions that bear on sexual harassment, Vafai (2010) and Bac (2015), deal with quid pro quo sexual harassment involving threats by a supervisor to modify the agent’s terms of employment. Vafai (2010) allows for collusion and sexual extortion, a form of abuse of authority, assuming that the supervisor’s threat not to report hard effort evidence is credible. Bac (2015) explicitly models a harassment stage where the supervisor diverts organizational resources to “buy” the agent’s silence for sex. Both papers show that the principal must leave rents to the employees in containing quid pro quo harassment. This paper is the first formal study of co-worker sexual harassment. Its hierarchical setup is rich enough to capture the institutional measures promoted by the EEOC compliance structures which emphasize the active role of victims as source of information. Besides enhancing our understanding as to how these internal compliance structures support the contractual measures, the paper highlights the range of harassment levels that can be induced without sacrificing from perfor-
formance objectives and generates predictions about the relationship between wages and the level of harassment.

The wage-harassment relationship is discussed in Posner (1999) and Basu (2003) with reference to employer liability. Ruling out wage discrimination (by the Equal Pay Act), the wages of the protected group are lower and the victims are instead compensated by the right to tort damages. While this function of the law is not contested, vicarious liability may also lead employers to supplement internal compliance measures with nondiscriminatory higher wages, better deter sexual harassment and economize on the potential damages they would pay if sued under Title VII. In a recent empirical study of the link between wages and sexual harassment, Hersch (2011) finds that women in jobs with an average probability of sexual harassment are paid 25 cents per hour more relative to comparable women employed in jobs with no risk of sexual harassment and interprets this as a compensating wage differential. The interpretation is valid in organizations with ineffective internal compliance structures where wages serve as ex-ante compensations for, rather than impacting on, harassment. Wages affect sexual harassment only in the presence of internal mechanisms which victims trust and easily use to seek redress, without fear. So, wages and harassment levels should be negatively correlated across organizations with similar and effective internal mechanisms, whereas organizations with better internal mechanisms can reduce both the level of harassment and the wage bill.

The next section presents the model. Section 3 derives the optimal contracts in the benchmark case without sexual harassment, followed by the analysis in sections 4, 5 and 6. Section 7 presents concluding remarks.

\footnote{Whether the EEOC-imposed internal structures are in practice effective or cosmetic is debated; see Krawiec (2005) and Dobbin and Kelly (2007). Ideally, empirical studies should incorporate and control for the quality of internal compliance structures in the wage-harassment equation. The sexual harassment data in Hersch (2011) consists of individual charges filed with federal or local authorities, the EEOC or the Fair Employment Practices Agency. If the compliance structures in these individuals’ institutions were effective, many of the victims in the data may not have charged a file with authorities outside their institutions. In organizations with ineffective internal compliance structures the impact of wages on sexual harassment is nil, or minimal.}
2 The Model

A hierarchical organization comprises a principal, a supervisor S and two agents, X and Y. All parties are risk-neutral, with outside options normalized to zero. The interaction between X, Y and S evolves through two stages, an inspection stage followed by a harassment stage.

In the inspection stage effort and inspection decisions are made simultaneously. Each agent chooses between high and low effort. High effort costs $e$ and low effort costs zero. The supervisor’s task is to inspect and report the outcome to the principal. Inspection costs $m$ and its outcome is stochastic, producing verifiable evidence about the effort pair with probability $\mu$, no evidence with probability $1 - \mu$. If the supervisor does not inspect, no effort evidence can be obtained. The inspection outcome is observed by X, Y and S.

Following the inspection stage but before S submits the report, Y may sexually harass X. The net harm resulting from this act depends on the parties’ privately known characteristics, i.e., types. The type $b$ of Y represents his private benefit drawn from a distribution with cdf $F(.)$ and support $R_+$. The type of X is denoted $h$, representing her personal harm from the act, drawn from a distribution with cdf $G(.)$, with support $R_+$.\(^5\) These benefits and harms are realized when the act is carried out.

If Y does not harass, X does nothing and the harassment stage ends. If Y harasses, X chooses between keeping silent and filing an internal complaint. Initiating the grievance procedure costs $L$ to X and the outcome of the investigation is uncertain: X prevails and harassment is confirmed with probability $\pi$. When that happens, I assume, X recovers a fraction $(1 - \beta)$ of the harm she suffered, where $0 < \beta < 1$, capturing the positive psychological effect of redress and justice.\(^6\) Of

\(^5\)F(.) and G(.) are continuous and differentiable. Negative $b$ values could be admitted to represent the Y types who derive a disutility from harassment. Differences in proclivity to sexually harass is well documented in the personality psychology literature. An early and influential measure developed by Pryor (1987) is the Likelihood to Sexually Harass Scale, constructed through self-reports as to how male respondents would behave in ten hypothetical harassment scenarios if assured that their behavior would not result in reprisals. The survey by Pina et. al. (2009) provides references to the extensive literature researching the traits of harassers.

\(^6\)The parameters $L$, $\pi$ and $\beta$ relate to what the EEOC and the Organizational Sentencing Guidelines refer as “internal compliance structures,” which impose standards on the organization’s internal anti-harassment policies and employee training as well as rules on formal harassment grievance procedures. These structures are costly but have a powerful effect on workplace
the four outcomes of the harassment stage, i.e., no harassment, harassment not reported, harassment reported but not confirmed, harassment reported and confirmed, the principal cannot differentiate between the first two, and given the fact that the third is legally equivalent to the first two, essentially the contractible outcomes of the harassment stage are, a confirmation of harassment or not. Wages depend only on contractible outcomes.

- **Sequence of events**

  At the outset the principal offers two contracts, one for the supervisor position and the other for the two agent positions. Following the binary acceptance/rejection choices, the principal picks a supervisor and two agents from those who accept and the relationship begins. X and Y agents are observationally distinguishable by gender. One agent employed must be an X, the other a Y.\(^7\) The sequence of events is as follows.

  - Contracts are offered; X, Y and S chosen.
  - X, Y and S play the inspection game and observe the outcome.
  - X and Y make their choices in the harassment stage.
  - S submits a performance report and, if X has filed a sexual harassment complaint, the investigation outcome is realized. Contracts are executed.

Later in the analysis I introduce collusion possibilities at an interim date between the second and the third bullets (inspection and harassment stages). Collusion is an agreement between S, X and Y on side transfers to misrepresent the inspection outcome; namely, S may suppress the effort evidence \((L \text{ or } H)\) and report \(\emptyset\) as inspection outcome. I rule out, however, the possibility of an intra-agent deal where Y buys out X’s future silence for money. Though in practice this kind of agreement is not unthinkable, the transaction costs involved should be extremely high because the parties are incompletely informed about each other’s harm and harassment, if properly maintained and enforced. They may also constitute evidence of non-negligence for the employer in failing to discover harassment. Better compliance structures have lower levels of \(L\) and \(\beta\) and higher levels of \(\pi\).

\(^7\)The one X - one Y assumption serves to simplify the exposition. The agents are employed in the same unit and same rank but may be performing two different tasks, one more suitable for the X agents and the other for Y agents.
benefit, but more importantly because it is extremely difficult for X to commit herself to silence after harassment.

- **Contracts**
  
  The last component of the model is the contract space. Contracts can be based solely on the supervisor’s report and the outcome of the internal harassment procedure, if initiated. Denote by $C_A = \{w^L_A, w^0_A, w^H_A\}$ and $C_S = \{w^L_S, w^0_S, w^H_S\}$ the agents’ and the supervisor’s contracts, comprising the wages in each of the three inspection outcomes. I assume limited liability, so, $w^k_i \geq 0$. Note that the agent contract does not depend on the agent’s identity: there is one agent contract. Discriminating between employees at the same rank and position on the basis of gender or other observationally distinguishable traits is prohibited (in the U.S., by the Equal Pay Act).\(^8\)

  The principal must institute a sanction to deter sexual harassment. Enforcing disciplinary actions accords with EEOC’s requirements and has a positive effect in defense of non-negligence in a potential Title VII law suit. The contracts accordingly include a clause specifying sexual harassment as cause for some form of financial punishment. As a working assumption, the punishment in the model is dismissal. Thus if agent Y is confirmed to have harassed X, his wage is zero no matter the supervisor’s effort report.\(^9\) Otherwise, Y’s contract $C_A = \{w^L_A, w^0_A, w^H_A\}$ is executed according to the supervisor’s effort report.

  To keep the focus on the interaction between performance incentives and the probability of sexual harassment, the analysis ignores the potential law suits that may follow. The victim may earn extra payoffs in damage awards by filing suit (to the employer under Title VII, possibly also to the harasser under civil law) which could reduce the probability of harassment and increase the probability of a complaint upon harassment.

\(^8\)In the majority of the cases the harasser is male and the victim is female. The law bans wage discrimination but protects potential victims with a right to tort damages by holding the employer liable for co-worker sexual harassment. See Posner (1999) for a discussion.

\(^9\)The victim’s wages, on the other hand, do not depend on the outcome of the investigation; paying a contractual reward/compensation for successful harassment complaints would obviously lead to an inflation of frivolous harassment charges. There are alternative instruments to encourage complaints (studied in Section 6) such as improving the efficacy of the internal grievance procedures, in accordance with the EEOC guidelines.
3 A Benchmark

The model without the possibility of harassment sets a useful benchmark in evaluating the expected cost of controlling harassment and the modification it entails on the optimal wage contracts. As the game now consists solely of the inspection stage, each agent’s participation and effort incentive constraint are respectively

\[
\mu w^H_A + (1 - \mu)w^0_A - e \geq 0; \quad \mu(w^H_A - w^0_A) \geq e. \tag{1}
\]

The left hand side of the first condition, the participation constraint, represents the expected wage net of the effort cost. The incentive constraint in the second condition states that to prevent a deviation to low effort the wage differential under high and low effort evidence should exceed the cost of effort augmented by the probability of verifiable effort evidence.

The participation and incentive constraints of the supervisor are

\[
\mu w^H_S + (1 - \mu)w^0_S - m \geq 0; \quad \mu(w^H_S - w^0_S) \geq m. \tag{2}
\]

The principal’s objective is to minimize the expected cost of inducing high effort, \(EP_0 = \mu(w^H_A + w^H_S) + (1 - \mu)(w^0_A + w^0_S)\), subject to (1) and (2), and the limited liability constraints \(w^i_j \geq 0\), for \(i = L, \emptyset, H\) and \(j = A, S\). Denote the optimal contracts in this case by \(\{w^L_A, w^0_A, w^H_A\}\) and \(\{w^L_S, w^0_S, w^H_S\}\). The following result is straightforward.

**Proposition 0** If the possibility of harassment is eliminated, the optimal agent contract is \(\{w^L_A, w^0_A, w^H_A\} = \{0, 0, e/\mu\}\) and the optimal supervisor contract is \(\{w^L_S, w^0_S, w^H_S\} = \{0, 0, m/\mu\}\), binding the participation constraints. The expected wage bill is \(EP_0 = e + m\).

In this benchmark case the principal offers contracts that bind the participation constraints of the two agents and the supervisor. The cost of inducing high effort is then minimized, \(EP_0 = e + m\). Because low effort evidence is off the path of play, the corresponding contractual wages are not paid, but \(w^L_A\), must be kept as low as possible to prevent shirking, hence, \(w^L_A\). On the other hand \(w^L_S\) is indeterminate because it plays no role in the supervisor’s incentive constraints in (2). Without loss of generality, then, \(w^L_S = 0\).\(^{10}\)

\(^{10}\)It is easy to see that these contracts are collusion-proof. There is no surplus for the three parties from misrepresenting the inspection outcome \(L\) or \(H\) as \(\emptyset\).
4 Behavior and utilities in the harassment stage

Contracts are powerless in extracting the information about individual attributes that shape behavior in the harassment stage because the types $h$ and $b$ are unrelated to inspection outcomes. A high-$b$ type of Y would accept any contract that a low-$b$ type accepts and a low $h$ type of X would accept any contract that a high $h$ type accepts. Thus, to any contract $\{w^L_A, w^9_A, w^H_A\}$ satisfying the agent’s effort incentive constraint corresponds a critical X type, denoted $h_A$ and defined by $EU_X(h_A) = 0$, such that all $h < h_A$ accept and all $h > h_A$ reject the contract. In this model the cost of containing co-worker sexual harassment partly stems from the need to raise $h_A$ and attract high-$h$ types of X (the potential victim group) from safer jobs or alternative outside options that pay less.

The harassment stage is thus played under two-sided incomplete information. I study the harassment subgame in reverse order of moves and proceed upward, deriving along the path the optimal behavior of each agent type and use these to construct the parties’ expectations about future moves, consistent with the strategies. Denote by $p_k$ the probability of harassment, by $q$ the probability that X files complaint conditional on harassment in subgame $k$ (when the inspection outcome is $k \in \{L, \emptyset, H\}$). Individual types play pure strategies, so these probabilities represent the fraction of types who adopt the corresponding actions.

- The decision to file complaint and harassment

Given an inspection outcome $k$, suppose that Y harasses. The utility of X from silence is $w^k_A - h - e$ whereas filing complaint yields the expected utility $w^k_A - L - (\pi \beta + (1 - \pi))h - e$. Thus X prefers filing complaint if

$$h > \frac{L}{\pi(1 - \beta)} \equiv h_C. \quad (3)$$

Note that X’s choice depends, besides of course her type, on the efficiency of the internal complaint procedure, captured by the parameters $L$, $\beta$ and $\pi$. A lower cost of filing complaint, a speedy process that reduces $\beta$ and improved accuracy in decisions in the form of a higher $\pi$ reduce $h_C$ and increase the measure of victims who blow the whistle.

Consider now the choice of Y, beginning with inspection outcome $k = H$. Y gets the utility $w^H_A - e$ from not harassing, the expected utility $b + [1 - q + (1 - \pi)q]w^H_A - e$
from harassing. He will choose to harass if
\[ b > \pi qw_A^H \equiv b_C^H. \]  \hfill (4)

The other inspection outcome is \( k = \emptyset \): inspection fails, no effort evidence is available. Agent Y will choose to harass if
\[ b > \pi qw_A^\emptyset \equiv b_C^\emptyset. \]  \hfill (5)

Then the probability of harassment in the subgame following inspection outcome \( k \in \{H, \emptyset\} \) is
\[ p_k = 1 - F(b_k^C). \]

Given these strategies, under high effort the expected utilities of X and Y are:
\[
EU_X = \begin{cases} 
\mu w_A^H + (1 - \mu)w_A^\emptyset - e - \overline{p}h & \text{if } h \leq h_C \\
\mu w_A^H(1 - \mu)w_A^\emptyset - e - \overline{p}(L + h(1 - \pi(1 - \beta))) & \text{if } h > h_C
\end{cases}
\]  \hfill (6)

where
\[ \overline{p} = \mu p_H + (1 - \mu)p_\emptyset; \]  \hfill (7)

\[
EU_Y = \begin{cases} 
\mu w_A^H + (1 - \mu)w_A^\emptyset - e & \text{if } b < b_C^\emptyset \\
\mu w_A^H + (1 - \mu)(w_A^\emptyset(1 - \pi q) + b) - e & \text{if } b \in [b_C^\emptyset, b_C^H) \\
(\mu w_A^H + (1 - \mu)w_A^\emptyset)(1 - \pi q) + b - e & \text{if } b \geq b_C^H.
\end{cases}
\]  \hfill (8)

The specification of \( EU_Y \) above assumes \( b_C^H \geq b_C^\emptyset \), where the Y types \( b \in [b_C^\emptyset, b_C^H) \) prefer to harass in subgame \( k = \emptyset \) but not in subgame \( k = H \). If \( b_A^H < b_A^\emptyset \), suffice it in (8) to interchange the positions of \( b_C^H \) and \( b_C^\emptyset \) and modify the second-row expected utility as \( \mu(w_A^H(1 - \pi q) + b) + (1 - \mu)w_A^\emptyset - e. \)

It is easy to verify that the participation constraint of Y is satisfied whenever the contract induces acceptance from X types. That is, if \( EU_X(0) \geq 0 \), then \( EU_Y(b) \geq 0 \), with strict inequality holding for \( b > 0 \).

• The marginal X type, \( h_A \)

If the probability of complaint is zero, \( q = 0 \), then \( p_k = 1 \) for \( k = H, \emptyset \) and harassment is maximal. To reduce harassment, it is essential that victims file complaint, that the agent contract attract a marginal \( h_A > h_C \) who finds it optimal to file complaint if harassed. Using the second line utility expression in (6), from \( EU_X(h_A) = 0 \) we get
\[
h_A = \frac{\mu w_A^H + (1 - \mu)w_A^\emptyset - e - L\overline{p}}{\overline{p}(1 - \pi(1 - \beta))}.
\]  \hfill (9)
When \( q > 0 \), the choices of X types are as illustrated below.

\[
\begin{array}{ccc}
0 & \bullet & h_c \\
\text{silent} & \text{file complaint} & \text{reject contract} \\
\end{array}
\]

The principal can raise \( q \) by increasing the wages \( w^H_A \) and \( w^\emptyset_A \), which shifts \( h_A \) to the right and reduces the proportion of silent X types. This adjustment is costly. When \( h_A > h_C \) and a positive \( q \) is induced, the expected wage will exceed the cost of effort by an amount proportional to the expected cost of filing complaint adjusted by the probability of relief:

\[
h_A > h_C \iff \mu w^H_A + (1 - \mu) w^\emptyset_A > \frac{\pi \lambda}{\pi (1 - \beta)} + e = \pi h_C + e. \tag{10}
\]

Not surprisingly, the benchmark agent contract \( \{w^L_A, w^K_A, w^H_A\} = \{0, 0, e/\mu\} \) violates (10); X rejects this contract if the probability of harassment is positive. A discrete increase in the wages \( w^\emptyset_A \) and \( w^H_A \) is required to induce acceptance by X types \( h > h_C \) and impact on harassment in \( \emptyset \) and \( H \) subgames.

*Incentive constraints*

The incentive constraint (1) in the benchmark case continues to apply for the X types. If X switches to low effort, with probability \( 1 - \mu \) she gets the same wage \( w^\emptyset_A \) but with probability \( \mu \) her wage becomes \( w^L_A \). Since her complaint strategy in the harassment subgame does not depend on the inspection outcome \( q \) is the same in subgames \( k = L, \emptyset, H \) the deviation to low effort has a utility impact only if the supervisor obtains effort evidence. So the X agent’s incentive constraint is as stated in (1):

\[
\mu (w^H_A - w^L_A) \geq e. \tag{11}
\]

For Y types, however, the incentive constraint needs to be modified because the wages are also potentially the prices he pays for harassment. In addition to economizing on the cost of effort, shirking may change this potential price as well. Define \( b^L_C = \pi q w^L_A \), analogous to \( b^H_C \) and \( b^K_C \) in (4) and (5), such that in subgame \( k = L \) (when S has evidence of low effort) Y will harass if and only if \( b \geq b^L_C \). Thus, Y may adopt different harassment strategies in the three subgames depending on his type \( b \) and the relative positions of \( b^L_C, b^K_C \) and \( b^H_C \). We have \( b^L_C \leq b^C_C < b^H_C \) if \( w^L_A \leq w^K_A < w^H_A \), the case illustrated below with four \( b \) intervals, in each of which the incentive constraint may take a different form.
Observe that the relative position of \( b_L^C \) and \( b_\emptyset^C \) is irrelevant for the form of the incentive constraint because when inspection fails to produce evidence Y’s utility does not depend on his effort choice; the wage is \( w_\emptyset^A \). Shirking has utility impact only in the presence of effort evidence. I verify below the incentive constraints case by case, eliminate those implied by others and the constraint that guarantees Y will exert high effort no matter his type.

- \( b \leq b_L^C \): When Y does not harass as in this case, his incentive constraint is the same as that of X, in (11).

- \( b \in (b_L^C, b_H^C) \): These \( b \) types harass if the inspection outcome generates low effort evidence, don’t harass if inspection reveals high effort. In the left neighborhood of \( b_H^C \), for example, the incentive constraint is

\[
\mu w_A^H + (1 - \mu)(w_A^0 + b(1 - \pi q)) - e \geq \mu w_A^L + (1 - \mu)w_A^0 + b(1 - \pi q).
\]

Given that the expected utility component for the no-evidence outcome is identical, the incentive constraint of these \( b \)-types simplifies to \( \mu(w_A^H - w_A^L)(1 - \pi q) - b \geq e \).

The left-hand side of this condition is decreasing in \( b \), reaching its the smallest value at \( b = b_H^C = \pi w_A^H q \) in this interval. Evaluated at that point, the incentive constraint is

\[
\mu(w_A^H - w_A^L)(1 - \pi q) \geq e.
\]

- \( b > b_H^C \): Y harasses no matter the outcome of inspection, so, deviation to low effort has utility impact when inspection reveals the effort. Therefore for \( b > b_H^C \) the incentive constraint is as stated in (12).

Note that (11) is implied by (12), which thus suffices to ensure that all types of X and Y who accept the contract exert high effort.

5 The targets and induced equilibria

The principal’s objective is to induce (i) high effort by both agents and (ii) a pair of target probabilities of harassment \((p_H^*, p_\emptyset^*)\).\(^{11}\) An agent contract \( \{w_L^A, w_\emptyset^A, w_H^A\} \)

---

\(^{11}\)The targets \((p_H^*, p_\emptyset^*)\) may differ because the corresponding subgames are reached with different probabilities; for example, priority may be given to harassment prevention in subgame \( H \) if \( \mu \) is
that satisfies (10) and the incentive constraint (12), and a supervisor contract \( \{w^S_H, w^S_, w^H_H\} \) that satisfies the participation and inspection constraints in (2), induce an equilibrium pair of harassment probabilities

\[
p^*_H = 1 - F(\pi w^H_A q^*) \quad \text{and} \quad p^*_0 = 1 - F(\pi w^0_A q^*)
\]

and a probability of complaint upon harassment

\[
q^* = 1 - \frac{G(h_C)}{G(h_A)}
\]

where \( h_A \) is defined in (9). Note that \( q^* \) is bounded above by

\[
q_m = 1 - G(h_C),
\]

which corresponds to \( h_A \to \infty \), hence, arbitrarily large \( w^H_A \) and/or \( w^0_A \). Thus, \( q^* \in [0, q_m) \). In an induced equilibrium \( \{p^*_H, p^*_0, q^*\} \), each player type’s strategy is individually optimal and beliefs are updated according to Bayes’ rule along any path that is followed with positive probability. When \( X \) and \( Y \) exert high effort and \( S \) inspects with probability one, the inspection stage does not convey any type information, so the beliefs at the beginning of the harassment stage are determined by contract acceptance strategies.

- **Inducing the targets** \( p^*_H \) and \( p^*_0 \)

Because the supervisor component of the optimal contract is relatively straightforward, the analysis below deals predominantly with the agents’ contract. Substituting (14) in (13), the equilibrium conditions can be expressed as

\[
p^*_k = 1 - F(\pi w^k_A [1 - \frac{G(h_C)}{G(h_A)}]), \quad k = H, \emptyset.
\]

The agent contract must satisfy (12) and (16) to meet the two objectives of the principal. The questions of interest are whether the principal’s performance and anti-harassment targets are compatible and, in the affirmative, what modification is needed in the contracts characterized in Proposition 0 to induce the anti-harassment targets.

large, subgame \( \emptyset \) if \( \mu \) is small. The target harassment probabilities would be the solution to a higher level optimization problem which I do not address, incorporating reputational, legal and other organizational costs of harassment. Endogenizing \( \{p^*_H, p^*_0\} \) is analytically uninteresting unless one explicitly incorporates the components these costs, which might in particular be linked to legal liability, but also to visibility of harassment and the organizational harm.
Let us focus first on (16), which is a system of two implicit equations and two unknowns. Proposition 1, proved in the Appendix, states that this system has a unique solution \( \{w_{\emptyset}^*, w_{H}^*\} \), given \( p_H^* \) and \( p_{\emptyset}^* \). Therefore, assuming high effort, the agent wages under \( \emptyset \) and \( H \) reports are fully determined by the anti-harassment targets through (16).

**Proposition 1** A unique pair \( \{w_{\emptyset}^*, w_{H}^*\} \) satisfies (16) given target probabilities \( (p_H^*, p_{\emptyset}^*) < (1,1) \).

Before proceeding with comparative statics I impose a technical condition about the magnitude of the second-order effects of a disturbance to (16) on \( p_k^* \). The condition guarantees unambiguous comparative statics results. To illustrate, consider the impact of a fall in \( w_{H}^* \) on \( p_H^* \); Figure 1 is useful for the intuition. The initial direct effect is to increase \( p_H = 1 - F(\pi w_{H}^* q) \), at constant \( q \). The fall in \( w_{H}^* \) coupled with the direct effect on \( p_H \) will reduce \( h_A \), hence, \( q \). This in turn will trigger adjustments in \( p_{\emptyset} \) and \( p_H \). The magnitude of these feedbacks on \( p_H \), shown by the broken arrow in Figure 1, are of a second order and must be smaller than the initial direct effect. Formally,

\[
dp_{k}^* > - \sum_{k\in\{\emptyset, H\}} f(b_{kC}^*) \pi w_{A_{k}}^* \frac{\partial q}{\partial p_{k}} \ dp_{k}^*.
\]

The left-hand side is the direct effect whereas the right hand side represents the second-order effects (positive, because \( \partial q/\partial p_{k} < 0 \)). In other words, for a unit rise \( dp_{k} = 1 \) the feedbacks on \( p_{k} \) through the changes in \( q \) sum to less than one.

Although (17) may not be transparent, as a stability condition it basically rules out a disproportionate \( q \) response to a change in the probability of harassment. That is, the partial derivatives in (17), \( \partial q/\partial p_{k} \), should not be too large, or, the type density functions \( g(.) \) and \( f(.) \) should not display large variations in the equilibrium region. If these densities are not too large, the feedbacks will be small because the response of \( h_A \), hence of \( q \) as well as the resulting impact on \( p_{k} \) through the change in \( b_{C}^* \), will be contained within reasonable bounds.\(^{12}\)

\(^{12}\)The mass of \( X \) types who switch from acceptance to rejection when the probability of harassment increases depends on the density of \( G(.) \), and given the resulting fall in \( q \), the mass of \( Y \) types who switch to harassment depends on the density of \( F(.) \). The condition (17) thus ensures that the higher-order effects of a disturbance form a vanishing sequence.
The following proposition, proved in the Appendix, states the equilibrium wage-harassment relationship as well as the wage adjustments needed to change the probability of harassment in the subgames with and without effort evidence.

**Proposition 2** To induce a lower \( p_k^* \) with \( p_{-k}^* \) constant, the principal must raise \( w_k^* \) and reduce \( w_{-k}^* \). Raising \( w_k^* \) alone induces a lower probability pair \( (p_H^*, p_0^*) \).

Moreover, \( p_H^* > p_0^* \iff w_A^H < w_A^0 \) and \( p_H^* = p_0^* \iff w_A^H = w_A^0 \).

A change in \( w_A^k \) affects \( p_k \) via the two channels illustrated in Figure 1: directly, by changing Y’s cost of harassment in the corresponding subgame \( k \), indirectly, by changing the probability of a complaint \( q^* \), which feeds back to Y’s decision to harass, \( p_k \) as well as \( p_{-k} \). So an increase in \( w_A^H \) reduces \( p_H = 1 - F(\pi w_A^H q) \) and increases the contractual expected utility of X, attracting a larger measure of X agent types, which increases \( h_A \) and \( q \). A higher probability of complaint \( q \), in turn, will feed back and further reduce harassment in both \( k = H \) and \( k = \emptyset \) subgames.

The final impact of an increase in \( w_A^H \) will be larger on \( p_H \) than \( p_\emptyset \) because only the indirect effect operates in subgame \( \emptyset \).

It follows that if the principal’s objective is to reduce \( p_H^* \) only and keep \( p_0^* \) constant, \( w_A^H \) should be increased in conjunction with an appropriate decrease in \( w_A^\emptyset \).

The last part of Proposition 2 states that a uniform anti-harassment target \( p_H^* = p_0^* \) is induced through the same agent wages under performance reports \( H \) and \( \emptyset \). Since X files complaint with the same probability \( q^* \) in subgames \( \emptyset \) and \( H \), this result follows directly from (13). Asymmetric anti-harassment targets are induced by adjusting the corresponding wages in the opposite direction.

**Effort Incentives**

Let us now include the principal’s performance objectives into the picture. The targets \( (p_H^*, p_0^*) \) are compatible with effort incentives under the contracts \( \{w_A^L, w_\emptyset^L, w_A^H\} = \{0, 0, m/\mu\} \) and \( \{w_A^L, w_A^0, w_A^H\} = \{0, 0, w_A^0, w_A^H\} \) if the solution \( (w_A^H, w_A^0) \) to the system of equations in (16) satisfies (12). Note that the contract component that plays the critical role for incentive compatibility is \( w_A^H \), paid under a high-effort report. This wage cannot be smaller than \( e/\mu(1 - \pi q^*) \) to satisfy (12), when \( w_A^L = 0 \). Fix a target harassment probability \( p_0^* < 1 \) and define

\[
p_H(p_0^*) = 1 - F\left(\frac{\pi q^* e}{\mu (1 - \pi q^*)}\right), \tag{18}
\]

\[\text{13}\] The qualitative effect of an increase in \( w_A^0 \) alone is the same but the quantitative effects will be reversed, with a larger reduction in \( p_0 \) than in \( p_H \), for the same reasons.
Given $p^*_0$ and the wage $w^{H^*}_A = e/\mu(1 - \pi q^*)$, the probability $p_H(p^*_0)$ solves (16) along with a wage for the empty inspection report, $w^0_A$. By propositions 1 and 2, $p_H(p^*_0)$ is well-defined; it is the probability of harassment that can be induced by paying the minimum wage $w^{H^*}_A = e/\mu(1 - \pi q^*)$ compatible with inducing high effort. The exact level of $p_H(p^*_0)$ depends on model parameters and the shape of the probability functions $G(.)$ and $F(.)$. Proposition 3 characterizes the set of incentive compatible anti-harassment targets.

**Proposition 3** For each $p^*_0 < 1$, the range $p^*_H \in (p_H(p^*_0), 1]$ is not compatible with the objective of inducing high effort. The lower bound $p_H(p^*_0)$ of this target range is increasing in $p^*_0$, from $1 - F(\frac{\pi q^*_m \mu}{\mu(1 - \pi q^*_m)})$ as $p^*_0 \to 0$, approaching 1 as $p^*_0 \to 1$.

[Figure 2]

As shown in Figure 2 the target combinations $\{p^*_H, p^*_0\}$ such that $p^*_H \leq p_H(p^*_0)$ are incentive compatible under the contracts $C^*_A = \{0, w^0_A, w^{H^*}_A\}$ and $C^*_S = \{0, 0, m/\mu\}$ induced by the agent contract $\{0, w^0_A, e/\mu(1 - \pi q^*)\}$. A target pair $\{p^*_H, p^*_0\}$ such that $p^*_H < p_H(p^*_0)$ is also incentive compatible because by Proposition 2 a lower $p^*_H$ with constant $p^*_0$ can be induced only by raising $w^{H^*}_A$ above $e/\mu(1 - \pi q^*)$, which therefore continues to satisfy (12), given $w^{L^*}_A = 0$.

Basically, asymmetric harassment probability combinations, too high in sub-game $H$ relative to subgame $\emptyset$, are not incentive compatible because the wage $w^{H^*}_A$ that corresponds to a sufficiently high $p^*_H$ is too low to prevent shirking, even if $w^{L^*}_A = 0$. Symmetric probabilities $p_H = p_\emptyset$ (corresponding to equal wages $w^{H^*}_A = w^0_A$) might be of particular interest to the principal. Sufficiently low symmetric target pairs definitely satisfy (12) because the wage $w^{H^*}_A$ that induces these targets is large enough, whereas parts of the diagonal in Figure 2 may lie above the upper bound $p_H(p^*_0)$ and violate incentive compatibility. The fact that a range of anti-harassment targets cannot be induced without sacrificing from effort objectives prevents a gradual transition from a full harassment situation to the region of lower harassment levels. The organization may have to reduce harassment substantially, below the $p_H(p^*_0)$ schedule, by raising the wages accordingly to a pair $\{w^0_A, w^{H^*}_A\} > \{0, e/\mu(1 - \pi q^*)\}$.

- Collusion
Collusion, an interim agreement between S, X and Y, involves side transfers to misrepresent the inspection outcome. Because hard effort evidence cannot be forged, collusion may occur only when S obtains effort evidence, under inspection outcome \( H \) or \( L \). If the parties collude, S suppresses/destroys the effort evidence, side transfers are made and the agents proceed to the harassment stage. Finally, S reports \( \emptyset \) and contracts are executed.

There are two occasions for, or types of, collusion. Type-(i) collusion occurs off the high-effort path, when S has low effort evidence, which is possible only if the agents deviate to low effort. Type-(ii) collusion occurs along the high-effort path, when S obtains evidence of high effort. If the parties collude, the agents play the harassment stage under the knowledge that wages will be \( w^\emptyset_A \), not \( w^H_A \). Both types of collusion can be prevented by adjusting the supervisor’s contract. The formal analysis is relegated to the proof of Proposition 4 in the Appendix.

**Proposition 4**

(i) Type-(i) collusion is prevented by appropriately raising the supervisor’s wage \( w^L_S \) for reporting low effort evidence above the benchmark level \( w^{L0}_S = 0 \).

(ii) Given \( p^*_0 \in (0, 1) \), there exists a critical \( p^*_H \),

\[
p^*_C_H = p^*_0 + \frac{m/\mu - 2(w^\emptyset_A - w^{H*}_A)}{L + h_A(1 - \pi(1 - \beta))}
\]  

(19)

such that the targets \( p^*_H \leq p^*_C_H \) are type-(ii) collusion-proof under the contracts \( \{C^*_A, C^*_S\} \). We have \( p^*_C_H > p^*_0 \), thus all \( p^*_H \leq p^*_0 \) are type-(ii) collusion-proof. To prevent collusion at \( p^*_H > p^*_0 \) the principal raises the supervisor’s wage \( w^H_S \) above the benchmark level \( w^{H0}_S = m/\mu \).

Under type-(i) collusion, the agents deviate to low effort with an expectation to bribe the supervisor for suppressing hard evidence of low effort. If they succeed, the agents play the harassment stage as if inspection outcome were \( k = \emptyset \), replacing the wages \( w^k_i \) by \( w^\emptyset_i \). Part (i) of Proposition 4 states that the principal can avoid this outcome by appropriately raising \( w^L_S \). This wage adjustment comes at no extra cost because \( w^L_S \) is paid with probability zero, off the path of induced play.

Part (ii) of Proposition 4 describes the impact of introducing the possibility of type-(ii) collusion and its prevention. A target pair \( (p^*_H, p^*_0) \) is collusion-proof under the corresponding contracts \( \{C^*_A, C^*_S\} \) characterized in Proposition 2 if the total surplus from suppression of high effort evidence is non-positive for all possible
agent type configurations. The largest total surplus obtains for the type $h_A$ of $X$ (whose disutility from harassment is largest, always files complaint) and the types $b \in [0, b_H^C]$ of $Y$ (who prefer not to harass). The collusion-proof target probabilities are described in Figure 2 by the area below the schedule $p_H^C$, defined in (19). Above the diagonal, for $p_H^* > p_0^*$, the total wage difference $2(w_A^0 - w_A^H)$ and the type-$h_A$ agent’s disutility difference are increasing in proportion with $(p_H^* - p_0^*)$: There is a net surplus to share for the agents if $S$ joins the coalition. At $p_H^* = p_H^C$ this net surplus of the agents is exactly equal to the supervisor’s loss $m/\mu$ from reporting $\emptyset$ instead of $H$. So, under the contracts $\{C_A^*, C_S^*\}$ $S$ will join the coalition if $p_H^* > p_H^C$.

Thus, while collusion occurs (if not prevented) with positive probability for $p_H^* > p_H^C$, all harassment probabilities below the diagonal in Figure 2 are immune to type-(ii) collusion because all agent types prefer playing subgame $H$ to subgame $\emptyset$: the agent wage is larger under report $H$ than report $\emptyset$, moreover the probability of harassment is smaller under report $H$ than report $\emptyset$. At the boundary harassment level $p_H^C$ the parties’ (X, Y and S) total utilities under reports $\emptyset$ and $H$ are equal. When $p_H^*$ is raised above $p_H^C$, total utilities under report $\emptyset$ rise whereas total utilities under report $H$ fall. To eliminate the risk of collusion the principal has no choice but to raise the supervisor’s wage for a high effort report above the benchmark level $m/\mu$. In essence this adjustment increases the parties’ total utilities under report $H$ up to the (now larger) total utilities under report $\emptyset$, which means that the expected wage bill at $p_H^* > p_H^C$ is larger than the expected wage bill at $p_H^* = p_H^C$. As it is difficult to rationalize a higher wage bill for a higher harassment probability, I conclude that the principal would never induce target combinations $(\bar{p}_H, \bar{p}_\emptyset)$ above the $\bar{p}_{CP}$.

• Summary

Excluding from the principal’s objectives the harassment probabilities in the

---

14 More precisely, given constant $p_0^*$, to induce a $p_H^*$ above $p_H^C$, the principal must reduce $w_A^H$ and raise $w_A^0$ as stated in Proposition 2. This generates incentives to collude and change the inspection outcome $H$ as $\emptyset$ because $X$ and $Y$’s total expected utilities in subgame $\emptyset$ increases whereas their total expected utilities in subgame $H$ falls. This holds for all $b$ types of $Y$ and for all $h$ types of $X$ who accept the contract because $p_H^* > p_0^*$ implies $w_A^H > w_A^0$ while $q^*$ is the same in the two subgames. The agent’s contract cannot be modified because the wages $w_A^0$ and $w_A^H$ are tied to the target harassment probabilities. The principal has one instrument to prevent collusion, the wage $w_S^H$ paid when $S$ reports evidence of high effort, which must be raised to a level $w_S^{HC}$ such that the total utilities in subgame $H$ are at least as large as the total utilities in subgame $\emptyset$. 

19
type-(ii) collusion region and those that violate the effort incentive constraint leaves us with the feasible range of pairs \( (p_H^*, p_∅^*) \), that is, \( p_H^* \leq \min\{p_H(p_H^*), p_C^\} \). The corresponding contracts that induce high effort are \( C_S^* = \{w_S^C, 0, m/µ\} \) and \( C_A^* = \{0, w_A^0, w_A^H\} \) where the wage \( w_S^C \) may be positive or not, depending on the parties’ surplus from type-(i) collusion. Thus, in this feasible range of anti-harassment targets \( C_S^* \) is unchanged except for the possibility of type-(i) collusion, but \( C_A^* \) has higher wages than the reference case in Proposition 0. Introducing a feasible anti-harassment component into the principal’s effort objective entails a discrete jump in the expected wage bill. The principal must raise the wages \( w_A^0 \) and \( w_A^H \) sufficiently above their reference levels to attract those X types with \( h > h_C \) so that a complaint upon harassment is filed with positive probability. The high wages also impact on harassment by serving as opportunity costs for the harassers.

The model’s prediction about the wage-harassment relationship is in contrast with the theory of compensating wage differentials, namely, that wages should be higher for potential victims in jobs with higher probability of harassment. The law rules out compensating wage differentials for potential victims (protected groups) and institutes instead employer liability for workplace sexual harassment. That is, employer liability is viewed as a substitute for higher, compensating wages to potential victims (e.g., Posner 1999, Basu 2003). This argument is based on a labor market equilibrium with homogeneous victims and firms with exogenous harassment levels. It thus ignores heterogeneity of potential victims and the legal internal compliance mechanisms by which employers must abide to avoid vicarious liability in potential law suits. There is evidence that large-size firms in particular institute these compliance mechanisms, so it seems that the threat of legal liability is effective on employers. In the affirmative, this would lead them to induce low harassment levels by paying higher, not lower, wages, converting the high risk - low pay workplace into a low risk - high pay workplace. In the next section I probe into the impact of a betterment in the internal compliance structures which, as institutional responses to vicarious liability, aim at preventing the harms from harassment in the workplace.

\[15\] Large firms in particular widely comply with the legal standards on internal grievance procedures because they serve to contain harassment and may constitute evidence of non-negligence on the part of the firm. Allen and Montgomery (2001) report that seventy-five percent of Fortune 500 firms and thirty-six percent of other firms had structures of this sort by 1998.
6 Impact of internal compliance structures

According to the Organizational Sentencing Guidelines adopted in 1991, the employer has a duty to disseminate and apply a procedure by which victims of harassment can file complaint without fear of retaliation, as well as instituting an investigation mechanism which delivers without delay a decision on any report of possible harassment. In the present model the requirement that firms widely disseminate their anti-harassment policy and grievance procedures affects $L$, the victim’s cost of filing complaint. Reducing the delays in processing the complaints reduces $\beta h$, the final net harm of the victim as a function of her type $h$. Internal monitoring and auditing systems and the Equal Employment Opportunity guidelines that detail the enforcement standards on information-gathering and investigation procedures raise $\pi$ by improving the accuracy of investigation outcomes.

The impacts of the institutional changes of the types mentioned above work primarily through the utilities and incentives of potential victims. A reduction in $L$ or $\beta$, or a rise in $\pi$, generate two distinct effects that reinforce each other. The first effect comes as a fall in the critical $X$ type $h_C$ who is indifferent between silence and filing complaint upon harassment. Other things equal, the fall in $h_C$ raises the probability of a complaint, $q$. This is an inframarginal effect. The second effect works by raising the contractual utility of potential victims, which leads to an increase in the marginal $X$ type, $h_A$, who is indifferent between accepting and rejecting the given contract. The rise in $h_A$ enlarges the range of whistleblower $X$ types, $h_A - h_C$, hence, also increases $q$. As both effects lead to a higher $q$, the contract now implements a smaller pair of harassment probabilities. Alternatively the principal can modify the contracts and economize on the expected wage bill for the same pair of harassment probabilities.

**Proposition 5** A reduction in $L$ or $\beta$, or an increase in $\pi$ leads to:

(i) lower wages $w_A^*, w_H^*$ for a given target pair $(p_H^*, p_\emptyset^*)$, or lower harassment probabilities $(p_H^*, p_\emptyset^*)$ given the agent contract;

(ii) a fall in the upper bound $\bar{p}_H(\bar{p}_\emptyset)$ of the range of incentive-compatible targets.

The proof is in the Appendix. Part (ii) of Proposition 5 states that the same improvements in compliance structures which produce favorable effects on the wage bill and/or the level of harassment also narrow down the range of harassment probabilities compatible with high effort. To grasp the intuition, suppose that the principal continues to induce the same harassment probability but reduces the wages...
in response to the improvement in the compliance structure. Recall that when S obtains evidence of high effort, with probability $\pi q$ Y will pay the “price” $w_A^{H*}$ if he harasses X. As the principal reduces $w_A^{H*}$ in response to the improvement, so, as $q$ rises accordingly, $\pi q$ increases as well. Then the effort incentive constraint (12) may be violated, primarily for the high-$b$ types of Y. More precisely the high-$b$ harasser type may now prefer withholding effort and harassing in subgame $L$ rather than exerting effort and harassing subgame $H$ because, given his high personal benefit from harassment and the cost effort to economize, the reduction in the wage $w_A^{H*}$ and the rise in $\pi q$ reduce the potential price difference for harassment between subgames $H$ and $L$. As a result, the fixed harassment probability in subgame $H$ may no longer be compatible with effort incentives when $L$ or $\beta$ is decreased $\pi$ increased. This is true especially for high harassment probabilities close to the upper bound $p_H(p_0)$ of the feasible range.

Generally, a change in the workplace parameters favorable to potential victims brings about economies in the contractual costs but narrows down the range of targets compatible with the organization’s effort objectives. Changes of this kind thus encourage the organization to implement lower harassment targets with the same contracts, but organizations with low harassment levels (much below the upper bound $p_H(p_0)$ of the feasible range) can also exploit the gains by modifying the contracts to reduce their wage bill.

7 Conclusion

This paper studies the role of employment contracts in containing co-worker sexual harassment in a three-layer hierarchy model with moral hazard. Victims are the sources of information, contracts and the internal compliance structure are the enforcement tools for encouraging complaints and punishing sexual harassment. By extending the theory of incentives in organizations to incorporate co-worker harassment under the internal compliance structures promulgated by the laws, the paper offers new theoretical predictions and fills an important gap.

Contrary to the theory of compensating differentials which predicts higher wages in high-harassment environments, in this model higher wages serve to contain harassment in the presence of effective internal compliance structures. By raising the power of effort incentives the organization attracts whistleblower types and breaks the victims’ silence and increases, on the other hand, the opportunity cost for ha-
rassers. Higher wages for good performance are instruments of control, rather than mere compensations for enduring an exogenous workplace harassment level.

This negative correlation between wages and the induced frequency of harassment implies that high levels of harassment may be incompatible with the organization’s effort objective. The result is not an artifact of a postulated negative impact of sexual harassment on performance. The negative wage-harassment correlation arises because at high levels of harassment the wages for good performance are low, therefore the price for harassment under a high effort report is too low to prevent shirking. The high-benefit harasser types—who cannot be screened out because observationally cannot be distinguished from non-harassers—are the first to deviate to low effort. In addition, there is scope for collusion to misreport the inspection outcome at high levels of harassment. Even in a world in which sexual harassment does not directly harm effort incentives, an organization with a performing agent force is unlikely to display high harassment for incentive or collusion-prevention reasons, or both.

The paper also sheds light on the channels through which internal compliance structures affect co-worker sexual harassment. Better rules and grievance procedures serve to break the victims’ silence and punish the harassers, thereby, economize on contractual costs or reduce sexual harassment for a given set of contracts by raising the contractual utility of potential victims and the likelihood of complaints upon harassment. Interestingly, better compliance structures reduce the range of anti-harassment targets that are compatible with effort incentives. An implication of these findings is that an organization with a performing workforce and exemplary internal compliance structure must display extremely low levels of sexual harassment—must, because only then can the organization meet its performance objective. To preserve effort incentives when its internal compliance structures improve the organization may also have to raise the wage paid for high performance and induce a lower probability of harassment, rather than preserving its anti-harassment objectives to economize on the wage bill.

Appendix

Proof of Proposition 0: Consider the agent contract \( \{ w^L_A, w^0_A, w^H_A \} \). If positive, reducing \( w^L_A \) would relax the incentive constraint without affecting the participation constraint. Therefore \( w^{L0}_A = 0 \). Given this, observe that the partici-
Optimal contract is \( w_A^H + (1 - \mu)w_A^\theta - \varepsilon \geq 0 \) differs from the incentive constraint \( \mu w_A^H - \varepsilon \geq 0 \) in the extra term \( (1 - \mu)w_A^\theta \). Setting \( w_A^\theta = 0 \) is therefore optimal. Therefore \( \{w_A^{L0}, w_A^{\theta0}, w_A^{H0}\} = \{0, 0, \varepsilon/\mu\} \), as stated in the proposition.

Using the same arguments it is straightforward to verify that the supervisor’s optimal contract is \( \{0, 0, m/\mu\} \), binding the participation and incentive constraints. Q.E.D.

**Proof of Proposition 1.** Fix a target pair \((p_H, p_\theta)\) and consider the expressions in (16), reproduced below with substitutions for \( h_C, h_A \) and \( q \):

\[
p_H = 1 - F(\pi w_A^H [1 - \frac{G(L/\pi(1 - \beta))}{\frac{G\left(\frac{w_A^H + (1 - \mu)w_A^\theta - \varepsilon qL}{\bar{\rho}(1 - \pi(1 - \beta))}\right)}}]) \tag{20}
\]

\[
p_\theta = 1 - F(\pi w_A^\theta [1 - \frac{G(L/\pi(1 - \beta))}{\frac{G\left(\frac{w_A^H + (1 - \mu)w_A^\theta - \varepsilon qL}{\bar{\rho}(1 - \pi(1 - \beta))}\right)}}]) \tag{21}
\]

where \( \bar{\rho} = \mu p_H + (1 - \mu)p_\theta \). Observe that the expressions in (20) and (21) are continuous in \( w_A^H \) and \( w_A^\theta \). The system of two equations must therefore have a fixed point, \((w_A^{H*}, w_A^{\theta*})\).

To show uniqueness, consider the \( \mathbb{R}^2 \) plane representing \( w_A^H \) on the \( y \) axis and \( w_A^\theta \) on the \( x \) axis, where I define two schedules: Let \( w_A^H(w_A^\theta, p) \) denote the wage pairs \((w_A^H, w_A^\theta)\) that satisfies (20) given \( p = (\bar{p}_H, \bar{p}_\theta) \). Similarly, define the schedule \( w_A^\theta(w_A^H, p) \) through (21). Thus, a fixed point \( w_A^{H*} = w_A^H(w_A^{\theta*}, p) \) and \( w_A^{\theta*} = w_A^\theta(w_A^{H*}, p) \) induces the targets \( p = (p_H, p_\theta) \).

Consider the limit values of \( w_A^H(w_A^\theta, p) \) as \( w_A^\theta \) is changed in its domain, \([0, \infty)\), given fixed targets \( p \). Observe that as \( w_A^\theta \) becomes very large, the value of \( G(h_A) \) in (20) approaches 1 and therefore \( q \) approaches its maximal value \( q_m \). The same holds for the limit value of \( w_A^0(w_A^H, p) \) as \( w_A^H \) approaches \( \infty \). Therefore \( \lim_{w_A^H \to \infty} w_A^H(w_A^\theta, p) = w_A^h \), where \( w_A^h \) satisfies \( p_H = 1 - F(\pi w_A^h, q_m) \). Similarly, \( \lim_{w_A^H \to \infty} w_A^\theta(w_A^H, p) = w_A^\infty \), where \( w_A^\infty \) satisfies \( p_\theta = 1 - F(\pi w_A^\infty, q_m) \). Define also the limits at the other extreme, \( \lim_{w_A^\theta \to 0} w_A^H(w_A^\theta, p) = w_0^H \) and \( \lim_{w_A^H \to 0} w_A^\theta(w_A^H, p) = w_0^\theta \). The schedules \( w_A^H(w_A^k, p) \) are decreasing in \( w_A^{-k} \), therefore \( w_0^H > w_A^h \) and \( w_0^\theta > w_A^\theta \). Figure 3 illustrates these limits and the two schedules.

[Figure 3]
Differentiating (20) and (21) at a fixed point \((w^H_A, w^\emptyset_A)\) yields

\[
\frac{dw^H_A}{dw^\emptyset_A} = - \frac{w^H_A \frac{\partial q}{\partial w^\emptyset_A}}{q + w^H_A \frac{\partial q}{\partial w^\emptyset_A}} \quad \text{and} \quad \frac{dw^H_A}{dw^\emptyset_A} = - \frac{q + w^\emptyset_A \frac{\partial q}{\partial w^\emptyset_A}}{w^\emptyset_A \frac{\partial q}{\partial w^\emptyset_A}}. \tag{22}
\]

The expressions in (22) are the slopes (both negative) of the two schedules, respectively, \(w^H_A(w^\emptyset_A, p)\) and \(w^\emptyset_A(w^H_A, p)\), in the \((x, y) = (w^\emptyset_A, w^H_A)\) plane. Evaluated at a fixed point \((w^H_A, w^\emptyset_A)\) and given \(p\), the partial derivatives \(\partial q/\partial w^\emptyset_A\) at the right hand sides of the slopes in (22) must be identical. Using this fact in (22) yields immediately the result

\[
\frac{[dw^H_A/dw^\emptyset_A]_{k=H}}{[dw^H_A/dw^\emptyset_A]_{k=\emptyset}} = \frac{w^H_A (\partial q/\partial w^\emptyset_A) w^\emptyset_A (\partial q/\partial w^H_A)}{[q + w^H_A (\partial q/\partial w^H_A)][q + w^\emptyset_A (\partial q/\partial w^\emptyset_A)]} < 1.
\]

At any fixed point, the slope of \(w^H_A(w^\emptyset_A, p)\) is unambiguously smaller than the slope of \(w^\emptyset_A(w^H_A, p)\) in absolute value. Given the limit values of the two schedules, it follows that there can be at most one fixed point satisfying \(w^H_A = w^H_A(w^\emptyset_A, p)\) and \(w^\emptyset_A = w^\emptyset_A(w^H_A, p)\) as shown in Figure 3.

**Proof of Proposition 2**: Total differentiation of the equilibrium equation (16) yields:

\[
dp_k = -f(b^C_C)\pi[q.dw^A_A + w^k_A\{\frac{\partial q}{\partial w^A_A}dw^k_A + \frac{\partial q}{\partial p_H}dp_k + \frac{\partial q}{\partial p_k}dp_k\}] \tag{23}
\]

where \(-k = H\) if \(k = \emptyset\) and vice versa. The first part of the proposition is a statement about the wage adjustments needed to change \(p_H\) with constant \(p_\emptyset\).

Setting \(dp_\emptyset = 0\) and rearranging the terms in (23) for \(k = H\) and \(k = \emptyset\) separately yields the two equations below:

\[
(1 + f(b^C_C)\pi w^H_A \frac{\partial q}{\partial p_H})dp_H = -f(b^C_C)\pi[(q + w^H_A \frac{\partial q}{\partial w^H_A})dw^H_A + w^H_A \frac{\partial q}{\partial w^\emptyset_A}dw^\emptyset_A]
\]

\[
0 = (q + w^\emptyset_A \frac{\partial q}{\partial w^\emptyset_A})dw^\emptyset_A + w^\emptyset_A \frac{\partial q}{\partial p_H}dp_H + w^\emptyset_A \frac{\partial q}{\partial w^H_A}dw^H_A.
\]

Solving for \(dw^\emptyset_A\) from the second equation, substituting into the first and simplifying the terms we get

\[
[q + w^\emptyset_A \frac{\partial q}{\partial w^\emptyset_A} + f(b^C_C)\pi w^H_A \frac{\partial q}{\partial p_H} q]dp_H = -f(b^C_C)\pi q\widetilde{X} dw^H_A
\]

where \(X = q + w^\emptyset_A \frac{\partial q}{\partial w^\emptyset_A} + w^H_A \frac{\partial q}{\partial w^H_A}\). The coefficient of \(dw^H_A\) is negative because the partial derivatives in \(\widetilde{X}\) are positive, whereas the coefficient of \(dp_H\) at the left hand
side is positive because $1 + f(b_H^i)\pi w_A^H \frac{\partial q}{\partial p_H} > 0$, by (17). Therefore $dw_A^H > 0$ if $dp_H < 0$. Given this, $dw_A^\emptyset < 0$ follows from the total differentiation equation above.

The second part of the proposition is on the impact of a change in one of the wages on $p_H$ and $p_\emptyset$. Consider an increase in $w_A^H$ with constant $w_A^\emptyset$. Setting $dw_A^H > 0$ and $dw_A^\emptyset = 0$ in (23) yields the two equations below:

$$
(1 + f(b_H^i)\pi w_A^H \frac{\partial q}{\partial p_H}) dp_H = - f(b_H^i)\pi [q + w_A^H \frac{\partial q}{\partial w_A^H} dw_A^H + w_A^H \frac{\partial q}{\partial p_\emptyset} dp_\emptyset]
$$

$$
(1 + f(b_\emptyset^i)\pi w_A^\emptyset \frac{\partial q}{\partial p_\emptyset}) dp_\emptyset = - f(b_\emptyset^i)\pi w_A^\emptyset \frac{\partial q}{\partial w_A^H} dw_A^H + \frac{\partial q}{\partial p_H} dp_H.
$$

Substituting from the second equation for $dp_\emptyset$ and rearranging terms, the first equation becomes

$$
[1 + f(b_H^i)\pi w_A^H \frac{\partial q}{\partial p_H} + f(b_\emptyset^i)\pi w_A^\emptyset \frac{\partial q}{\partial p_\emptyset}] dp_H = - f(b_H^i)\pi [q + f(b_\emptyset^i)\pi w_A^\emptyset \frac{\partial q}{\partial w_A^H} + w_A^H \frac{\partial q}{\partial w_A^H}] dw_A^H.
$$

The coefficient of $dp_H$ at the left hand side is positive by (17), whereas the coefficient of $dw_A^H$ at the right hand side is negative. Therefore $dw_A^H > 0$ implies $dp_H < 0$. Using this result and the fact that $\partial q/\partial p_H < 0$ and $\partial q/\partial w_A^H > 0$ in the second total differentiation equation yields $dp_\emptyset < 0$. The proof of the symmetric case of a fall in $p_\emptyset$ with $p_H$ constant is analogous.

The last part of the proposition, $p_H \geq p_\emptyset \Rightarrow w_A^H \geq w_A^\emptyset$, follows directly from the equilibrium conditions in (13), reproduced below:

$$
\bar{p}_H = 1 - F(\pi q^* w_A^H); \quad \bar{p}_\emptyset = 1 - F(\pi q^* w_A^\emptyset),
$$

where $q^*$ is identical in both subgames, $k = H$ and $k = \emptyset$. Thus, if $p_H \geq p_\emptyset$ the solution to (16) must satisfy $w_A^H \geq w_A^\emptyset$, with equality holding if $p_H = p_\emptyset$.

Q.E.D.

Proof of Proposition 4: Consider first type-(i) collusion, where $S$ has evidence of low effort. Given the contracts $C_A^*$ and $C_S^*$, denote by $u_i^kd$ the continuation expected utility of agent $i = X, Y$ under low effort, viewed from the end of the inspection stage outcome $k = L, \emptyset$. Agent $i$ gets the surplus $u_i^kd - u_i^{Ld}$ if $S$ reports $\emptyset$ instead of $L$. This cannot happen if the loss of $S$ from misrepresenting the inspection outcome exceeds the agents’ gain, i.e., if

$$
u_S^L - u_S^\emptyset = \bar{w}_S^L - w_S^\emptyset \geq \sum_{i=X,Y} \text{max}\{0, (u_i^\emptyset - u_i^{Ld})\}.$$
The surplus \( u_{i}^{bd} - u_{i}^{Ld} \) is a finite magnitude under any feasible contract. If the above condition does not hold, the principal cannot modify the agent wages in the contract \( C_{A}^{*} \) because \( w_{A}^{H*} \) and \( w_{A}^{b*} \) are fixed by the target harassment probabilities through (16), whereas \( w_{A}^{L*} = 0 \) and therefore cannot be reduced any further. This leaves the principal with one instrument, the supervisor’s contract \( C_{S}^{*} \), to prevent type-(i) collusion. Accordingly, define a minimum \( w_{S}^{L} \) through

\[
 w_{S}^{Cd} = w_{S}^{0} + \sum_{i=X,Y} \max\{0, u_{i}^{bd} - u_{i}^{Ld}\} \tag{24}
\]

so that \( S \) truthfully reports \( k = L \) when \( w_{S}^{L} = w_{S}^{Cd} \). Clearly, the wage \( w_{S}^{Cd} \) in (24) prevents type-(i) collusion. It exceeds the wage \( w_{S}^{L0} = 0 \) whenever \( u_{i}^{bd} - u_{i}^{Ld} \) is positive, as stated in the proposition.

Consider now the possibility of type-(ii) collusion, when \( S \) obtains evidence of high effort. The net continuation surplus of \( S \) from reporting \( k = H \) over reporting \( k = \emptyset \) is

\[
 u_{S}^{H} - u_{S}^{0} = w_{S}^{H} - w_{S}^{0}.
\]

As for the agents’ net benefits from performance report \( k = H \) over \( k = \emptyset \), they depend on their types, as follows:

\[
 u_{X}^{H} - u_{X}^{0} = \begin{cases} 
 (w_{A}^{H} - w_{A}^{0}) - (p_{H} - p_{0})h & \text{if } h \in [0, h_{C}] \\
 (w_{A}^{H} - w_{A}^{0}) - (p_{H} - p_{0})[L + h(1 - \pi(1 - \beta))] & \text{if } h \in (h_{C}, h_{A}]
\end{cases}
\]

for \( X \), and by

\[
 u_{Y}^{H} - u_{Y}^{0} = \begin{cases} 
 (w_{A}^{H} - w_{A}^{0}) & \text{if } b < \min\{b_{C}^{H}, b_{C}^{0}\} \\
 (w_{A}^{H} - w_{A}^{0})(1 - \pi q^{*}) & \text{if } b \geq \max\{b_{C}^{H}, b_{C}^{0}\}
\end{cases}
\]

for \( Y \). The surplus expression for the \( Y \) types in the intermediate range missing above depends on whether \( b_{C}^{H} > b_{C}^{0} \) or \( b_{C}^{H} < b_{C}^{0} \). \( \text{See (4) and (5) and note that } b_{C}^{H} > b_{C}^{0} \iff w_{A}^{H} > w_{A}^{0}. \) If \( b_{C}^{H} > b_{C}^{0} \), the types \( b \in [b_{C}^{0}, b_{C}^{H}] \) harass in subgame \( H \) but not in subgame \( \emptyset \), whereas if \( b_{C}^{H} < b_{C}^{0} \), the intermediate-range types \( b \in [b_{C}^{H}, b_{C}^{0}] \) do the opposite, harassing in subgame \( \emptyset \) but not in \( H \). Whichever the case, then, for types \( b \in [\min\{b_{C}^{H}, b_{C}^{0}\}, \max\{b_{C}^{H}, b_{C}^{0}\}] \), the surplus \( u_{i}^{H} - u_{i}^{0} \) is contained between the two bounds in the surplus expressions above, \((w_{A}^{H} - w_{A}^{0})(1 - \pi q^{*})\) and \(w_{A}^{H} - w_{A}^{0}\).

\footnote{To see that \( Y \)’s surplus is within these bounds, consider the case \( b_{C}^{H} > b_{C}^{0} \), where the utility difference is \( u_{i}^{H} - u_{i}^{0} = w_{A}^{H} - w_{A}^{0}(1 - \pi q^{*}) - b \) for \( b \in [b_{C}^{0}, b_{C}^{H}] \), which is positive because \( w_{A}^{H} > w_{A}^{0} \) and decreasing in \( b \), therefore is maximal at \( b = b_{C}^{0} \). On the other hand if \( b_{C}^{H} < b_{C}^{0} \), for \( b \in [b_{C}^{H}, b_{C}^{0}] \), the surplus expression

\[
 (w_{A}^{H} - w_{A}^{0})(1 - \pi q^{*}) - b \]

is a finite magnitude under any feasible contract. If the above condition does not hold, the principal cannot modify the agent wages in the contract \( C_{A}^{*} \) because \( w_{A}^{H*} \) and \( w_{A}^{b*} \) are fixed by the target harassment probabilities through (16), whereas \( w_{A}^{L*} = 0 \) and therefore cannot be reduced any further.}
The contracts \( \{C_A^*, C_S^*\} \) are type-(ii) collusion-proof if \( \sum_{i=S,X,Y}(h_i^H - h_i^0) \geq 0 \) for all type combinations \( b \) and \( h \) who accept \( C_A^* \). We know that if \( w_{A}^{H*} \geq w_{A}^{0*} \) so that \( b_C^H \geq b_C^0 \), the net expected surplus of each agent type is positive, that is, \( u_i^H - u_i^0 \geq 0 \) for all \( i = X \) types \( h \in [0, h_A] \) and all \( i = Y \) types \( b \). Combined with the fact that \( w_{S}^H - w_{S}^0 = w_{A}^{H*} - w_{A}^{0*} = m/\mu > 0 \), the contracts \( \{C_A^*, C_S^*\} \) are type-(ii) collusion-proof if \( w_{A}^{H*} \geq w_{A}^{0*} \). By Proposition 2 this corresponds to anti-harassment targets \( p_H \leq p_0 \).

On the other hand, if \( w_{A}^{H*} < w_{A}^{0*} \), i.e., if \( b_C^H < b_C^0 \), we have \( u_i^H - u_i^0 < 0 \) for all agent types. The contract is then collusion-proof only if the surplus of \( S \) from reporting \( \emptyset \) is sufficiently large to exceed the maximum value of \( \sum_{i=X,Y}(u_i^0 - u_i^H) \) over the types \( h \leq h_A \) of \( X \) and \( b \geq 0 \) of \( Y \). The type combinations who obtain the largest net surplus from collusion are \( h_A \) for \( X \) and \( b \in [0, b_C^H) \) for \( Y \). Thus, collusion cannot happen if

\[
\text{(25) } w_{S}^{H*} - w_{S}^{0*} = m/\mu \geq 2(w_{A}^{0*} - w_{A}^{H*}) + (p_H - p_0)[L + h_A(1 - \pi(1 - \beta))].
\]

By Proposition 2, \( p_H > p_0 \Leftrightarrow w_{A}^{H*} < w_{A}^{0*} \). The right hand side of (25) is positive for \( w_{A}^{H*} < w_{A}^{0*} \) and corresponding targets \( p_H > p_0 \), increasing in \( p_H \) because by Proposition 2 this is possible only if the wage \( w_{A}^{H*} \) is decreased with an upward adjustment in \( w_{A}^{0*} \). For a fixed \( p_0 \), the critical \( p_H \) defined in (19) is obtained through (25) holding with equality:

\[
p_H^C = p_0 + \frac{m/\mu - 2(w_{A}^{0*} - w_{A}^{H*})}{L + h_A(1 - \pi(1 - \beta))}.
\]

Given \( p_0 \), this is the maximum \( p_H \) that can be induced in subgame \( H \) without producing a surplus for type-(ii) collusion under the contracts \( \{C_A^*, C_S^*\} \). To induce a \( p_H \) above \( p_H^C \), the contracts must be modified. It is the supervisor’s contract and not the agent contract that should be modified because a reduction in \( w_{A}^{H*} \) and, to keep \( p_0 \) constant, a rise in \( w_{A}^{0*} \), violates (25).

For \( p_H \geq p_H^C \), define a supervisor wage for reporting \( H \) through (25) holding with equality,

\[
w_{S}^{HC} = 2(w_{A}^{0*} - w_{A}^{H*}) + (p_H - p_0)[L + h_A(1 - \pi(1 - \beta))].
\]

We have \( w_{Y}^H - w_{Y}^0 = w_{A}^{H*}(1 - \pi q^*) - w_{A}^{0} + b \), which is negative because \( w_{A}^{H*} < w_{A}^{0} \) and increasing in \( b \), therefore is maximal at \( b = b_C^0 \) as well. Using \( b_C^0 = \pi q^* w_A^0 \) in these expressions yields as maximal net surplus for \( b \) in this intermediate interval: \( (w_{A}^{H} - w_{A}^{0}) > 0 \) in the case \( b_C^H > b_C^0 \) and \( (w_{A}^{H} - w_{A}^{0})(1 - \pi q^*) < 0 \) in the case \( b_C^H < b_C^0 \).
By definition, $w_{HC}^S = m/\mu$ if $p_H = p_H^C$. The larger is the target $p_H$ relative to $p_H^C$, the higher is $w_{HC}^S$, the rent that should be left to the supervisor to prevent collusion in inspection outcome $H$.

**Proof of Proposition 5**: Consider a change $dZ$ in $Z = \pi, L, \beta$ for fixed pair of targets $(p_H, p_\emptyset)$, hence, fixed $b_H^k$ and $b_\emptyset^k$. Differentiating the equilibrium condition $p_k = 1 - F(b_k^C)$ where $b_k^C = \pi w_A^k q^*$ yields

$$\frac{\partial b_k^C}{\partial w_A^k} dw_A^k + \frac{\partial b_k^C}{\partial w_A^{-k}} dw_A^{-k} = -\frac{\partial b_k^C}{\partial Z} dZ$$

for $k = \emptyset, H$.

Note that, from (16),

$$\frac{\partial b_k^C}{\partial w_A^k} > 0; \quad \frac{\partial b_k^C}{\partial w_A^{-k}} > 0; \quad \frac{\partial b_k^C}{\partial \pi} > 0; \quad \frac{\partial b_k^C}{\partial L} < 0; \quad \frac{\partial b_k^C}{\partial \beta} < 0.$$

Using the signs of the partial derivatives above yields $dw_\emptyset^k/dZ < 0$, $dw_H^k/dZ < 0$, for $Z = \pi$, and $dw_\emptyset^k/dZ > 0$, $dw_H^k/dZ > 0$ for $Z = L, \beta$.

Consider now the incentive compatibility constraint (12), that is, $w_H^* \geq e/\mu(1 - \pi q^*)$. An increase in $\pi$ raises the right hand side of the inequality directly, and indirectly by leading to an increase in $q^*$, whereas it leads to a fall in the left hand side, $w_H^*$, for constant targets $(p_H, p_\emptyset)$. Therefore, following the rise in $\pi$ the critical target $p_H(p_\emptyset)$ decreases, implying a smaller range $[0, p_H(p_\emptyset)]$ of targets in subgame $H$ compatible with effort incentives, for any fixed $p_\emptyset$.

The impact of $L$ and $\beta$ on (12) is qualitatively similar to that of $\pi$, operating through the induced change in $q^*$ and $w_H^*$.

**Q.E.D.**

**References**


