Monitoring Harassment in Organizations

Laura Boudreau
Columbia University

Sylvain Chassang
Princeton University

Ada González-Torres
Ben Gurion University

Rachel Heath*,†,‡
University of Washington

August 23, 2023

Abstract

We evaluate secure survey methods designed for the ongoing monitoring of harassment in organizations. To do so, we partner with a large Bangladeshi garment manufacturer and experiment with different designs of phone-based worker surveys. "Hard" garbling (HG) responses to sensitive questions, i.e., *automatically* recording a random subset as complaints, increases reporting of physical harassment by 290%, sexual harassment by 271%, and threatening behavior by 45%, from reporting rates of 1.5%, 1.8%, and 9.9%, respectively, under the status quo of direct elicitation. Rapport-building and removing team identifiers from responses do not significantly increase reporting. We show that garbled reports can be used to consistently estimate policy-relevant statistics of harassment, including: How prevalent is it? What share of managers is responsible for the misbehavior? and, How isolated are its victims? In our data, harassment is widespread, the problem is not restricted to a minority of managers, and victims are often isolated within teams.

KEYWORDS: Harassment, whistleblowing, garbling, secure survey design, gender, garments, Bangladesh

*This project is funded by the Private Enterprise Development in Low-income Countries (PEDL) Initiative, by Columbia University's Provost's Diversity Grants Program for Junior Faculty and by the Israeli Science Foundation. We are grateful to Ferdausi Sumana, Raied Arman, and Krishna Kamepalli for their excellent research assistance.

†We are grateful to Nava Ashraf, Jana Gallen, Rocco Macchiavello, and Benjamin Roth for detailed discussions. We are indebted to Dan Ben-Moshe, Laura Doval, Florian Englmaier, Nathaniel Hendren, Emily Nix, Danielle Li, Tomasso Porzio, and Andrea Prat for helpful comments. We are grateful to seminar participants at Bar-Ilan, Berkeley, Chicago-Booth, Columbia, CUNEF, Haifa, Hebrew U, LMU Munich, Reichman, Rochester, UBC, UCSD, USC, UW, as well as participants at the Barcelona Summer Forum, the BREAD & MIT Conference, the CEPR/CESifo/Imo-ENT Conference, the Economics of Firms and Labor Conference, the German-Israeli Frontiers of Humanities Symposium, the Organising Development & the Development of Organisations Conference, the NBER Organizational Economics Meeting, the NBER Personnel SI, the SIOE Conference, the Stockholm Workshop on Diversity & Workplace Inclusion, and the Women in Applied Microeconomics Conference for stimulating comments.

‡This research was pre-registered in the AER registry (AEARCTR-0007103), and received IRB approval from Ben-Gurion University (2001-1; 2021-02-17), Columbia University (IRB-AAAT2938; 2021-02-23), and University of Washington (STUDY00010219; 2020-12-22).

1 Introduction

Organizations' ability to take action against harassment is limited by their ability to elicit information from relevant parties. Reporting harassment is a difficult step for individuals who have been victimized and for witnesses concerned with possible retaliation and reputational costs. This prevents organizations from responding to individual issues, but also from assessing the scope and nature of their harassment problem. Using a phone-based survey experiment implemented at a large Bangladeshi apparel manufacturer, we study the impact of survey methods that seek to offer plausible deniability, increase trust in the survey enumerator, and reduce the perceived likelihood of leaks, on information transmission. We also show how such reporting data can be used to answer policy-relevant questions about the nature harassment, including: How widespread is it? What share of managers are responsible for what share of the damage? How isolated are victims?

Our theoretical framework builds on a principal-agent-monitor model (Chassang and Padró i Miquel, 2018, Chassang and Zehnder, 2019). A monitor, here the victim, is asked to report harassment behavior by the agent to the principal. The difficulty is that the agent can engage in retaliation, and victims may be concerned that reports could be leaked. Leakages may be the result of legitimate steps taken by the principal to investigate or to address the issue, as well as malicious or erroneous revelation by either the principal or survey collectors. The theoretical framework predicts that steps that increase plausible deniability, i.e., that make it harder to infer a respondent's intended message, as well as steps that increase trust in the enumerator and reduce the perceived likelihood of leaks, can increase reporting by reducing the perceived risk of retaliation.

This motivates three concrete treatments. First, hard garbling (HG) recorded information by automatically setting a random subset of reports as reports that harassment took place, which provides respondents with plausible deniability in the event that they file an incriminating report (Warner, 1965, Chassang and Padró i Miquel, 2018, Chassang and Zehnder, 2019). Second, rapport building (RB) by the survey enumerator, i.e., chatting about family and hobbies in a natural but pre-specified manner beyond the minimum small talk typical in a social science survey, which may increase the respondents' trust in the enumerator, as well as their trust in the surveyor's commitment to follow th protocol. Third, reducing the amount of personally identifying information collected in the survey (Low PII), including the name of workers' direct supervisor and their production team, which may alleviate the

concern that leaked data could be traced back to the respondent.

In all three approaches, the possible benefit of increased willingness to report comes at a cost: HG provides a noisy signal of misbehavior, which constrains the severity of organizational responses to reports; RB requires careful planning of the RB process, additional training of survey enumerators, and more time to conduct the survey; removing team-level information precludes computation of manager-level statistics that are important to characterize the nature of an organization's harassment problem.

We highlight a number of policy-relevant statistics of harassment that a decision-maker may wish to estimate, including: How prevalent is harassment? What share of managers is responsible for the bulk of the misbehavior?[1] How isolated are victims? How do harassment rates compare for men and women? The answers to these questions are crucial inputs to determining the policies that can be used to address harassment. For example, if a small share of managers is responsible for the harassment, the organization could investigate and fire them. In contrast, if most managers are involved, firing them all is likely impossible, and other remedial actions need to be taken. We show these statistics can be consistently estimated from garbled reports, and illustrate their use with our survey data.

We conducted phone-based surveys with 2,245 workers at two of the apparel producer's plants and had a response rate of 63%.[2] We randomly assigned survey respondents to 9 different combinations of the treatment conditions: HG, RB, and Low PII. The status quo, or baseline treatment arm, entailed direct elicitation (DE) of respondents' experience of harassment, no RB, and elicitation of team-level PII. We examine the effects of our survey design interventions on three pre-specified outcomes: reporting of threatening behavior, physical harassment, and sexual harassment by respondents' direct supervisors.

We find that reporting rates in the survey's control group are low, especially for physical and sexual harassment: 9.9% of respondents report threatening behavior, 1.53% report physical harassment, and 1.78% report sexual harassment. HG increased reporting of threatening behavior by 45%, sexual harassment by about 271%, and physical harassment by 290%. We also find that low PII and RB had positive but weak effects. There is suggestive evidence of complementarity between treatment arms: combining hard garbling with rapport-building and low PII increases reporting compared to the sum of individual treatment effects.

[1] In the context that we study, harassment by managers perpetrated against workers is the primary concern. Section 2 provides more information on the context.

[2] Nearly all non-response was due to our inability to reach workers by phone.

We find a surprising pattern of heterogeneous treatment effects (HTEs) by respondents' sex. Compared to women, men's baseline reporting rates were higher for threatening behavior and physical harassment and lower for sexual harassment. The effects of HG were substantially larger for men compared to women for both threatening behavior and sexual harassment, although for sexual harassment, we lack power to detect the statistical differences between the effects for men and women.

Next, we use the improved reporting data to assess the scope and nature of harassment in the apparel producer's organization. Estimating statistics using garbled data requires constructing estimators that depend on respondents' *intended* reports. Warner (1965) derives a consistent estimator for the mean intended reporting rate using garbled data. We extend this result and derive consistent estimators of team-level statistics under different HG schemes, including independent and identically distributed (i.i.d.) HG and what we refer to as blocked HG. With blocked HG, the surveyor ensures that a target number of reports are set to automated "yeses," either in the overall sample or per team. Blocked HG, in particular at the team-level, substantially reduces the variances of estimators.[3]

Using data from treatment arms using HG, we estimate that in this organization, 13.5% of workers reported threatening behavior, 5.7% reported physical harassment, and 7.7% reported sexual harassment. On average, there are seven workers per production team in arms that use HG and collect PII. In this sample, we find that 72% of teams had at least one worker who had been threatened, 40% had at least one who had been sexually harassed, and 25% had at least one who had been physically harassed. These statistics indicate that harassment is widespread, and a policy of firing all misbehaving supervisors is unlikely to be feasible. Conditional on a type of harassment, victims tend to be isolated, and more so for graver types of harassment. The probability of having at least two victims on the team, conditional on having at least one, are respectively 17% for threatening behavior, 14% for physical harassment, and 8% for sexual harassment. These results shed light on the implications of setting different burdens of proof for harassment. In contexts where victims are isolated, requiring multiple victims to come forward, for example, to avoid "he said, she said" situations, will miss the majority of cases, and eradicating harassment will require having processes in place that can be initiated when only one victim comes forward.

This paper contributes to an emerging literature in economics on workplace harassment,

[3]It also affords workers with less protection in case of a data leakage, which could be an important consideration in many contexts.

in particular sexual harassment, and its implications for labor markets. Cheng and Hsiaw (2020) consider reasons for underreporting of sexual harassment; they develop a model in which harassment is underreported if there are multiple victimized individuals because of coordination problems. Dahl and Knepper (2021) provide evidence that U.S. employers use the threat of retaliatory firing to coerce workers not to report sexual harassment. Adams-Prassl et al. (2022) document that experiencing harassment leads to adverse employment outcomes for victims and perpetrators and Folke and Rickne (2022) show that sexual harassment contributes to gender inequality in the labor market. We contribute evidence that lack of plausible deniability causally negatively affects reporting of workplace harassment. Our findings suggest that estimates of labor supply and other responses to harassment may be severely biased when harassment is measured using formal complaints: it may be that reporting is most suppressed in workplaces where harassment is most problematic.

In the context of developing countries, sexual harassment is considered a key barrier to women's labor market participation (Jayachandran, 2021).[4] There is a dearth of evidence, however, on the effects of sexual harassment in the workplace on workers' labor supply and well-being.[5] Further, in light of workers' lack of access to secure internal reporting channels (Boudreau, 2022) and to recourse through criminal justice systems, as well as relatively stronger gender norms, we expect underreporting to be even more of a concern in many developing countries. We contribute to our understanding of the prevalence and nature of harassment in a low-skill manufacturing sector that is common to many developing countries. Our evidence confirms that harassment against women by managers who are men is common, and it shows that harassment by men against subordinate men is also substantial.

This research also contributes to the literature on the detection and deterrence of collusion, corruption, and other forms of misbehavior in organizational settings. A large body of contract theory literature with principal-agent-monitor set-ups considers the possibility of bribes in collusive relationships between monitors and agents to limit information transmission to the principal (Tirole, 1986, Laffont and Martimort, 1997, 2000, Prendergast, 2000, Faure-Grimaud et al., 2003, Ortner and Chassang, 2018). More recently, a smaller strand of

[4]One stream of literature establishes that harassment is prevalent in public spaces and transit systems in cities ranging from Rio de Janeiro to Delhi and that it reduces women's educational investments and labor supply (e.g., Aguilar et al. (2021), Kondylis et al. (2020), Borker (2018)).

[5]The poor working conditions (Boudreau et al., 2022) and extreme gender imbalances between managers and workers (Macchiavello et al., 2020) documented in the literature on Bangladesh's garments sector are suggestive of possible harassment concerns.

literature considers that collusion may come in the form of punishments against informants, or whistleblowers (Heyes and Kapur, 2009, Bac, 2009, Makowsky and Wang, 2018). Chassang and Padró i Miquel (2018) develop a model in which misbehaving agents can commit to a retaliation strategy. They show that garbled intervention policies are needed to discipline their behavior. They also clarify how to experimentally evaluate such policies even in the hypothetical presence of malicious workers wrongfully reporting well-behaved managers. We contribute by bringing HG into a real-world organizational setting. The large experimental effect of HG on information transmission in our setting suggests that this class of mechanisms deserves further exploration in other environments where credible threats or reputation costs limit information transmission.

Finally, this research contributes to a literature on garbled survey designs and on inference from garbled surveys dating back to Warner (1965). Warner (1965) proposed randomized response (RR) as a way to offer survey respondents a form of plausible deniability when answering sensitive questions. Under RR, the surveyor instructs respondents to roll a dice, and answer the question truthfully or not depending on the outcome. For instance, a respondent may be instructed to submit the response "Yes" if the dice lands on 1 or 2, and to answer "Have you experienced harassment?" truthfully if the dice lands on 3-6. The surveyor does not observe the respondent's dice roll. RR admits several variants, which we discuss later in the paper. Provided that people comply with the surveyor's instructions, RR offers plausible deniability: a recorded response "Yes" may be due to the fact that the dice landed on 1 or 2. The empirical literature on survey design for sensitive questions has found that RR performs better than DE, at least in single shot, large scale surveys (Rosenfeld et al., 2016).

We argue that RR and related designs, such as list experiments (LE), are poorly suited for ongoing use in organizations. Because the randomization is entirely under the respondent's control, respondents can ignore instructions to randomly respond "Yes" if they are worried about retaliation. In equilibrium, this causes plausible deniability to unravel altogether. This concern is empirically validated by Chuang et al. (2020): survey respondents often do not comply with the protocol to garble and systematically provide the least sensitive response. Because the garbling in RR relies on respondents' compliance, we refer to mechanisms in this class as soft garbling. Instead, in our design, responses are mechanically switched at an exogenous rate, which is why we refer to it as hard garbling. Chassang and Zehnder (2019) show that in contrast to RR, HG does not unravel in equilibrium. For this reason, we believe

it is better suited for ongoing use in organizations. We make two further contributions. First, we derive consistent estimators of team-level statistics of intended responses using garbled data, extending the estimator of population-level reporting rates proposed by Warner (1965). Second, we show that using sequences of garbling errors that satisfy a small law of large numbers – i.e., blocking – considerably improves inference. This is especially important when baseline reporting rates are low so that sampling error can dwarf the statistic of interest.

The remainder of the paper is organized as follows. Section 2 provides background on Bangladesh's garments sector and the anonymous apparel producer whom we partner with. Section 3 provides a simple theoretical framework that clarifies incentives for information transmission under various designs. We explicitly discuss the pros and cons of HG vs. RR or LE and provide estimators for team-level statistics based on garbled reports. Section 4 presents the research design. Section 5 presents the results of the reporting experiment. Section 6 uses the garbled survey data to characterize the apparel producer's harassment problem. Section 7 discusses our findings and concludes.

2 Context

We conducted this research in collaboration with a large apparel producer in Bangladesh, employing upwards of 25,000 workers in roughly half a dozen factories.[6] The manufacturer's senior leadership team sought a collaboration with our research team because it wished to improve relations with its workers and to improve workers' well-being. To achieve this, it aimed to directly collect feedback from workers on their experiences in the workplace and relationships with their managers. It then aimed to use this feedback to inform its HR policies. For the purpose of the experiment, we agreed to survey workers at 2 of its plants. In the longer-term, the senior management team's goal was to set-up a reporting system for workers to provide continuous feedback in real-time.

Ethnographic evidence and evidence from community-based surveys suggests that harassment is a long-running problem in Bangladesh's garments sector (Siddiqi, 2003, Sumon et al., 2018, Kabeer et al., 2020). Workers' precarious livelihoods and lack of legal recourse, as well as conservative societal norms around gender and sex, contribute to an enabling environment for managers with power over workers to harass them (Siddiqi, 2003). While there

[6] We have a confidentiality agreement with the apparel manufacturer.

is reason to believe that harassment is widespread, measuring and constructing informative statistics of harassment is extremely challenging, even in social science research conducted outside of the workplace. For example, using data from Kabeer et al. (2020)'s community-based survey of garment workers, we find that while 20% (11%) of workers report witnessing physical (sexual) harassment, only 1% (0%) report experiencing it themselves.

The manufacturer's operations are representative of garment manufacturing in Bangladesh. Production is organized into cutting, sewing, and finishing sections; some factories also have wet and dry washing sections, which adds texture and/or fading to sewn garments (e.g., denim jeans). Within these sections, workers are organized into production teams or lines, with team assignments that are largely stable over time. The organizational structure is very hierarchical: teams of workers are typically overseen by 2 supervisors, followed by line chiefs or team incharges, floor-supervisors and/or assistant production managers, production manager(s), and finally, the managing director. Production sections vary considerably in their sex composition: cutting and wet washing sections typically exclusively employ men, sewing and finishing sections mostly employ women, and dry washing sections are often more mixed. In contrast, more than 90% of managers in all sections are men.

Within the two plants that were surveyed, 34-42% of workers are employed on sewing lines, 16-18% are employed in finishing, and 10-14% are employed in washing. The remaining workers are employed in smaller, supporting production sections. 93% of managers are men.

3 Framework

We begin by identifying a number of policy-relevant statistics of harassment that we are interested in estimating. Some statistics, such as the share of victimized workers, do not require collecting information about workers' teams (i.e., their team id). In contrast, statistics associated with team-level patterns do: for instance, assessing whether victimized workers are isolated or assessing the number of managers engaging in misbehavior.

The difficulty is that harassment is not directly observable, so that the statistics must be estimated based on reporting data. Using a principal-agent-monitor framework, we show how the gap between true statistics of harassment and their counterparts based on intended reports can be affected by different survey features, including garbling. We then show how to consistently estimate statistics of intended reports based on garbled reports alone. We

discuss how different garbling structures affect statistical power and clarify the pros and cons of using different versions of HG versus common alternatives, such as RR and LE.

3.1 Policy-relevant statistics of harassment

Consider an organization consisting of $m \in \mathbb{N}$ teams. Each team $a \in M \equiv \{1, \cdots, m\}$ consists of a manager (also denoted by a) and L workers indexed by $i \in I \equiv \{1, \cdots, L\}$. Altogether, the organization consists of $n \equiv m \times L$ workers and m managers.

We assume for simplicity that all harassment is performed by managers against workers under their span of control. For any manager a and worker i, we denote by $h_{i,a} = 1$ the event that manager a harassed worker i, and by $h_{i,a} = 0$ the event that they did not. We denote by $h_a \in \{0,1\}^L$ the profile of harassment choices made by manager a. Throughout the paper, we take as given the behavior of managers, and we seek to elicit information about patterns of harassment $(h_a)_{a \in M}$ in the organization.

We are interested in identifying four statistics that can help organizations assess policy options. We emphasize that these statistics are not directly computable because they depend on harassment patterns that are not directly observed by the decision-maker. We discuss workers' decisions to report harassment below. We are interested in computing:

$$\begin{aligned} \mathsf{S}_V &\equiv \frac{1}{n} \sum_{a,i \in M \times I} h_{i,a}, \\ \mathsf{S}_{PM} &\equiv \frac{1}{m} \sum_{a \in M} \max_{i \in I} h_{i,a}, \\ \forall k \in \{1, \cdots, L\}, \quad \mathsf{S}_{TV \geq k} &\equiv \frac{1}{m} \sum_{a \in M} \mathbf{1}_{\sum_{i \in I} h_{i,a} \geq k}, \\ \mathsf{E}_{2V|1V} &\equiv \frac{\mathsf{S}_{TV \geq 2}}{\mathsf{S}_{TV \geq 1}}. \end{aligned}$$

Statistic S_V measures the share of victimized workers. This allows decision-makers to gauge the magnitude of the harassment problem in their organization, which allows stakeholders to correctly prioritize the issue and to allocate suitable resources.

Statistic S_{PM} measures the share of problematic managers, in other words, managers who have harassed at least one person. It is a special case of statistic $\mathsf{S}_{TV \geq k}$, which measures the share of managers who have harassed at least k workers, for $k = 1$. The behavior of $\mathsf{S}_{TV \geq k}$

as k increases clarifies policy options. For example, if there exists a k large such that $\mathsf{S}_{TV\geq k}$ is small, but $k \times \mathsf{S}_{TV\geq k}$ is large, then this means that a relatively small share of managers is responsible for a large amount of the damage. This means that investigating and firing repeat offenders may be a viable policy option. If instead S_{PM} is large, but $k\mathsf{S}_{TV\geq k}$ is small for k large, then this means that many managers are involved in harassment, and it is not possible to address a significant number of cases by firing a small number of managers. Since firing many managers is likely impossible for the organization, this means that other remedial action will have to be taken, such as improved training or better monitoring.

Finally, $\mathsf{E}_{2V|1V}$ measures the likelihood that a manager has at least 2 victims given that they have at least one. This allows decision-makers to assess how isolated victims are. If $\mathsf{E}_{2V|1V}$ is small, then victims are isolated. This implies that escrow mechanisms along the lines of Ayres and Unkovic (2012), which seek to help coordinate the reports of multiple victims, are unlikely to be helpful in such cases. In addition, rules limiting investigations to cases where multiple victims come forward would lead the organization to ignore the majority of problem cases. In contrast, if $\mathsf{E}_{2V|1V}$ is close to 1, then victims are rarely isolated. This means that escrow mechanisms could be helpful, and that once someone complains, it may be possible to cross-validate reports of misbehavior, permitting more effective action.

Sensitivity of statistics. These statistics differ in the sensitivity of information required to compute them. It is not necessary to know a particular worker's team to compute S_V. In contrast, S_{PM}, $\mathsf{S}_{TV\geq k}$, and $\mathsf{E}_{2V|1V}$ all require the respondent to associate some team identifier to their report. Otherwise, it is not possible to match the reports of different workers on the same team. For this reason, these statistics are intrinsically more sensitive than S_V: surveys needed to compute these sensitive statistics will need to include both team ids and harassment reports. We will return to this consideration in our discussion of workers' decision to report harassment and the design of our survey experiment.

Third-party witnesses. In principle, harassment may be observed by workers other than the victim, and decision makers may be interested in statistics of harassment calculated using information furnished by witnesses. In this paper, our focus is on reporting of one's own harassment status. We leave the question of witnesses' role in detecting and counter-acting harassment to future research.[7]

[7]We note that witnesses are exposed to the same retaliation risk as victims, and may derive lower personal benefits than victims from informing about a problem manager who has not harassed them directly.

3.2 A worker's reporting decision

Because the actual harassment status $h_{i,a}$ of workers is typically not observed, organizations must proxy the true statistics of interest with reported harassment. One difficulty is that victims are often unwilling to come forward. This may be due to explicit or implicit threats of retaliation, concerns over one's own reputation, or negative impacts on one's career and private life, even if the organization takes action against the perpetrator.

We consider a set-up in which worker i in team a completes a binary survey, meaning that they can submit an intended response $r_{i,a} \in \{0, 1\}$. In our setting, rates of reported harassment are low, and the implicit stigma associated with reports of harassment (especially of a sexual nature) is high. For this reason, we assume there are no false positives: $r_{i,a} \in \{0, h_{i,a}\}$. We discuss the possibility of false positives, as well as equilibrium responses to garbling by managers, in Section 7. We argue using the framework of Chassang and Padró i Miquel (2018) that getting people to complain is a necessary first step, even if false positives become an issue. Following Chassang and Padró i Miquel (2018) and Chassang and Zehnder (2019), we consider garbled survey methods that add noise to the report sent by a worker. An intended report $r \in \{0, 1\}$ is associated with a potentially random recorded report $\widetilde{r}$ distributed according to $\phi(r) \in \Delta(\{0, 1\})$.

Concretely, we are interested in the following survey designs:

- *Direct Elicitation*, in which $\phi(r) = r$: the recorded report matches the intended report.
- *Hard Garbling*, in which $\phi(1) = 1$, but

$$\phi(0) = \begin{cases} 0 & \text{with probability } 1 - \pi \\ 1 & \text{with probability } \pi \end{cases}$$

  where $\pi \in (0, 1)$. In words, reports of harassment are always recorded, but reports of no harassment are switched to reports of harassment with an interior probability π.

For the remainder of this section, unless otherwise noted, we refer to hard garbling as "garbling." The rationale for garbling surveys is to guarantee the worker plausible deniability in the event that their record is leaked. In particular, we assume that the worker assigns subjective probability $p \in [0, 1]$ on their recorded report $\widetilde{r}_{i,a}$ being leaked. We do not take a stance on whether leaks actually occur or not. In our experimental application, leaks of

individual reports exist only in the mind of respondents. However, we are interested in the use of reporting systems for ongoing monitoring in organizations. In such a context, leaks may simply correspond to the fact that some action is taken by the organization on the basis of the recorded report.[8] Leaks are inevitable even under ideal governance.

Worker i's utility U_i associated with an intended report r depends on their true harassment status and consists of direct benefits from reporting as well as potential reputational and/or retaliation costs:

$$U_i(r|h_{i,a}) = \mathsf{PB}(r|h_{i,a}) + \mathsf{SB}(\widetilde{r}|h_{i,a}) + p \times \mathsf{RC}(\widetilde{r})$$

where:

PB is a psychological benefit from taking action such that $\mathsf{PB}(1|1) > 0$ and for simplicity $\mathsf{PB}(1|0) = \mathsf{PB}(0|1) = \mathsf{PB}(0|0) = 0$. Respondents only derive psychological benefits from taking action against a misbehaving manager.

SB is a social benefit from realized report $\widetilde{r}$ as perceived by the worker, either because it triggers an investigation, or because it helps the organization design better policies. For simplicity, we assume that $\mathsf{SB}(1|1) > 0$, $\mathsf{SB}(1|0) < 0$ and $\mathsf{SB}(0|1) = \mathsf{SB}(0|0) = 0$.[9] Recorded complaints only yield social benefits if they are associated with a misbehaving manager.

$\mathsf{RC}(\widetilde{r})$ is a reputational and/or retaliation cost in case recorded report $\widetilde{r}$ is leaked. We assume it takes the form $\mathsf{RC}(\widetilde{r}_{i,a}) = -\mathsf{K}(\text{prob}(r_{i,a} = 1|\widetilde{r}_{i,a}))$ where: K is a positive strictly increasing function; $\text{prob}(r_{i,a} = 1|\widetilde{r}_{i,a})$ is the posterior belief that worker i intended to submit a complaint $r_{i,a} = 1$ about manager a, conditional on recorded report $\widetilde{r}_{i,a} = 1$.[10]

In equilibrium, a non-harassed worker always finds it optimal to submit intended report $r_{i,a} = 0$. Given harassment status $h_{i,a} = 0$, the expected payoffs from sending reports $r_{i,a} = 1$ and $r_{i,a} = 0$ are

$$U_i(1|0) = \mathsf{SB}(1|0) - p \times \mathsf{K}(\text{prob}(\mathrm{r_{i,a}} = 1|\widetilde{\mathrm{r}}_{\mathrm{i,a}} = 1)) < 0$$
$$U_i(0|0) = \pi \times (\mathsf{SB}(1|0) - p \times \mathsf{K}(\text{prob}(\mathrm{r_{i,a}} = 1|\widetilde{\mathrm{r}}_{\mathrm{i,a}} = 1))) = \pi \times U_i(1|0).$$

[8] For instance, the manager is sent to a training seminar.

[9] The assumption that $\mathsf{SB}(1|0) < 0$ implies that arbitrarily high garbling rates π are not a priori desirable.

[10] This functional form captures concerns over ex post retaliation by managers and related career concerns.

In turn, a harassed worker's payoffs are

$$U_i(1|1) = \mathsf{PB}(1|1) + \mathsf{SB}(1|1) - p \times K(\text{prob}(\text{r}_{\text{i,a}} = 1|\widetilde{\text{r}}_{\text{i,a}} = 1))$$
$$U_i(0|1) = \pi \times [\mathsf{SB}(1|1) - p \times K(\text{prob}(\text{r}_{\text{i,a}} = 1|\widetilde{\text{r}}_{\text{i,a}} = 1))] .$$

Hence, a harassed worker is willing to send intended report $r = 1$ if and only if

$$\mathsf{PB}(1|1) + (1-\pi)[\mathsf{SB}(1|1) - p \times K(\text{prob}(\text{r}_{\text{i,a}} = 1|\widetilde{\text{r}}_{\text{i,a}} = 1))] \geq 0. \tag{1}$$

where the posterior belief $\text{prob}(r_{i,a} = 1|\widetilde{r}_{i,a} = 1)$ takes the form

$$\text{prob}(r_{i,a} = 1|\widetilde{r}_{i,a} = 1) = \frac{1}{1 + \pi \frac{\text{prob}(r_{i,a}=0)}{1-\text{prob}(r_{i,a}=0)}}. \tag{2}$$

Taking as given the share of null reports $\text{prob}(r_{i,a} = 0)$, it follows from (1) and (2) that increasing π increases a worker's propensity to send report $r_{i,a} = 1$. Equation (2) captures the fact that increasing π reduces the impact of a positive recorded report $\widetilde{r}_{i,a}$ in terms of the reputational and/or retaliation cost. In addition, coefficient $(1 - \pi)$ in (1) captures the fact that increasing π shrinks the reputational and/or retaliation cost savings associated with sending a null report $r_{i,a} = 0$. As a result, the left-hand side of (1) exhibits single crossing in the garbling rate π: whenever its value is negative, it is increasing in π.

Proposition 1 (the value of survey design)**.** *Taking as given the behavior of managers,*

(i) intended reports underreport true harassment: $r_{i,a} \leq h_{i,a}$*;*

(ii) equilibrium reporting weakly increases with garbling rate π*;*

(iii) equilibrium reporting weakly decreases with perceived leakage probability p*.*

A corollary of Proposition 1 is that both garbling and reducing the perceived leakage probability increase the accuracy of intended reports.[11] We note that the impact of garbling π on reporting is an equilibrium effect. For HG to be effective in the long run, respondents

[11] In principle, one may worry that HG could reduce the welfare of individuals who would prefer to report that they have not been harassed, but under HG, may be associated with a positive recorded report. In our setting, the evidence suggests that HG does not reduce the welfare of workers who do not submit a positive intended report. First, we do *not* find that individuals assigned to HG are less likely to respond compared to those assigned to DE. Individuals were not aware of their treatment assignment at the recruitment stage, so they could not select in or out of the survey based on their assignment. During the survey, all individuals

as well as managers must understand that reported complaints are not necessarily associated with intended complaints.

Let S^r_V, S^r_{PM}, and $\mathsf{S}^r_{TV\geq k}$ denote analogues of S_V, S_{PM}, and $\mathsf{S}_{TV\geq k}$ computed using intended reports $r_{i,a}$ instead of actual harassment status $h_{i,a}$.

Corollary 1. *Measurement errors* $|\mathsf{S}_V - \mathsf{S}^r_V|$, $|\mathsf{S}_{PM} - \mathsf{S}^r_{PM}|$, *and* $|\mathsf{S}_{TV\geq k} - \mathsf{S}^r_{TV\geq k}|$ *are decreasing in garbling rate* π *and increasing in the perceived leakage probability* p.

The value of survey design. Proposition 1 suggests two approaches to encourage reporting through survey design. First, increase the garbling rate π. Second, reduce the worker's subjective probability p of a leak. In our survey design, we aim to do this in two ways. First, we vary whether team identifiers (needed to compute team statistics) are elicited or not. Second, we vary the extent of rapport built with enumerators prior to the sensitive module. Removing team identifiers reduces the likelihood that a leaked report may be linked to a specific worker, thereby reducing the worker's perceived expected reputational or retaliation cost. Similarly, building rapport may increase the worker's trust that survey enumerators are trustworthy, and in particular, unlikely to leak any information. If rapport affects workers through trust, there may be complementarities between rapport and HG: HG is only effective if workers trust that it is implemented as described; increasing trust may therefore increase the impact of the HG treatment.

3.3 Measurement

Corollary 1 argues that garbling reduces bias in the measurement of policy-relevant statistics based on intended reports. However, intended reports are not directly available to the analyst when garbling is used. We now discuss how to infer S^r_V, S^r_{PM}, and $\mathsf{S}^r_{TV\geq k}$ from garbled reporting data under different garbling schemes, including i.i.d. and blocked garbling. For simplicity, we state identification results under the assumption that team size is constant.

in both the DE and HG arms completed the survey, so assignment to HG did not reduce willingness to respond to sensitive questions (during the consent process, individuals were told that they could stop the survey at any time without punishment). Second, as we show in the Supplementary Materials, assignment to HG did not affect attrition from a follow-up survey conducted two weeks after our main survey. Finally, in the follow-up survey, we measured workers' mental health and job satisfaction, and we find weakly positive reduced form effects of assignment to HG (not statistically significant). Given the relatively small share of workers induced to submit positive intended reports by the HG mechanism, if assignment to HG reduced workers' welfare, we would expect these reduced form effects to be negative (Table B.1).

In practice, team size varies in our application, and we discuss below how we address this in our estimation.

Inference from garbled reports. A key insight of Warner (1965) is that the share of workers reporting harassment S_V^r can be consistently estimated from garbled data, even though it depends on intended reports. The following estimator is consistent

$$\mathsf{S}_V^{\widetilde{r}} \equiv \frac{\frac{1}{n}\sum_{a,i\in M\times I}\widetilde{r}_{i,a} - \pi}{1-\pi}. \tag{3}$$

It turns out that the same is true for other statistics of intended reports, such as $\mathsf{S}_{TV\geq k}^{\widetilde{r}}$, but the precision of estimators depends on the specific garbling scheme used, and some trade-offs may have to be made.

Because team-members are anonymous, for any team a, the number or intended and recorded reports $r_a \equiv \sum_{i\in I} r_{i,a}$ and $\widetilde{r}_a \equiv \sum_{i\in I}\widetilde{r}_{i,a}$ are sufficient statistics for the profile of intended and recorded reports in team a. Let $\mu \in \Delta(\{0,1,\cdots,L\})$ and $\widetilde{\mu} \in \Delta(\{0,1,\cdots,L\})$ respectively denote the sample distribution of number of team-level intended and recorded reports, $(r_a)_{a\in M}$ and $(\widetilde{r}_a)_{a\in M}$ across teams:

$$\forall\, r \in \{0,1\cdots,L\}, \quad \mu(r) \equiv \frac{1}{m}\sum_{a\in M}\mathbf{1}_{r_a=r} \quad \text{and} \quad \widetilde{\mu}(r) \equiv \frac{1}{m}\sum_{a\in M}\mathbf{1}_{\widetilde{r}_a=r}.$$

We are interested in recovering μ from $\widetilde{\mu}$. Let us express garbled reports as

$$\widetilde{r}_{i,a} = r_{i,a} + (1-r_{i,a})\eta_{i,a} \tag{4}$$

where $\eta_{i,a} \in \{0,1\}$ is a Bernoulli random variable equal to 1 with probability π. As we discuss below, the correlation structure across shocks $\eta_{i,a}$ will turn out to matter for power. We distinguish three cases:

- i.i.d. garbling, in which $(\eta_{i,a})_{i\in I,a\in M}$ are i.i.d. across i,a.
- population-blocked garbling, in which errors $(\eta_{i,a})_{i\in I,a\in M}$ are exchangeable across workers in the population and $\sum_{i\in I,a\in M}\eta_{i,a} = \pi n$. By exchangeable across workers in the population, we mean that the distribution of $(\eta_{i,a})_{i\in I,a\in M}$ is unchanged by permutations of labels (i,a), and a fraction of messages exactly equal to π is automatically switched to a complaint.

- team-blocked garbling, in which, for every team a, errors $(\eta_{i,a})_{i\in I}$ are exchangeable across workers in team a, and satisfy $\sum_{i\in I}\eta_{i,a} = \pi L$.[12]

Proposition 2 (identification)**.** *Without additional assumptions for both i.i.d. and population-blocked garbling, and under the assumption that $\widetilde{\mu}(L) = 0$ under team-blocked garbling, there exists a fixed $(L+1, L+1)$ matrix M such that $M\widetilde{\mu}$ is an unbiased estimator of the distribution of intended reports μ that is consistent as the number of teams m grows large.*

Note that the appropriate matrix M depends on the garbling scheme. An explicit construction is provided in Appendix 7.2. This generalization of Warner (1965) allows the computation of consistent estimates of statistics S^r_V, S^r_{PM}, and $\mathsf{S}^r_{TV\geq k}$, all of which are functions of distribution μ.

Because our estimators are unbiased for a given team-size, averaging out estimators across teams yields a more precise estimate of team-population statistics. Standard errors can be computed using bootstrap samples. We follow this approach in Section 6.

Note that identification does not always hold under team-blocked garbling. The reason for this is that μ admits L degrees of freedom, while $\widetilde{\mu}$ admits only $L-\pi L$ degrees of freedom under team-blocked garbling: mechanically, $\widetilde{\mu}(0) = \cdots = \widetilde{\mu}(\pi L - 1) = 0$. In words, team-blocked garbling forces each team to have a minimum of πL recorded reports of complaints, so there is zero probability of observing team profiles with $\pi L - 1$ or fewer recorded reports of complaints. Hence, πL additional restrictions are needed. The assumption that $\widetilde{\mu}(L) = 0$ provides such a restriction. It implies that $\mu(L) = \mu(L-1) = \cdots = \mu(L - \pi L) = 0$. This assumption is true in our sample of recorded responses.

Blocked garbling improves precision. The reason team-blocked garbling is of interest, although it requires additional assumptions for identification, is that it can considerably increase precision. This is especially useful when underlying reporting rates are low. The intuition for this is made especially clear by comparing the precision of the estimator $\mathsf{S}^{\widetilde{r}}_V$, defined in (3), under i.i.d. and either population- or team-blocked garbling.

For concision, we index workers by $j \in \{1, \cdots, n\}$ rather than $a, i \in M \times I$. The sum of

[12] In settings with varying team size, or if πL is not an integer, the number of garbled reports by team may vary. In that case, inference with blocked garbling is equivalent to inference with known team-level numbers of reports assigned to be garbled. This is the informational setting under which we perform inference in Section 6. Note that blocking is performed ex ante, independently of participants' intended response.

garbled reports can be expressed as

$$\sum_{j=1}^{n} \widetilde{r}_j = \sum_{j=1}^{n} r_j + \underbrace{\sum_{j=1}^{n} \eta_j}_{A} - \underbrace{\sum_{j=1}^{n} r_j \eta_j}_{B}.$$

Take as given a vector of intended reports, $\mathbf{r}$, and denote by $\overline{r}$ its sample mean. When garbling terms η_i are i.i.d. across workers, then the variance of the sum of garbled reports is

$$\mathsf{Var}\left(\sum_{j=1}^{n} \widetilde{r}_j \middle| \mathbf{r}\right) = (1-\overline{r})\pi(1-\pi)n.$$

When the average reporting rate $\overline{r}$ is small, most of the variance is due to sampling error in aggregate garbling term A (its variance is $\pi(1-\pi)n$).

For this reason, whenever the mean reporting rate $\overline{r}$ is small, blocked garbling lowers the variance of $\sum_{j=1}^{n} \widetilde{r}_j$: term A is a constant so that the only remaining uncertainty is assigned to term B. For instance, under population-blocked garbling $\mathsf{Cov}(\eta_j, \eta_{j'}) = -\frac{\pi(1-\pi)}{n-1}$ and

$$\mathsf{Var}\left(\sum_{j=1}^{n} \widetilde{r}_j \middle| \mathbf{r}\right) = \mathsf{Var}\left(\sum_{j=1}^{n} r_j \eta_j \middle| \mathbf{r}\right) = \overline{r}\left(1 - \frac{\overline{r}n - 1}{n-1}\right)\pi(1-\pi)n.^{13} \tag{5}$$

Whenever mean reporting rate $\overline{r}$ is small, as is the case in our setting, this induces a significant reduction in the variance of the estimator $\mathsf{S}^{\widetilde{r}}_V$.

While population-blocking is enough to improve the estimation of the mean number of victims per team, it does not improve the estimation of team statistics. This is why blocking at the team level is valuable. We illustrate this point using simulations in the paper's Supplementary Materials. In our application, we attempted to implement team-blocked garbling by ordering respondents assigned to HG by production team and gender and ensuring that 2 out of every 10 consecutive reports were recorded as complaints (i.e. $\pi = 2/10$).[14] Because team sizes are not multiples of 10, we do not achieve exact team-blocking. This leads us to track the number of reports pre-assigned to be garbled at the team-level and to perform inference given this data.

We note that in settings where retaliation and leakages (through legitimate action, or

[13] See Appendix 7.2 for a proof of (5).

[14] See the Supplementary Materials for details on HG implementation in our application.

malicious channels) are an especially high concern, then i.i.d. garbling may be preferred to blocked (or known team-level) garbling. In our context, where leakages are a subjective concern, and where we are especially interested in learning about team level statistics, the benefits of blocked (or known team level) garbling outweighed its costs.

3.4 Alternative indirect survey response methods

Starting with the pioneering work of Warner (1965) on randomized response (RR), many indirect response methods have been developed to guarantee survey respondents plausible deniability, including the unrelated question approach (Greenberg et al., 1969), list experiments (LEs, Raghavarao and Federer (1979)), and most recently, crosswise RR methods (Blair et al., 2015). We use HG instead of these alternatives for several reasons.

A common feature of these approaches, including RR and LE, is that the respondent controls the report being sent, and plausible deniability occurs only if respondents comply with the instructions. For this reason, we refer to these approaches as soft garbling. As Tourangeau and Yan (2007) and Chuang et al. (2020) highlight, a difficulty with soft garbling is that respondents frequently do not comply with instructions, and they selectively deviate when instructed to submit a more sensitive answer. Chuang et al. (2020) propose tests of non-compliance and show that non-compliance is large and problematic in both RR and LEs.

This implies that it is not possible to use standard inference formulas to recover intended response rates. Because the number of randomly induced sensitive responses is not known, one cannot normalize the number of recorded sensitive responses to obtain estimators of intended response rates as in (3). Further, as Chassang and Zehnder (2019) highlight in a laboratory setting, over time, noncompliance means that the plausible deniability associated with these methods unravels. This is very important in organizational settings, since participants repeatedly interact with the survey method and learn to play in equilibrium. For instance, under LEs, if providing a higher number incriminates one's manager, then employees may be told to systematically agree with the smallest plausible number of statements.

HG addresses both concerns. The survey tool performs the garbling, so the nature of the noise is known, permitting inference. Second, plausible deniability does not unravel even if all agents submit non-sensitive reports: some non-sensitive reports are mechanically switched to sensitive reports. For this reason, and especially in view of the evidence provided by Chuang et al. (2020), we think that HG is better suited for steady state monitoring in

organizations. HG also allows for blocked designs that deliver more precise estimates than i.i.d. garbling, which is the only option under RR. This is especially valuable when baseline reporting rates are low and sampling error can dwarf the statistic of interest.

This is not to say that HG does not have drawbacks. Because the survey tool performs the garbling, respondents need to trust that the survey organization follows the protocol it announces. This is feasible in organizational settings where longer-term relationships allow third-parties to build a reputation with respondents (as is the case in our application), but it may be more difficult in one shot, large scale surveys where the survey organization does not have high trust within the respondent population. In such settings, although compliance is an issue, RR may be preferred since it allows respondents to control the noise.

4 Experiment Design

We surveyed workers at 2 of the apparel producer's plants. Prior to the survey's launch, the plants' HR departments announced on the PA system that workers may be invited to participate in a survey the firm was running in collaboration with independent researchers. The BRAC Institute for Governance and Development (BIGD), a respected arm of BRAC University in Bangladesh, conducted all data collection. We prepared a pre-analysis plan (PAP) for the experiment and registered it on the AEA's RCT registry. We overwhelmingly adhere to our PAP and acknowledge any deviations in the text. The survey process entailed 3 phone calls conducted outside of work hours. The first call introduced the survey, established a baseline level of trust, and recruited the prospective respondent. The second completed the main survey. The third, 2 weeks later, conducted a follow-up survey. During the first call, workers who consented to participate were asked to suggest a time for the survey when they could find a private place where they felt comfortable talking about difficult workplace issues. We informed participants that aggregated results would be shared with senior management and would inform HR policy. All survey enumerators for the study were women.[15]

4.1 Harassment Outcomes

The research team was interested in measuring workers' experience of three types of harassment: threatening behavior, physical harassment, and sexual harassment. For each type of

[15]Budget constraints prohibited random assignment of the survey enumerator's sex after stratifying respondents by their sex. Based on contextual knowledge and guidance from local staff, we expected that it would be more acceptable for enumerators who are women to survey men than the reverse.

harassment, we asked, "In the past year, has your line supervisor taken any of the following actions toward you against your will?" We then listed, for each type of harassment, the actions in the second column of Table 1. Respondents were instructed to answer "Yes" if they had experienced *any* of the actions, without revealing which of the specific actions they had experienced. Ex ante, we hypothesized that threatening behavior would be the least sensitive to report and that sexual harassment would be the most sensitive to report.

Table 1: Harassment definitions

Type of harassment	Examples of harassment actions read aloud to respondent
Threatening behavior	Threatened you; Told you that they will harm you if you do not agree to or fulfill their demands.
Physical	Hit, slapped, or punched you; Cut or stabbed you; Tripped you; Otherwise intentionally caused you physical harm.
Sexual	Made remarks about you in a sexual manner; Asked you to enter into a love or sexual relationship; Asked or forced you to perform sexual favors; Asked or forced you to meet outside of the factory or meet them alone in a way that made you feel uncomfortable; Touched you in a sexual manner or in a way that made you feel uncomfortable or scared; Shown you pictures of sexual activities.

Notes: For each type of harassment, respondents were asked, "In the past year, has your line supervisor taken any of the following actions toward you against your will?" Respondents were instructed to respond "Yes" if they experienced *any* of the actions against their will.

4.2 Treatment Conditions

We randomly assigned survey participants to different combinations of treatment conditions. We varied whether the survey method garbled respondents' intended reports. We varied the extent to which the survey enumerator built rapport with the surveyed individual. Finally, we varied the level of identifiability of a workers' team and manager. As discussed in Section 3.4, the latter two conditions aim to reduce the worker's subjective probability of a leak p. More specifically, the status quo and alternative treatment conditions were as follows.

Survey method for harassment-related questions:

- Direct elicitation (DE): directly ask the survey respondent about sensitive information.
- Hard garbling (HG): for a yes or no question, where *yes* is the more sensitive answer, exogenously flip *no* answers to *yes* with probability $\pi = 2/10$.

DE is the status quo survey method and the control condition. HG is the treatment condition: it provides respondents with plausible deniability if they submit a sensitive answer. For HG, we set the flipping rate to 20% and use blocking to ensure that 2 out of every 10 consecutive reports, after ordering respondents by production team and gender, are garbled.[16] We decided upon a 20% flipping rate based on several contextual considerations, including reporting rates from prior surveys that we and other researchers had run, focus groups and piloting that suggested that this rate was perceived as providing material protection, and the relative ease of explaining a 1-in-5 flipping rate. The paper's Supplementary Materials provide implementation details. We explained HG as follows.

> "We are now going to ask you several questions about the way your manager treats you and other employees. For instance: 'Has your manager shouted at you in the last month? Yes or No?'
>
> Each of the questions has a Yes or No answer.
>
> Our system is set up so that it's safe to report an issue.
>
> If you choose to respond YES (there is an issue), our system will record it as a YES for sure. Importantly, if someone responds NO, the system will sometimes record the response as YES.
>
> This means that if you respond YES, we can guarantee that you won't be the only one saying YES. For every 5 responses from workers, at least 1 will be recorded as YES.
>
> The researchers are only interested in the total number of yes/no responses from all surveys. If you respond YES, aside from me, no one will ever be able to know that this was your answer, not even the researchers. Your answers are fully protected with us."

We report the main effects of HG and our other treatments in Section 5. It is legitimate to ask whether any potential effect of HG on responses is actually due to respondents understanding the value of plausible deniability. An alternative hypothesis is that respondents

[16] We implemented the blocked garbling ex ante, so it is not conditional on the intended response.

trust the assertion that "Our system is set up so that it's safe to report an issue" but do not understand the garbling mechanism. We address this challenge in detail in Section 7.1.

Rapport-building (RB):

- Status quo approach: survey enumerators follow a typical social science research introduction script before beginning the survey and then ask the survey questions.
- RB approach: survey enumerators allocate survey time to build rapport, or trust, with the participant. RB entails chatting about family and hobbies in a natural but prespecified manner, beyond the minimum small talk typical in social science surveys.[17]

We developed our RB treatment modules by combining insights from practitioners and policy-makers conducting surveys on sensitive issues, such as sexual abuse and gender-based violence (e.g. United Nations Human Rights Office, 2011, United Nations Statistical Office, 2014, Muraglia et al., 2020) and from research focused on protocols for criminal investigations of sexual abuse allegations (e.g. Cowles, 1988, Vallano and Compo, 2011, Hershkowitz et al., 2014). For details on the development of our RB approach and our RB modules, see the paper's Supplementary Materials.

The status quo approach is the control condition. RB is the treatment condition. We conduct a shorter and a longer version of RB to test for the possibility that the marginal returns of building rapport decrease quickly. RB1 is the baseline rapport-building section, in which the enumerator signals that they care about the worker, getting to know the respondent, using emotional mirroring and acknowledging them. RB2 is the extended rapport-building section, in which the enumerator becomes personable with the worker, who has the chance to ask them questions. The enumerator also shares a related experience.

Removing personally-identifying information (Low-PII):

- Status quo approach: ask survey respondents to answer questions that reveal relatively more PII; questions include production section or line number and direct supervisor.
- Low PII approach: limit the amount of PII requested from the survey respondent; no questions asked about production section or line number or direct supervisor.

[17] During training, survey enumerators developed and practiced the RB approach using role plays. The senior research associate had to approve each enumerator on their RB approach before the survey launch.

Asking questions that reveal relatively more PII is the status quo approach because surveys in organizational settings often explicitly or de facto reveal respondents' identities. Note that identifying respondents' teams is necessary to compute team-level statistics such as the number of manager involved in harassment, the number of victims associated to repeat offenders, and the degree of isolation of victims. This represents an unavoidable trade-off.

Table 2: Treatment Arms & Surveyed (Planned) sample sizes

		No Rapport	Rapport 1	Rapport 2	TOTAL
Direct elicitation	PII	Arm 1 412(476)	Arm 2a 190(225)	Arm 2b 188(229)	790(930)
	Low PII	Arm 3 197(226)	Arm 4 189(220)		386(446)
Hard garbling	PII	Arm 5 416(487)	Arm 6a 188(225)	Arm 6b 195(227)	799(939)
	Low PII		Arm 7 270(305)		270(305)
	Total	1025(1189)	837(975)	383 (456)	2245(2620)

Table 2 summarizes the combinations of the experimental treatment arms that we tested. Treatment arm 1 is the benchmark, as it represents the status quo survey approach. Ex ante, we identified treatment arm 7 as the most protective. This may not be the case, however, if RB, which entails asking the respondent for more information about themselves that is not recorded in the survey, erodes the benefit of not asking for respondents' PII. We shed light on this possibility by comparing Arms 3 and 4. The experimental conditions were introduced after respondents completed all non-harassment related survey modules. Online Appendix (OA) Figure A.1 displays the survey modules and location of treatment interventions.

There are small variations across HG treatment arms in the realized garbling rate, as it was blocked within team but not within treatment arm. Consequently, we use the realized garbling rate for each HG treatment arm in the analysis to ensure that differences in the realized garbling rate across treatment arms do not affect the treatment effect estimates.

4.3 Sampling and Assignment to Treatment Arms

Sampling. We conducted a stratified random selection of workers. We sampled workers from 4 types of production teams using the employee lists for both plants: sewing lines;

finishing teams; dry washing teams; and wet washing teams. Among these teams, we chose teams with 15 or more workers because we aimed to stratify the treatment assignment by team and gender. We were left with 112 eligible teams and a total of 5,948 eligible workers out of a workforce of 7,727 workers (77% of workers). Within these teams, we stratified workers by their sex, which we identified based on name (male, female, uncertain).[18] In a small number of teams with few members of one sex, we aggregated this sex to the smallest level that yielded a group size suitable for stratified assignment (e.g., production section-floor). We selected 9 workers per stratum, which aimed to ensure a minimum of 1 per stratum assigned to each arm. We then sampled larger strata in proportion to their share of the overall eligible worker population.

Based on power calculations, we targeted a sample size of 2,620 workers. Because we had access to the complete population at the 2 plants, we were able to replace workers who were unreachable or who declined to participate. We attempted to recruit a total of 3,578 workers by phone, and we achieved a final sample size of 2,245 workers (63% response rate) from the 112 teams. The main reason for non-response was that we were not able to reach workers by phone (76% of cases); of workers whom we reached, 96% agreed to participate, and we were ultimately able to survey 86%.[19] We obtained workers' phone numbers from the apparel manufacturer's HR department, so it is likely that the main cause of our inability to reach workers was outdated phone numbers. The response rates by treatment arm are all very similar (see the Supplementary Materials).[20]

We did not achieve our target sample size despite our ability to replace workers because we stratified our selection by team and gender, and for some strata, we ran out of candidate replacement workers. The median (mean) team in the data has 39% (41%) of its workers surveyed. The median and mean team-level response rate to the survey was 65% (see the Supplementary Materials for the distribution of response rates across teams).[21]

[18] Names in Bangladesh are highly gendered. As such, we were able to categorize names as male or female for 99.7% of eligible workers. Among surveyed workers with categorized names, the classification error rate was 2.76% for men (11 respondents) and 1% for women (17 respondents).

[19] We attempted to call workers a total of 9 times to recruit them.

[20] Workers were not aware of their treatment assignment when deciding whether to participate. In the Supplementary Materials, we predict survey response with variables available in the producer's HR data and with team-level harassment rates estimated using data from the survey. The team-level harassment rates are not significant predictors of survey response.

[21] One may be concerned that supervisors who engage in more harassment may have pressured workers not to participate. This would not affect the experiment's internal validity, as selection into the sample happens before a respondent knows their treatment assignment. However this would bias downward the inferences

Assignment to Treatment Arms. The unit of randomization is a worker, stratified by plant-production team and sex. As detailed under sampling above, in cases where there were too few men or women on a production team, we aggregated to the next highest level that yielded a sufficiently large stratum size. We implemented the randomization in Stata. We first randomly assigned one worker per stratum to each treatment arm because we wanted to ensure that all strata were represented in all treatment arms. For larger strata, we then randomly assigned workers to each treatment arm with probabilities of assignment that corresponded to the treatment arm's target share of the overall sample size. We used the *randtreat* package by Carril (2017) to address misfits across strata. To improve balance, we proceeded along the lines suggested by Banerjee et al. (2020). We conducted 10 randomizations and selected the one that performed best in terms of balance on two covariates available to the research team: tenure and skill group (high- versus low-).

Integrity of the experiment & baseline balance. As detailed in the Supplementary Materials, we implemented HG using a system of preassigned, encrypted 1-2 digit numeric codes for "Yes" and "No" responses. Our approach relied on enumerators' adherence to a protocol for inputting the code assigned to a respondent and to their intended report. During the data quality checking process, we found that 1 survey enumerator did not adhere to this protocol. Upon further questioning, it was confirmed that the enumerator understood the HG data entry protocol but had not adhered to it. This enumerator conducted 53 DE and 48 HG surveys, all of which we drop. We also drop one observation of a worker under age 18 who was accidentally surveyed. As a result, the sample size is 2,143 observations for the remainder of the paper. Table 3 presents summary statistics of our sample. OA Table A.1 presents team-level summary statistics for the teams represented in the survey, calculated for overall teams, including workers not sampled in the survey. OA Table A.2 shows balance tests for workers' characteristics across the main treatment conditions. OA Table A.3 presents balance tests for workers' characteristics across no rapport, short rapport, and long rapport treatment arms. There are no statistically significant differences across treatment conditions.

we draw about the organization in Section 6. OA Table A.13 reports the correlations between the team-level response rate and the team-level reporting rates for harassment with DE, HG, and the difference between the team-level reporting rates under the two. On the whole, the correlations are small or zero, and most are not statistically significant. For threatening behavior, the correlations are weakly negative, suggesting teams with higher rates of threatening behavior have slightly lower response rates. Altogether, we do not believe that more problematic supervisors are systematically pressuring workers not to participate.

Table 3: Summary Statistics

	Mean	SD	Min	p25	p50	p75	Max
Female	0.81	0.39	0	1	1	1	1
Currently Working	0.96	0.20	0	1	1	1	1
Age	26.8	5.14	18	23	26	30	55
Experience (yrs)	5.19	3.57	0	2.83	4.42	7.17	28.8
Tenure (yrs)	2.89	2.43	0.052	0.65	2.82	4.17	17.0
Tenure in Team (yrs)† [n=1515]	2.57	2.52	0	0.50	1.83	3.92	14.5
Years of Education	6.70	3.39	0	5	6.50	9	16
Marital Status (1=Yes)	0.82	0.38	0	1	1	1	1
Children (1=Yes)	0.74	0.44	0	0	1	1	1
Sewing Section	0.49	0.50	0	0	0	1	1
Finishing Section	0.34	0.47	0	0	0	1	1
Washing Section	0.17	0.38	0	0	0	0	1
Position: Helper	0.17	0.38	0	0	0	0	1
Position: Ironing/Folding	0.086	0.28	0	0	0	0	1
Position: Operator	0.60	0.49	0	0	1	1	1
Position: Packer	0.044	0.20	0	0	0	0	1
Position: Quality	0.097	0.30	0	0	0	0	1

Notes: This table reports summary statistics on workers' characteristics. Unless otherwise noted, the sample includes 2,143 workers who participated in our survey. †This variable is available for the 1,515 respondents who were assigned to status quo PII collection treatment arms, in which we collected respondents' team id, manager name, and tenure on their team.

5 The Impact of Survey Design

This section reports findings from the survey experiment. We describe outcomes for the main treatment conditions. We then assess HTEs by gender, and examine whether HG, RB, and Low PII treatments are substitutes or complements. Finally, we conduct robustness checks.

5.1 Specifications

We seek to estimate coefficients in the following regression:

$$r_{is} = \alpha\, HG_i + \beta\, Rapport_i + \gamma\, LowPII_i + \mu_s + \theta X_i + \epsilon_{is} \tag{6}$$

where r_{is} is the intended reporting outcome of interest for individual i in stratum s. HG_i, $Rapport_i$ and $LowPII_i$ are indicators for hard-garbling, rapport, and not asking for team-related identifying information. μ_s are stratum fixed-effects. We present results without and

with controls for individuals' characteristics X_i, which are selected using the post double selection lasso (Belloni et al., 2014, referred to as PDS going forward). The vector of eligible control variables includes all worker characteristics reported in Table 3.

Identification using garbled responses. For individuals in the HG arms, we observe garbled response $\widetilde{r}_i$ instead of intended response r_i. However, following Blair et al. (2015), we note that recorded reports can be expressed as

$$\widetilde{r}_i = r_i + (1 - r_i)(\pi + \varepsilon_i)$$

where ε_i is a mean-zero error, equal to $1 - \pi$ with probability π and equal to $-\pi$ with probability $1 - \pi$. We defined normalized recorded reports $\hat{r}_i$ by

$$\hat{r}_i \equiv \frac{\widetilde{r}_i - \pi}{1 - \pi} = r_i + \underbrace{\frac{1 - r_i}{1 - \pi}\varepsilon_i}_{\equiv \xi_i}$$

with $\pi = .2$ for the HG group and $\pi = 0$ for the DE group. Normalized report $\hat{r}_i$ is equal to the intended report plus a heteroskedastic error term. If r_i satisfies (6), then $\hat{r}_i$ satisfies a similar regression (6b) with heteroskedastic, mean-zero errors (conditional on covariates). Consequently, OLS is consistent, and robust standard errors are correct. With blocked-garbling, error terms ε_i are negatively correlated within blocks, and uncorrelated across. We estimate the following equation, in which ξ_{is} is now the residual, and report standard errors clustered by HG batch (HG respondents) or respondent (DE respondents):[22,23]

$$\hat{r}_{is} = \alpha\, HG_i + \beta\, Rapport_i + \gamma\, LowPII_i + \mu_s + \theta X_i + \xi_{is} \tag{6b}$$

HTE analysis by respondents' sex. We also estimate treatment effects separately for women and for men:

$$\begin{aligned}\hat{r}_{is} = {} & \alpha_f\, HG_i * Female_i + \alpha_m\, HG_i * Male_i + \beta_f\, Rapport_i * Female_i + \beta_m\, Rapport_i * Male_i \\ & + \gamma_f\, LowPII_i * Female_i + \gamma_m\, LowPII_i * Male_i + \lambda\, Female_i + \mu_s + \theta X_i + \xi_{is}\end{aligned} \tag{7}$$

[22] We pre-specified that we would report robust standard errors, so reporting clustered standard errors is a deviation from our PAP. Since sampling errors are negatively correlated within HG batch, this deviation is justified. See NBER Working Paper No. 31011 for results estimated using robust standard errors.

[23] For regressions that do not use intended reports, we report robust standard errors as per our PAP.

Complementarity across treatments. We test for complementarity vs. substitutability across treatments by estimating the effects separately for each arm. The omitted category is $\mathbb{1}(\text{DE} \times \text{PII} \times \text{No RB})_i = 1$, which is treatment arm 1, the control condition.

$$\begin{aligned}\hat{r}_{is} = {} & \alpha_1 \mathbb{1}(\text{DE} \times \text{PII} \times \text{RB 1})_i + \alpha_2 \mathbb{1}(\text{DE} \times \text{PII} \times \text{RB 2})_i + \alpha_3 \mathbb{1}(\text{DE} \times \text{Low PII} \times \text{RB 1})_i \\ & + \alpha_4 \mathbb{1}(\text{DE} \times \text{Low PII} \times \text{No RB})_i + \beta_1 \mathbb{1}(\text{HG} \times \text{PII} \times \text{No RB})_i + \beta_2 \mathbb{1}(\text{HG} \times \text{PII} \times \text{RB 1})_i \\ & + \beta_3 \mathbb{1}(\text{HG} \times \text{PII} \times \text{RB 2})_i + \beta_4 \mathbb{1}(\text{HG} \times \text{Low PII} \times \text{RB 1})_i + \mu_s + \theta X_i + \xi_{is} \end{aligned} \tag{8}$$

5.2 Results

Main effects of survey design on reporting. Table 4 reports the main treatment effects.[24] In regression tables throughout the paper, odd-numbered columns display the results from the baseline specification, while even-numbered columns display the results with PDS lasso-selected controls (based on all worker characteristics reported in Table 3).

In the control arm, (DE × PII × No RB), 9.9% of workers report experiencing threatening behavior, 1.53% report being physically harassed, and 1.78% report being sexually harassed by their supervisor. Among workers who report being harassed under DE, meaning respondents in arms 1-5, 43% who experienced threatening behavior reported it through one of their factory's internal channels, 52% who were physically harassed reported it, and 68% of those who were sexually harassed did. Based on the mean reporting rates for arms 1-5, from the producer's perspective, it would have detected that 3.7%, 0.98%, and 1.87% of workers, respectively, experienced threatening behavior, physical harassment, and sexual harassment by their supervisor in the past year.

We now turn to the effect of survey design. In percentage points (ppts), HG increases the reporting of threatening behavior, physical harassment, and sexual harassment, respectively, by 4.5 ppts (or 45.4%, p <0.01), 4.4 ppts (or 290%, p <0.01), and 4.8 ppts (or 271%, p <0.01).[25] Removing questions about respondents' supervisor (Low PII) increases the reporting of physical harassment by a marginally statistically significant 2.9 ppts (or 188%, p =0.108) but has no effect on the reporting of threatening behavior or sexual harassment. Building rapport appears to have a positive effect on the reporting of threatening behavior (1.20 ppts, or a 11.6% increase) and sexual harassment (1.90 ppts, or a 107% increase), but

[24] OA Table A.4 reports the main results with separate indicators for short- and long-RB conditions.

[25] The HG coefficients are similar across columns, but the effects are heterogeneous by gender (Table 5).

it is not statistically significant. Rapport has no detectable effect on physical harassment.

Table 4: Effects of Survey Design on Reporting of Harassment

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
HG Treatment	0.0445***	0.0448***	0.0438***	0.0451***	0.0478***	0.0487***
	(0.0150)	(0.0145)	(0.0121)	(0.0117)	(0.0114)	(0.0111)
Rapport Treatment	0.0113	0.0140	-0.0094	-0.0082	0.0188	0.0186
	(0.0202)	(0.0195)	(0.0200)	(0.0192)	(0.0183)	(0.0176)
Low PII Treatment	0.0102	0.0097	0.0280	0.0299*	0.0045	0.0067
	(0.0245)	(0.0239)	(0.0184)	(0.0178)	(0.0203)	(0.0201)
Control Group Mean	.0992	.0992	.0153	.0153	.0178	.0178
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes	No	Yes
Observations	2140	2140	2140	2140	2140	2140

Notes: This table reports OLS estimates of treatment effects on workers' reporting. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Together, these results make a compelling case for the importance of plausible deniability in the design of information transmission mechanisms in organizational settings. The costs and benefits of removing team-level identifying questions are much less clear. Low PII appears to increase the reporting of physical harassment, but there is no effect on threatening behavior or sexual harassment. It also comes at the cost of not being able to calculate manager-level statistics that may be valuable to decision-makers. Finally, we cannot reject that RB has no effect on reporting. It is possible that this null result masks heterogeneous effects across respondents or that RB's effect depends tightly on the survey design.

We discuss the robustness of these findings in Section 7. We first show that confusion among respondents does not explain the impact of HG on reporting. We also show that treatment status does not affect reporting behavior associated with a placebo question using DE for all treatment arms. This suggests that respondents' behavior reflects a real understanding of the benefits afforded by plausible deniability. We then discuss how to interpret our findings if there are concerns over false reporting by workers.

Effects of survey design on reporting by men and women. Motivated by the possibility that the experience of harassment and the utility generated by reporting harassment is different for men and women, we estimate the main effects separately by sex in Table 5. In our control arm, 19.12% of men report experiencing threatening behavior, 4.41% report experiencing physical harassment, and 1.47% report experiencing sexual harassment. Reporting rates among women are very different: 8.0% report threats, 0.92% report physical harassment, and 1.85% report sexual harassment. We cannot disentangle whether these differences are due to differential incidences of harassment or differential reporting. Among respondents who report being harassed under DE, across all forms of harassment, women are more likely to say that they reported their experience through an internal channel.[26]

As in the analysis presented Table 4, HG continues to increase reporting across the board. Interestingly, the point estimates of the effects are particularly large for men, but because our sample of men is small, with the exception of threatening behavior, we cannot reject that the effects are the same for men and women. The impact of removing PII appears to be weakly more positive for women, although standard errors increase, and we cannot reject that the effects for both groups are zero or the same. Table 5 suggests that the impact of rapport may be different on men and women. For both threatening behavior and sexual harassment, rapport appears to have increased reporting among women and may have backfired for men. This is plausible: survey enumerators were women, and being forced into small talk with an unknown woman may have raised men's suspicion regarding the survey. This suggests that further experimentation, and better tailoring of RB, is needed to assess its value.

Interactions among treatment conditions. We examine the possibility that the treatment conditions may substitute or complement each other using regression equation (8). Figure 1 summarizes the results, which are presented in OA Table A.5. The omitted category is the control arm, DE $\times$ PII $\times$ No RB.

The top 4 treatment conditions illustrated by Figure 1 correspond to reports elicited using DE. To a first order, removing PII and building rapport do not seem to have a large impact. There maybe an impact of extended rapport on the reporting of sexual harassment. While this individual coefficient is significant, the overall picture invites caution. OA Figure A.2

[26] In the DE arms, we asked respondents who reported being harassed whether they had reported it through internal channels. Due to the garbling, we were unable to ask this question in the HG arms.

Table 5: Effects of Survey Design on Reporting of Harassment, Heterogeneity by Sex

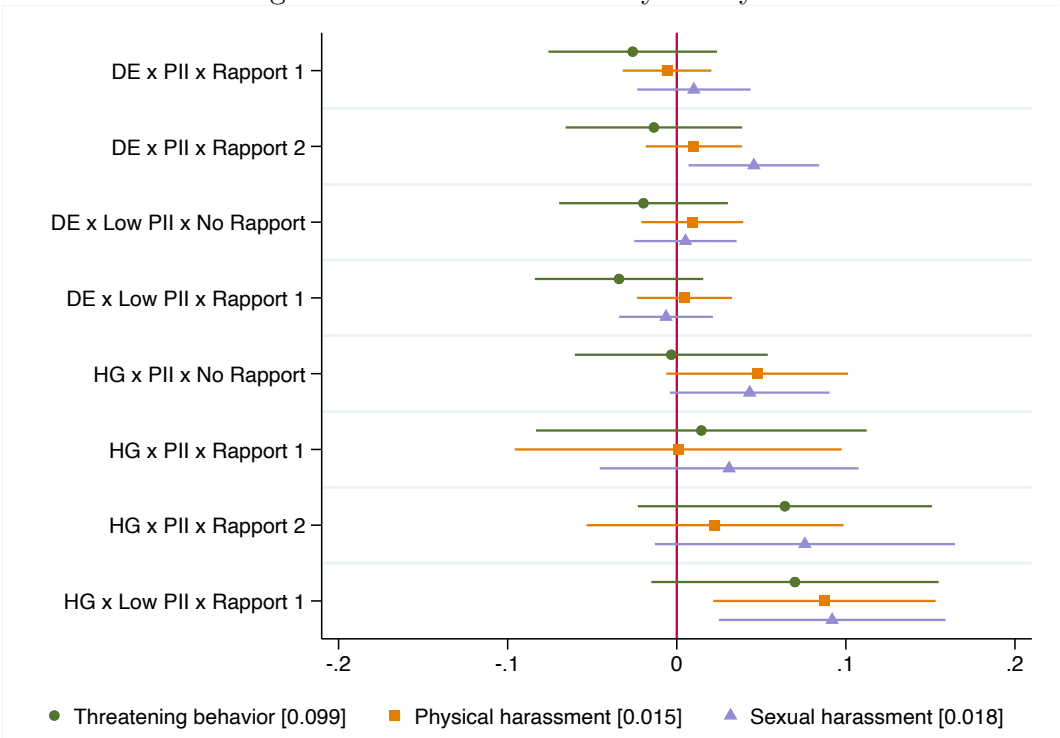
	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
HG Treatment × Female	0.0274	0.0276*	0.0405***	0.0420***	0.0371***	0.0387***
	(0.0171)	(0.0165)	(0.0132)	(0.0127)	(0.0133)	(0.0131)
HG Treatment × Male	0.1224***	0.1199***	0.0597*	0.0587*	0.0917***	0.0928***
	(0.0418)	(0.0408)	(0.0347)	(0.0335)	(0.0351)	(0.0344)
Rapport × Female	0.0193	0.0218	-0.0173	-0.0151	0.0304	0.0305
	(0.0228)	(0.0219)	(0.0234)	(0.0225)	(0.0204)	(0.0194)
Rapport × Male	-0.0233	-0.0230	0.0243	0.0204	-0.0371	-0.0360
	(0.0467)	(0.0449)	(0.0370)	(0.0360)	(0.0460)	(0.0459)
Low PII Treatment × Female	0.0132	0.0137	0.0326	0.0343*	0.0105	0.0127
	(0.0263)	(0.0258)	(0.0208)	(0.0199)	(0.0229)	(0.0224)
Low PII Treatment × Male	-0.0067	-0.0111	0.0120	0.0140	-0.0259	-0.0245
	(0.0549)	(0.0532)	(0.0457)	(0.0455)	(0.0402)	(0.0392)
Female	-0.0900	-0.0991	-0.0211	-0.0112	0.0682	0.0886
	(0.1059)	(0.1020)	(0.0751)	(0.0750)	(0.0776)	(0.0745)
Control Mean - Female	.08	.08	.0092	.0092	.0185	.0185
Control Mean - Male	.1912	.1912	.0441	.0441	.0147	.0147
p(HGxFemale - HGxMale)	[0.045]	[0.045]	[0.614]	[0.649]	[0.175]	[0.172]
p(RapportxFemale - RapportxMale)	[0.419]	[0.374]	[0.351]	[0.414]	[0.188]	[0.190]
p(NoPIIxFemale - NoPIIxMale)	[0.735]	[0.663]	[0.689]	[0.689]	[0.421]	[0.393]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Lasso	No	Lasso	No	Lasso
Observations	2140	2140	2140	2140	2140	2140

Notes: This table reports OLS estimates of treatment effects by gender heterogeneity on workers' reporting. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the gender interactions of treatment variables and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

shows that the negative effect of RB on reporting of threatening behavior is driven by men, while its positive effect on reporting of sexual harassment is driven by women. This suggests that RB does impact reporting, but that it's effect is subtle and varies across participants. Our interpretation is that RB needs to be carefully tailored to the respondent.

The bottom 4 treatment conditions illustrated by Figure 1 correspond to reports elicited using HG, first on its own, then introducing RB and Low PII. Estimated treatment effects appear to be rising as additional trust-enhancing steps are taken. This contrasts with patterns under DE, where trust-enhancing steps are not accompanied with increasing treatment

Figure 1: Treatment effects by survey arm



Notes: This figure reports coefficients from separate regressions of the outcome variable on the treatment arm indicators, strata fixed effects, and controls selected using the PDS lasso. The regression specification is equation (8). The whiskers are 95% confidence intervals estimated using robust standard errors. The omitted category is treatment arm 1, $\mathbb{1}(\text{DE} \times \text{PII} \times \text{No RB})_i = 1$, which is the control condition. The number in square brackets is the reporting rate for this group.

effects. Altogether, this suggests that there exist complementarities between HG and other steps that foster trust in the survey protocol. This is intuitive since HG can only be effective if respondents trust that protocol will be followed.

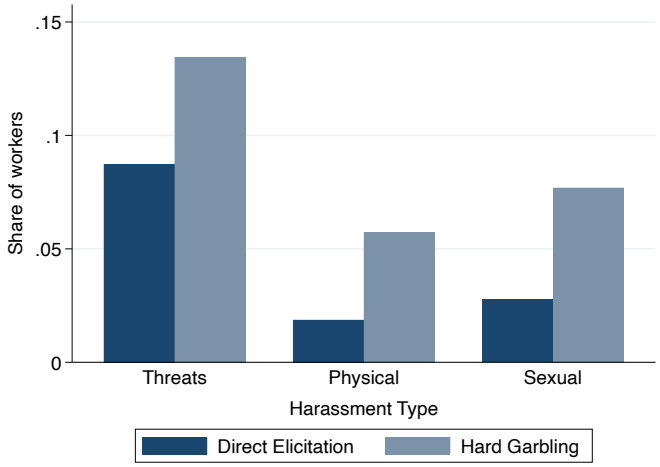
This visual intuition corresponds to the fact that the point estimate for the effect of (HG $\times$ Low PII $\times$ RB 1) is larger than the sum of the point estimates for (DE $\times$ PII $\times$ RB 1)+ (DE $\times$ Low PII $\times$ No RB) + (HG $\times$ PII $\times$ No RB) for all three harassment outcomes. We test the null hypothesis of no complementarity among HG, removing team-level identifying information, and RB in the complementarity test reported at the bottom of OA Table A.6, focusing on even-numbered columns, which include PDS-lasso-selected controls). The test is rejected for threatening behavior (p=0.035) but is too imprecise to be rejected for physical harassment (p=0.247) and sexual harassment (p=0.256). We thus interpret this as suggestive evidence of complementarity among the design features.

6 Understanding Harassment

In this section, we use our improved survey data to assess the scope and nature of the harassment problem in the apparel producer's organization. Given the large effect of HG on reporting, we apply the estimators from Section 3 to compute team-level statistics with data pooled across treatment arms that use HG and collect PII (when PII are needed).

We begin by describing the patterns of harassment in the organization, and then discuss their policy implications.[27] Our main goal is to illustrate the value of inferring team-level statistics from garbled reports, but the specific patterns of harassment estimated for our partner organization suggest policy insights that plausibly apply to other organizations.

Figure 2: Share of workers who have been victimized ($\mathbf{S}_V$) by survey method



Notes: This figure reports harassment rates estimated using reporting with DE and HG, respectively. We compute the harassment rates using (3) under DE and HG. For both DE and HG, we pool across all treatment arms, including the RB arms and the arms in which we do not collect team-level identifying information.

Scale of the issue and potential gains. Figure 2 illustrates the estimated share of victimized workers, computed using (3) under DE and HG. Since this statistic does not require PII, we pool data across all arms.

As we have already discussed, HG considerably increases workers' propensity to report harassment: 13.5% reported threatening behavior with HG compared to 8.7% with DE, 5.7% reported physical harassment with HG compared to 1.9% with DE, and 7.7% reported sexual harassment with HG compared to 2.8% with DE.

[27] One may worry that supervisors who engage in more harassment may have pressured workers not to take the survey, which would mean that our statistics are downward biased. See Footnote 21 for a discussion.

The primary takeaway is that harassment is meaningfully more widespread than standard surveys, or the firm's internal reporting channels would suggest. This means that addressing harassment may have a much more positive impact on overall employee welfare than what previously available data would lead one to conclude. We also note that since both men and women report significant levels of harassment under HG, addressing harassment would likely benefit both groups.

Problem managers and isolated victims. We now turn to team-level characteristics of interest, $\mathsf{S}_{TV\geq k}$ and $\mathsf{E}_{2V|1V}$, computed by pooling garbled reporting data from treatment arms that use HG and collect PII. We estimate the distribution of number of intended reports in the sample population of teams using the unbiased consistent estimator of Proposition 2. In our setting, team size for workers in treatment arms that use HG and collect PII varies from 3 - 12 workers, with a median and mean of 7 workers. As such, we apply the estimator for each team size in the data and average the estimates to obtain team-population statistics.

To use Proposition 2 under blocked garbling, the assumption that $\widetilde{\mu}(L) = 0$ needs to be satisfied. It is true in the sample of recorded reports for all team sizes in our data. This is illustrated in OA Table A.7. Even though the modal team size is 7, no team in our sample has more than 5 recorded "Yes" responses across different harassment types.

Table 6: Estimated team-level statistics $\mathsf{S}_{TV\geq k}$ for $k \in \{1, 2\}$ and $\mathsf{E}_{2V|1V}$

Statistic	Threatening Behavior	Physical Harassment	Sexual Harassment
$\mathsf{S}_{TV\geq 1}$	0.719	0.254	0.402
	(0.068)	(0.045)	(0.053)
$\mathsf{S}_{TV\geq 2}$	0.112	0.035	0.033
	(0.038)	(0.019)	(0.022)
$\mathsf{E}_{2V|1V}$	0.166	0.137	0.083
	(0.055)	(0.076)	(0.056)

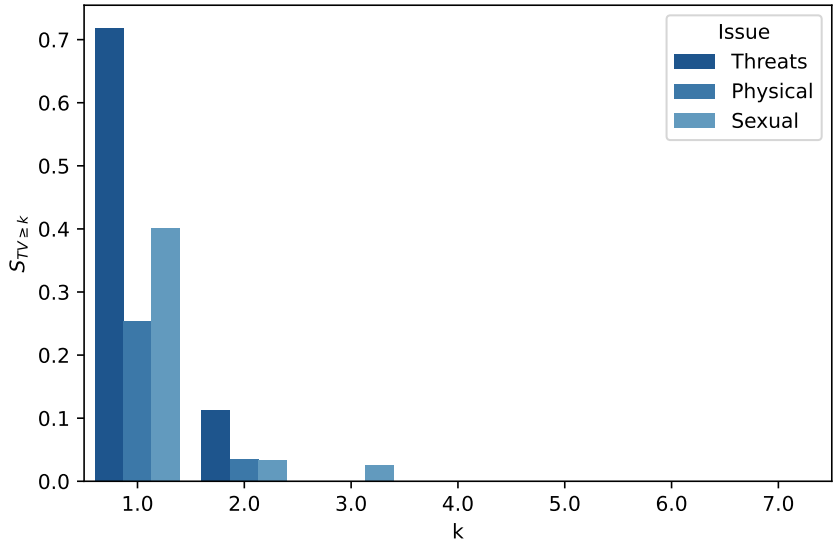
Notes: This table reports team-level statistics $\mathsf{S}_{TV\geq k}$ for $k \in 1, 2$ and $\mathsf{E}_{2V|1V}$, estimated using pooled data from treatment arms that use HG and collect PII. We estimate the statistics using the unbiased consistent estimator of Proposition 2. The sample includes 112 teams.

Table 6 reports $\mathsf{S}_{TV\geq k}$ for $k \in \{1, 2\}$ and $\mathsf{E}_{2V|1V}$. The estimated share $\mathsf{S}_{TV\geq 1}$ of problem teams in which at least one worker has been harassed is 72% for threatening behavior, 25% for physical harassment, and 40% for sexual harassment. The share of teams with at least

2 victimized workers is 11% for threats, 3.5% for physical harassment, and 3.3% for sexual harassment. This confirms that the bulk of the challenge consists in dealing with fairly widespread medium intensity harassment, rather than dealing with a fairly circumscribed group of high intensity offenders: $k\mathsf{S}_{TV\geq k}$ is not high for k large.

Across all types of harassment, victims are relatively isolated: conditional on getting a report, the likelihood of getting at least another one is equal to 17% for threats, 14% for physical harassment, and 8% for sexual harassment. Figure 3 visualizes this finding; it plots the full distribution of $\mathsf{S}_{TV\geq k}$ for each type of harassment and shows that no teams have 3 or more victims for threats or physical harassment and a small share of teams have 3 victims for sexual harassment. A caveat for this statement is that our estimate of intended reports is likely a lower bound to the real amount of harassment. For this reason, the finding is better stated as: victims *willing to report under HG* are isolated.

Figure 3: Share of teams with k or more victimized workers.



Notes: This figure reports the full distribution of $\mathsf{S}_{TV\geq k}$ for each type of harassment, estimated using pooled data from treatment arms that use HG and collect PII. We estimate the statistics using the unbiased consistent estimator of Proposition 2. The sample includes 112 teams.

Policy implications of estimated prevalence of harassment. Our findings suggest three policy takeaways. First, harassment is much more prevalent than filed complaints would suggest, and it affects both men and women. Second, harassment occurs at a moderate intensity but is widespread across teams. This means that firing a few bad apples cannot be the sole policy option. Instead, the behavior of existing managers must be changed.

Nonetheless, it would be beneficial to prioritize the worst offenders. Third, the extent to which victims are isolated in teams varies substantially by type of harassment. This has implications for setting different burdens of proof for harassment: when victims are more isolated, requiring multiple victims to come forward, for example, to avoid "he said, she said" situations, may miss the majority of cases. Remedying harassment likely requires having actions available that can be taken in cases when only one victim comes forward.

7 Discussion

Our work makes two main contributions. First, we evaluate the impact of different aspects of survey design – HG, RB, and removing team level information – on respondents' propensity to report harassment in a real-life organizational context. Second, we identify policy-relevant statistics of harassment and derive consistent estimators for these statistics when survey data are generated using HG. Applying these estimators to our data clarifies how they can guide policy responses to harassment in organizations.

Our experimental results show that lack of plausible deniability causes severe under-reporting of harassment in this organizational setting. Lack of trust in the integrity of the reporting system may also contribute, though our results suggest that the process of trust-building is highly contextual and may backfire when not well-targeted. In our context, harassment appears to be widespread, a majority of managers exhibit some propensity to harass workers, and victims are frequently isolated.

7.1 Robustness

In the remainder of this section we address some of the robustness concerns related to our findings and discuss how they might be further explored.

7.1.1 How did respondents understand the HG treatment?

It is legitimate to ask whether any potential effect of HG on responses is due to respondents understanding the implications of garbling on ex post updating by managers and subsequent retaliation or due to other mechanisms. We investigate two alternative mechanisms. First, we assess whether confusion in the HG condition could explain our results. Second, we consider whether the HG script increased trust among respondents in the HG condition.

Confusion in the HG condition. One concern with our findings is that HG is a more complicated mechanism than DE.[28] This means that respondents may be confused by HG, and that confusion may be more likely under HG compared to DE. This concern is especially relevant in our context, in which the average survey respondent has 6.70 years of schooling (Table 3). We were concerned about this possibility, so we included two comprehension questions in the HG module.[29] Respondents answered these prior to being asked the questions about harassment, and survey enumerators explained the answers to the comprehension questions after asking them.

8.8% of HG respondents answer at least 1 comprehension question incorrectly, while 4.8% answer 2 incorrectly. Women and men answer incorrectly at somewhat similar rates: 9.6% of men and 8.6% of women in HG answer at least 1 question incorrectly ($p = 0.685$), while 6.9% of men and 4.3% of women answer 2 questions incorrectly ($p = 0.133$).

While the surveyor would desire for respondents who are confused by HG to answer "no" to avoid false positives, in practice, reporting rates are weakly higher among confused respondents. Consequently, we evaluate whether asymmetric confusion among HG versus DE respondents could explain our results. We adopt a very conservative approach and re-estimate our main results setting to "No" the *recorded* answers to harassment questions of respondents who answer any comprehension question incorrectly. Panel A of OA Table A.9 reports the results; focusing on the HG effects and comparing them to the estimates in Table 4, for threatening behavior, column (2) shows that the effect is now a 3.6 ppt increase ($p < 0.05$) compared to a 4.5 ppt increase ($p < 0.01$). For physical harassment, the effect is a 3.8 ppt increase ($p < 0.01$) compared to a 4.4 ppt increase ($p < 0.01$). For sexual harassment, the effect is a 3.7 ppt increase ($p < 0.01$) compared to a 4.8 ppt increase ($p < 0.01$). Even under this conservative treatment of reports from confused respondents, the effects of HG are positive, large, and statistically significant.[30] Turning to Panel B, while the point estimates for both sexes are attenuated by similar magnitudes as in Panel A, there is not a differential pattern of attenuation by sex, and the patterns of heterogeneity are unchanged.

[28] Explaining HG takes more time. This increases average survey duration by 4% (OA Table A.8).

[29] The questions were, "Can you please tell me whether the following statements are true or false: (a) If I respond 'Yes,' no one can ever know this for sure. (b) The system will record at least one out of every five workers' responses as 'Yes.' " The script explaining HG, including the comprehension questions, is included in the paper's Supplementary Materials.

[30] Less conservative approaches would be to exclude these respondents from the analysis or to set their intended response to "no" and simulate out their recorded responses.

General assurances in HG condition. The script explaining HG included statements such as, "our system is set up so that it's safe to report an issue." This raises the question of whether respondents trust these assertions, and this increases their willingness to report with HG. We think that this is unlikely for two reasons. First, DE and HG respondents were both assured of a strong commitment to confidentiality as part of the IRB-required informed consent process, but even under this very strong general security assurance, baseline reporting rates for more sensitive types of harassment are very low among DE respondents.[31]

Second, the high rate of correct responses to comprehension questions suggests that respondents were attentive to the explanation of HG and comprehended the mechanics of how it works. Alternatively, it is possible that respondents perceived the HG script as a signal of the research team's general trustworthiness, and this made them more willing to report in general. If so, we would expect higher reporting rates in the fourth harassment question asked using DE for all respondents, which is not the case (OA Table A.10).[32]

Placebo question. Our survey module included a sensitive question that appeared directly following the three harassment questions and was asked using DE for all respondents. Specifically, we asked whether their supervisor had refused to forward their request for leave to their factory's administration or required them to perform certain tasks to submit it. If the effect of treatment is driven by respondents' understanding of the benefits of plausible deniability, then responses to this question should not be affected by treatment. OA Table A.10 shows that the share of HG respondents reporting "Yes" to this question is indeed indistinguishable from that of the DE respondents. This suggests that HG respondents understood that the survey mechanism had changed and was less secure.

In sum, the evidence is consistent with the vast majority of respondents in HG having some comprehension of how the HG mechanism works and responding to the specific assurances that it provides. It is less consistent with the key mechanism being general purpose reassurances provided in the HG script and is inconsistent with increased confusion under HG. That said, we acknowledge that we cannot provide a definitive answer to whether the effects are due to the specific or to the general purpose reassurances provided in the HG

[31] The text of our informed consent script is included in the Supplementary Materials.

[32] We examine enumerators' assessments of respondents after the survey. Enumerators answered 5 questions about respondents' behavior during the (entire) survey; we test within-enumerator differences in perceived respondent behavior across treatment conditions. Enumerators perceived HG respondents to be more honest, but no different from DE respondents on other behaviors, including trust (OA Table A.11).

script, and we think this is a natural question to ask in follow-up studies. For instance, one could run trial with a *weak-HG* treatment, in which the same language is used, but only 1-in-100 responses is required to be a "Yes." If this treatment is more effective than DE, then this is evidence that respondents react to the intent conveyed by the description of HG, rather than to quantitative aspects of the design. Finally, we note that even if respondents simply trust our wording in a one-shot setting, it is important for us to be able to implement the plausible deniability features we advertise. In the ongoing monitoring scenario we have in mind trust needs to be maintained and continuously earned.[33]

7.1.2 Strategic misreporting by workers and follow-up actions

In our conceptual framework, we assume that there are no false positives in reporting; workers either report their true harassment status or they report that they have not been harassed. As discussed in Section 3, we think that this is an appropriate assumption for our setting, at least in the short-run. One may still be concerned that this is a strong assumption. For example, workers who are motivated by career concerns may take advantage of the plausible deniability provided by garbling to try and take down innocent supervisors. This may especially be true for men, who are much more likely to be promoted into supervisor positions. If so, it provides an alternative explanation for the patterns of HTEs that we find.

Empirical evidence. This possibility is a priori unlikely in our context given the very low baseline rates of reporting and the stigma that victims face. We provide empirical evidence consistent with this view. To do so, we split our sample by sex and by whether the respondent has at least 8 years of schooling, an informal cutoff used by factories to determine workers' eligibility to become a supervisor.[34] If workers are strategically misreporting, we expect that our effects will be driven by workers with at least 8 years of schooling, who are disproportionately more eligible to become supervisors, especially among men. The results, in OA Table A.12, show that there is no consistent pattern of HTEs for men or women with more or less than 8 years of schooling. Sometimes the effects are larger for the group with less schooling, sometimes smaller, and sometimes the same. This evidence goes against the hypothesis that strategic misreporting due to career concerns is driving our results.

[33] Chassang and Zehnder (2019) provide preliminary lab evidence that HG has a lasting effect on reporting, in contrast to RR, which unravels because respondents don't comply with instructions.

[34] In a survey conducted with supervisors and other lower-level managers employed by the apparel producer, 87% of supervisors had at least 8 years of schooling. 22% had exactly 8 years of schooling, a large jump up from the 8% of managers reporting having the next lower category, "some middle school education."

Model guidance. Chassang and Padró i Miquel (2018) explicitly study equilibrium whistleblowing in a model where managers make endogenous retaliation choices, and workers may have malicious incentives to submit false accusations. They show that it is possible to achieve robust bounds on the underlying level of misbehavior by:

1. Starting from a low level of enforcement, reduce the information content of reports up to a point where workers are willing to complain.
2. Keeping the information content of reports the same, scale up enforcement.

In our context enforcement is minimal: the only action associated with a report of harassment is a change in the aggregate statistic reported to firm executives. Our paper can be viewed as achieving step (1). Investigating step (2) is a central question for future research.

7.2 Directions for future research

How to take governance actions as a function of garbled reports strikes us as a particularly valuable direction for future research. The choice of follow-up actions must reflect several considerations. First, actions need to be an acceptable, legitimate response to an inherently noisy signal. Sending the manager to a training seminar, initiating a more thorough yearly review, or moving the worker associated with the report to a new team may be appropriate responses to noisy evidence, whereas firing the manager would not be. Note that moderation in responses may be a plus from the perspective of the organization's leadership. Stronger evidence may lead to costly repercussions out of the organization's control, leading organizations to avoid information in the first place. Second, some follow-up actions are more likely than others to attract the interest of malicious workers. Sending a manager to a training seminar is unlikely to benefit a worker interested in sabotaging a manager's career, but linking recorded reports to managers' promotion opportunities would.

Building on recent research on the labor market implications of harassment by Adams-Prassl et al. (2022), Folke and Rickne (2022), and Dahl and Knepper (2021), we believe that evaluating the mental health and broader welfare effects of reporting harassment for workers, managers and producers is another import direction for research. Sociological studies suggests that the act of confiding secrets can improve an individual's well-being through improving one's perceived coping ability and reducing one's mental load associated with the

secret (Slepian and Moulton-Tetlock, 2019). To explore this possibility, we resurveyed workers two weeks after the survey experiment to test whether reporting harassment improved workers' mental well-being and job satisfaction. We estimate a 2SLS model with the randomized assignment to the treatments as our instruments for reporting. OA B details our empirical strategy and results. It provides suggestive evidence that reporting harassment improves workers' mental well-being and job satisfaction. The effects are large, and consistent across questions, but imprecisely estimated. The effect on job satisfaction also suggests one possible mechanism for beneficial effects to flow to the producer.

Proofs Appendix

Proof of Proposition 2. Recall that μ and $\widetilde{\mu}$ denote the distribution of team-level intended and recorded complaints. For all $k \in \{1, \cdots, L\}$ let us define $p_k \equiv \text{prob}_{\mu}\left(\sum_{i \in I} r_i = k\right)$ and $\widetilde{p}_k \equiv \text{prob}_{\widetilde{\mu}}\left(\sum_{i \in I} \widetilde{r}_i = k\right)$. Under i.i.d. garbling with garbling rate π, distribution parameters $(p_k)_{k \in \{1,\cdots,L\}}$ and $(\widetilde{p}_k)_{k \in \{1,\cdots,L\}}$ are related as follows:

$$\begin{aligned}
\widetilde{p}_0 &= p_0(1-\pi)^L \\
\widetilde{p}_1 &= p_0 \binom{L}{1} \pi (1-\pi)^{L-1} + p_1 (1-\pi)^{L-1} \\
\widetilde{p}_2 &= p_0 \binom{L}{2} \pi^2 (1-\pi)^{L-2} + p_1 \binom{L-1}{1} \pi (1-\pi)^{L-2} + p_2 (1-\pi)^{L-2} \\
\forall k \in \{1, \cdots, L\}, \quad \widetilde{p}_k &= \sum_{n=0}^{k} p_n \binom{L-n}{k-n} \pi^{k-n} (1-\pi)^{L-k}.
\end{aligned}$$

This is a triangular system of linear equation which means we can infer p_ks using observed $\widetilde{p}_k$s using the following recursion:

$$\begin{aligned}
p_0 &= \frac{1}{(1-\pi)^L} \widetilde{p}_0 \\
p_1 &= \frac{1}{(1-\pi)^{L-1}} \widetilde{p}_1 - p_0 \binom{L}{1} \pi \\
\forall k \in \{2, \cdots, L\}, \quad p_k &= \frac{1}{(1-\pi)^{L-k}} \widetilde{p}_k - \sum_{n=0}^{k-1} p_n \binom{L-n}{k-n} \pi^{k-n}.
\end{aligned}$$

This concludes the proof that μ is identified given $\widetilde{\mu}$. The same result holds under population-blocked garbling since the distribution of $\widetilde{\mu}$ conditional on μ are asymptotically identical

under i.i.d. garbling and population-blocked garbling as m grows large.

We now consider the case of blocked garbling, with teams of size L and g responses forced to be equal to YES, with $0 < g < L$. Given a distribution $\mu \in \Delta(0, \cdots, L)$ for the number of intended YES responses, the distribution $\widetilde{\mu}$ of garbled reports is given by

$$\begin{aligned} \forall k \in \{0, \cdots, g-1\}, \quad & \widetilde{p}_k = 0 \\ \forall k \in \{0, \cdots, L-g\}, \quad & \widetilde{p}_{g+k} = \sum_{h=0}^{g} p_{k+h} \binom{k+h}{h} \binom{L-k-h}{g-h} \end{aligned} \tag{.1}$$

In general the system of equations (.1) is not invertible. However it is invertible whenever $\widetilde{p}_L = 0$. Given that terms $(p_k)_{k \in \{0,\cdots,L\}}$ are positive, this implies that $p_L = p_{L-1} = \cdots = p_{L-g} = 0$. These additional conditions turn (.1) into an invertible triangular system of equations, which concludes the proof.

In all cases, given L and g, we get a matrix M such that $\mu = M\widetilde{\mu}$. In finite samples, we can replace $\widetilde{\mu}$ with its sample counterpart $\widehat{\widetilde{\mu}}$. Since $\widehat{\widetilde{\mu}}$ is an unbiased consistent estimator of $\widetilde{\mu}$, it follows that $\widehat{\mu} \equiv M\widehat{\widetilde{\mu}}$ is a consistent and unbiased estimator of μ. ■

Proof of Equation (5).

$$\begin{aligned} \mathsf{Var}\left(\sum_{j=1}^{n} r_j \eta_j \middle| \mathbf{r}\right) &= \sum_j r_j \mathsf{Var}(\eta_j) + \sum_{j \neq j'} r_j r_{j'} \mathsf{Cov}(\eta_j, \eta_{j'}) \\ &= \overline{r} n \pi(1-\pi) - \sum_{j,j'} r_j r'_j \frac{\pi(1-\pi)}{n-1} + \sum_j r_j \frac{\pi(1-\pi)}{n-1} \\ &= \overline{r} n \pi(1-\pi) - [\overline{r} n]^2 \pi(1-\pi) + \overline{r} n \frac{\pi(1-\pi)}{n-1} \\ &= \overline{r} \left(1 - \frac{\overline{r} n - 1}{n-1}\right) \pi(1-\pi) n. \end{aligned}$$

■

References

Adams-Prassl, A., K. Huttunen, E. Nix, and N. Zhang (2022): "Violence Against Women at Work," Tech. rep., mimeo.

Aguilar, A., E. Gutiérrez, and P. S. Villagrán (2021): "Benefits and Unintended

Consequences of Gender Segregation in Public Transportation: Evidence from Mexico City's Subway System," *Economic Development and Cultural Change*, 69, 1379–1410.

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481 –1495.

Ayres, I. and C. Unkovic (2012): "Information escrows," *Mich. L. Rev.*, 111, 145.

Bac, M. (2009): "An economic rationale for firing whistleblowers," *European Journal of Law and Economics*, 27, 233–256.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020): "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*, 110, 1206–30.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Contr," *Review of Economic Studies*, 81, 608–650.

Blair, G., K. Imai, and Y.-Y. Zhou (2015): "Design and analysis of the randomized response technique," *Journal of the American Statistical Association*, 110, 1304–1319.

Borker, G. (2018): "Safety First: Perceived Risk of Street Harassment and Educational Choices of Women," Tech. rep., mimeo.

Boudreau, L. (2022): "Multinational enforcement of labor law: Experimental evidence on strengthening occupational safety and health (OSH) committees," Tech. rep., mimeo.

Boudreau, L., R. Heath, and T. H. McCormick (2022): "Migrants, Experience, and Working Conditions in Bangladeshi Garment Factories," Tech. rep., mimeo.

Carril, A. (2017): "Dealing with misfits in random treatment assignment," *Stata Journal*, 17, 652–667.

Chakraborty, T., A. Mukherjee, S. R. Rachapalli, and S. Saha (2018): "Stigma of sexual violence and women's decision to work," *World Development*, 103, 226–238.

Chassang, S. and G. Padró i Miquel (2018): "Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports," *Review of Economic Studies*, 86, 2530–2553.

Chassang, S. and C. Zehnder (2019): "Secure Survey Design in Organizations: Theory and Experiments," .

Cheng, I.-H. and A. Hsiaw (2020): "Reporting Sexual Misconduct in the MeToo Era," Tech. rep., mimeo.

Chuang, E., P. Dupas, E. Huillery, and J. Seban (2020): "Sex, Lies, and Measurement: Do Indirect Response survey methods work?" .

Cowles, K. V. (1988): "Issues in qualitative research on sensitive topics," *Western Journal of Nursing Research*, 10, 163–179.

Dahl, G. B. and M. Knepper (2021): "Why is Workplace Sexual Harassment Underreported? The Value of Outside Options Amid the Threat of Retaliation," Tech. rep., mimeo.

Davison, A. C. and D. V. Hinkley (1997): *Bootstrap methods and their application*, 1, Cambridge university press.

Efron, B. (1987): "Better bootstrap confidence intervals," *Journal of the American statistical Association*, 82, 171–185.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort (2003): "Collusion, delegation and supervision with soft information," *The Review of Economic Studies*, 70, 253–279.

Folke, O. and J. Rickne (2022): "Sexual Harassment and Gender Inequality in the Labor Market," *Quarterly Journal of Economics*, 1–50.

Greenberg, B. G., A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz (1969): "The unrelated question randomized response model: Theoretical framework," *Journal of the American Statistical Association*, 64, 520–539.

Hershkowitz, I., M. E. Lamb, and C. Katz (2014): "Allegation rates in forensic child abuse investigations: Comparing the revised and standard NICHD protocols." *Psychology, Public Policy, and Law*, 20, 336.

Heyes, A. and S. Kapur (2009): "An economic model of whistle-blower policy," *The Journal of Law, Economics, & Organization*, 25, 157–182.

Jayachandran, S. (2021): "Social Norms as a Barrier to Women's Employment in Developing Countries," *IMF Economic Review*, 69, 576–595.

Kabeer, N., L. Huq, and M. Sulaiman (2020): "Paradigm Shift or Business as Usual? Workers' Views on Multi-stakeholder Initiatives in Bangladesh," *Development and Change*, 0, 1–39.

Kondylis, F., A. Legovini, K. Vyborny, A. Zwager, and L. Andrade (2020): "Demand for "Safe Space": Avoiding Harassment and Stigma," Tech. rep., mimeo.

Laffont, J.-J. and D. Martimort (1997): "Collusion under asymmetric information," *Econometrica*, 65, 875–911.

——— (2000): "Mechanism design with collusion and correlation," *Econometrica*, 68, 309–342.

Macchiavello, R., A. Menzel, A. Rabbani, and C. Woodruff (2020): "Challenges of Change: An Experiment Promoting Women to Managerial Roles in the Bangladeshi Garment Sector," Tech. rep., mimeo.

Makowsky, M. D. and S. Wang (2018): "Embezzlement, whistleblowing, and organizational architecture: An experimental investigation," *Journal of Economic Behavior & Organization*, 147, 58–75.

Muraglia, S., A. Vasquez, and J. Reichert (2020): "Conducting research interviews on sensitive topics," *Illinois Criminal Justice Information Authority (ICJIA)*.

Ortner, J. and S. Chassang (2018): "Making corruption harder: Asymmetric information, collusion, and crime," *Journal of Political Economy*, 126, 2108–2133.

Poi, B. P. (2004): "From the help desk: Some bootstrapping techniques," *The Stata Journal*, 4, 312–328.

Prendergast, C. (2000): "Investigating Corruption," Tech. rep., Working Paper, World Bank Development Group.

Raghavarao, D. and W. Federer (1979): "Block Total Response as an Alternative to the Randomized Response Method in Surveys," *Journal of the Royal Statistical Society*, B, 40–45.

Rosenfeld, B., K. Imai, and J. N. Shapiro (2016): "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions," *American Journal of Political Science*, 60, 783–802.

Siddiqi, D. M. (2003): "The Sexual Harassment of Industrial Workers: Strategies for Intervention in the Workplace and Beyond," Tech. rep., Center for Policy Dialogue, Dhaka, Bangladesh.

Slepian, M. L. and E. Moulton-Tetlock (2019): "Confiding Secrets and Well-Being," *Social Psychology and Personality Science*, 10, 472–484.

Sumon, M. H., A. Borhan, and N. Shifa (2018): "Garment Workers' Rights: Situation analysis in Dhaka, Gazipur, Narayanganj, and Chittagong," Tech. rep., Manusher Jonno Foundation, Dhaka, Bangladesh.

Tirole, J. (1986): "Hierarchies and bureaucracies: On the role of collusion in organizations," *Journal of Law, Economics, & Organizations*, 2, 181–214.

Tourangeau, R. and T. Yan (2007): "Sensitive questions in surveys." *Psychological bulletin*, 133, 859.

United Nations Human Rights Office (2011): "Manual on human rights monitoring," *OHCHR UN Publications*.

United Nations Statistical Office (2014): "Guidelines for producing statistics on violence against women," *United Nations, Department of Economic and Social Affairs Statistics*.

Vallano, J. P. and N. S. Compo (2011): "A comfortable witness is a good witness: Rapport-building and susceptibility to misinformation in an investigative mock-crime interview," *Applied cognitive psychology*, 25, 960–970.

Warner, S. L. (1965): "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, 63–69.

Online Appendix

A Figures & Tables

Figure A.1: Survey Modules & Treatment Interventions

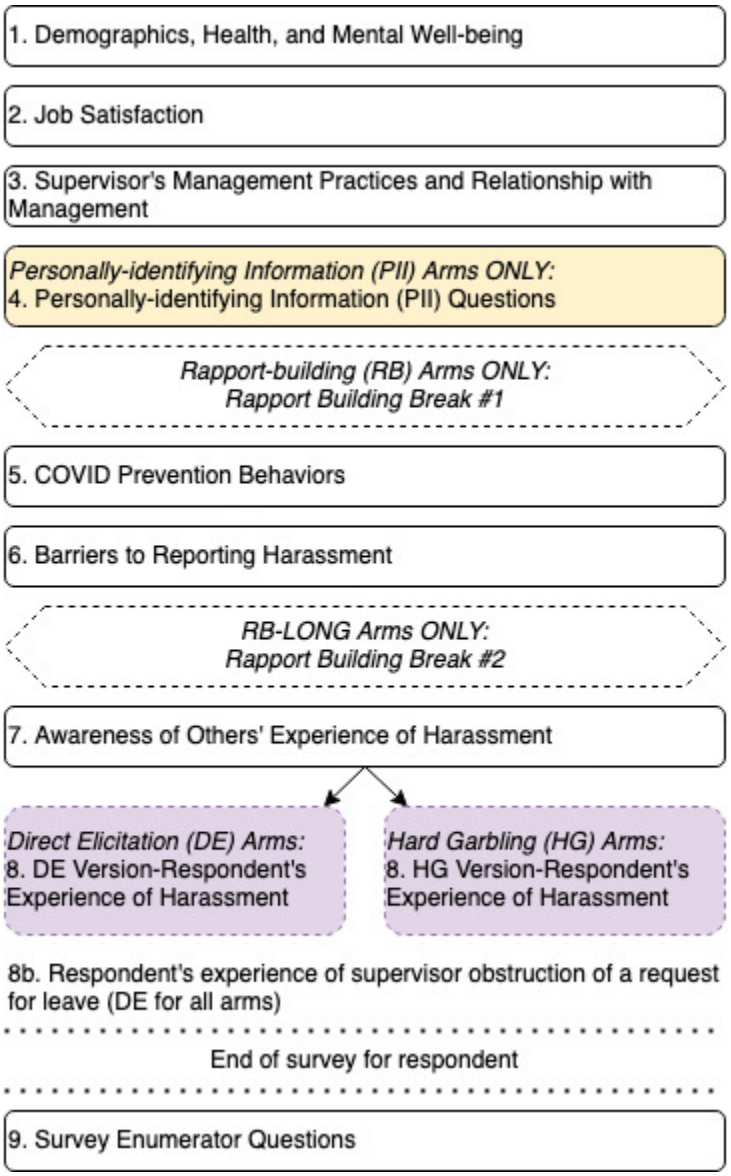
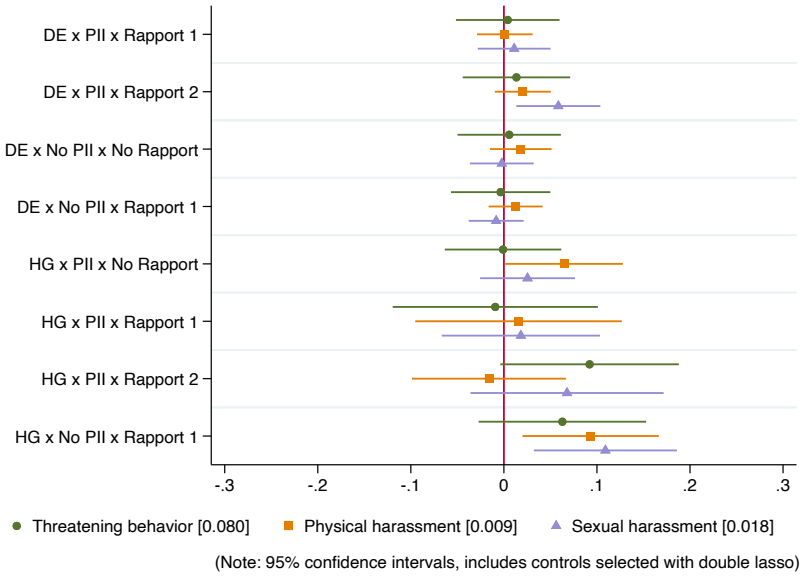
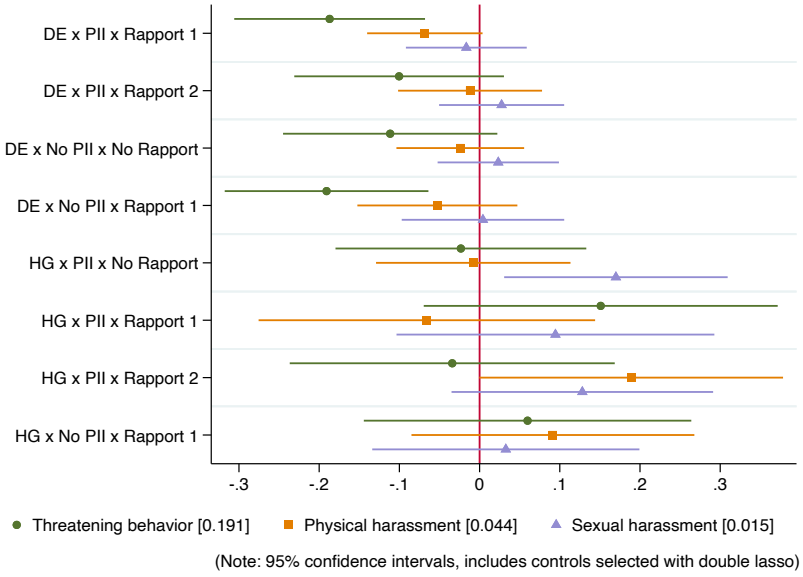


Figure A.2: Treatment effects by survey arm, separately by sex

(a) For Women

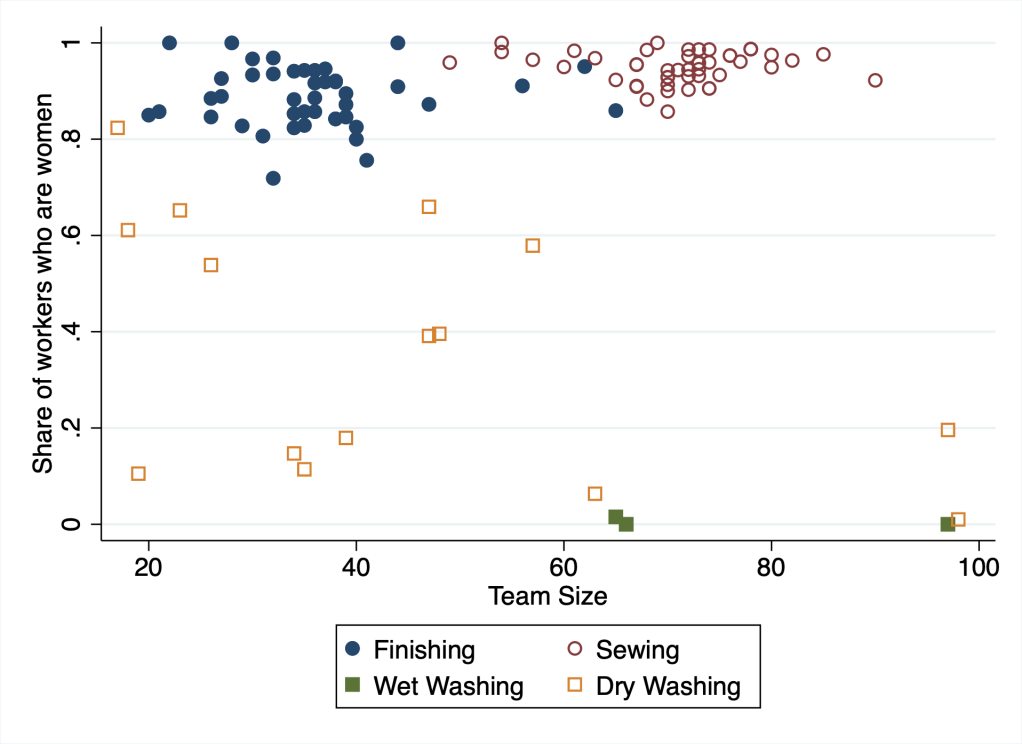


(b) For Men



Notes: This figure reports coefficients from separate regressions of the outcome variable on the treatment arm indicators, strata fixed effects, and controls selected using the PDS lasso, by sex. The regression specification is equation (8), conditional on each sex. The whiskers are 95% confidence intervals estimated using robust standard errors. The omitted category is treatment arm 1, $\mathbb{1}(\text{DE} \times \text{PII} \times \text{No RB})_i = 1$, which is the control condition. The number in square brackets is the reporting rate for each group, by sex.

Figure A.3: Team size & gender composition by production section



Notes: This figure plots the share of workers who are women against team size, with teams coded by production section. *Team Size* refers to the total number of workers on the production teams that surveyed individuals were sampled from (i.e., including workers who did *not* participate in our survey experiment). These numbers differ from the team size used in our analysis, which is based on the median number of team members included in HG/PII treatment arms. *Share of workers who are female* refers to the percent of female workers on the production teams that surveyed individuals were sampled from (i.e., including workers who did *not* participate in our survey experiment). The figure shows that there is little variation in the gender composition and size of teams within each production section (except for dry washing teams).

Table A.1: Team-level Summary Statistics (including all workers on team, including those not sampled)

	Mean	SD	Min	p25	p50	p75	Max	N
Panel A: Number of workers in a team								
Team Size: Overall	53.1	20.8	17	35	54	72	98	112
Team Size: Factory 1	54.9	23.1	19	32	55.5	74.5	98	60
Team Size: Factory 2	51	17.7	17	37	47.5	69	74	52
Team Size: Sewing Section	70.9	7.75	49	67.5	72	74.5	90	48
Team Size: Finishing Section	35.8	8.98	20	30	35.5	39	65	46
Team Size: Washing Section	49.8	27.0	17	26	47	65	98	18
Panel B: Share of workers in a team who are women								
Team's Female Share: Overall	0.82	0.26	0	0.84	0.92	0.96	1	112
Team's Female Share: Factory 1	0.85	0.26	0	0.88	0.94	0.97	1	60
Team's Female Share: Factory 2	0.79	0.25	0	0.81	0.88	0.93	1	52
Team's Female Share: Sewing Section	0.95	0.033	0.86	0.93	0.96	0.98	1	48
Team's Female Share: Finishing Section	0.89	0.062	0.72	0.85	0.89	0.93	1	46
Team's Female Share: Washing Section	0.30	0.28	0	0.063	0.19	0.58	0.82	18

Notes: This table provides summary statistics on the teams that surveyed workers are employed in. In Panel A, the *Number of workers in a team* refers to the total number of workers on the production teams from which we sampled workers from to participate in our survey. In other words, they are inclusive of workers who were randomly selected to be invited to participate and workers who were not randomly selected to be invited to participate in the survey. The median team size is larger than the team size in the *Understanding Harassment* analysis because the latter is the median number of team-members in the sample included in the treatment arms with HG and PII. In Panel B, the *Share of workers in a team who are women* refers to the share of workers who are women on the production teams from which we sampled workers from to participate in our survey.

Table A.2: Balance tests: main treatment conditions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Mean / (SD)				**Difference in means / [p-value]**		
Variable	DE	HG	No Rapport	Rapport	PII	Low PII	HG-DE	Diff Rapport	Diff PII
Female	0.811	0.816	0.815	0.812	0.815	0.809	0.007	0.005	-0.001
	(0.392)	(0.388)	(0.389)	(0.391)	(0.388)	(0.393)	[0.152]	[0.280]	[0.855]
Currently Working	0.957	0.961	0.955	0.962	0.960	0.957	0.003	0.006	-0.004
	(0.202)	(0.194)	(0.207)	(0.191)	(0.197)	(0.203)	[0.745]	[0.524]	[0.661]
Age	26.686	26.881	26.672	26.870	26.818	26.686	0.194	0.104	-0.117
	(5.042)	(5.254)	(5.060)	(5.214)	(5.210)	(4.982)	[0.371]	[0.635]	[0.616]
Experience (yrs)	5.173	5.204	5.133	5.234	5.192	5.178	-0.015	0.063	0.007
	(3.633)	(3.510)	(3.536)	(3.607)	(3.591)	(3.536)	[0.920]	[0.669]	[0.964]
Tenure (yrs)	2.880	2.900	2.868	2.907	2.900	2.864	0.033	-0.068	-0.033
	(2.431)	(2.429)	(2.431)	(2.429)	(2.420)	(2.454)	[0.704]	[0.429]	[0.732]
Years of Education	6.761	6.640	6.697	6.708	6.725	6.650	-0.097	0.047	-0.103
	(3.403)	(3.386)	(3.386)	(3.403)	(3.362)	(3.473)	[0.491]	[0.737]	[0.504]
Marital Status (1=Yes)	0.835	0.811	0.825	0.822	0.821	0.830	-0.026	-0.008	0.007
	(0.371)	(0.392)	(0.380)	(0.382)	(0.383)	(0.376)	[0.114]	[0.643]	[0.691]
Children (1=Yes)	0.738	0.744	0.743	0.739	0.740	0.744	0.004	-0.008	0.007
	(0.440)	(0.436)	(0.437)	(0.439)	(0.439)	(0.437)	[0.810]	[0.681]	[0.724]
Observations	1,122	1,021	978	1,165	1,515	628	2,143	2,143	2,143
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table summarizes workers' characteristics in each treatment condition. Columns (1)-(6) report the means and standard deviations of each variable separately by treatment condition. In column (4), Rapport pools the short and long rapport conditions. Columns (7)-(9) report the differences in means between each treatment condition, estimated from a regression of the covariate on the treatment indicator and stratification variables. Robust standard errors are reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.3: Balance tests: no rapport, short rapport, and long rapport arms

	Mean / (SD)			**Difference in means / [p-value]**		
Variable	No Rapport	Short Rapport	Long Rapport	(1) - (0)	(2) - (0)	(2) - (1)
	(0)	(1)	(2)			
Female	0.815	0.820	0.795	0.006	0.003	-0.005
	(0.389)	(0.385)	(0.404)	[0.253]	[0.635]	[0.409]
Currently Working	0.955	0.965	0.956	0.008	-0.000	-0.009
	(0.207)	(0.184)	(0.205)	[0.403]	[0.991]	[0.488]
Age	26.672	26.860	26.891	0.120	0.095	-0.029
	(5.060)	(5.124)	(5.411)	[0.614]	[0.767]	[0.930]
Experience (yrs)	5.133	5.323	5.040	0.163	-0.172	-0.341
	(3.536)	(3.589)	(3.644)	[0.311]	[0.421]	[0.121]
Tenure (yrs)	2.868	2.932	2.854	-0.020	-0.184	-0.115
	(2.431)	(2.419)	(2.452)	[0.832]	[0.127]	[0.354]
Years of Education	6.697	6.683	6.762	0.028	0.113	0.069
	(3.386)	(3.430)	(3.348)	[0.854]	[0.576]	[0.745]
Marital Status (1=Yes)	0.825	0.825	0.817	-0.006	-0.011	-0.007
	(0.380)	(0.380)	(0.387)	[0.759]	[0.642]	[0.787]
Children (1=Yes)	0.743	0.746	0.724	0.001	-0.024	-0.023
	(0.437)	(0.436)	(0.448)	[0.965]	[0.367]	[0.405]
Observations	978	799	366	1,777	1,344	1,165
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table summarizes workers' characteristics in each rapport building treatment condition. Columns (0)-(2) report the means and standard deviations of each variable separately by treatment condition. The next three columns report the differences in means between each treatment condition, estimated from a regression of the covariate on the treatment indicator and stratification variables. Robust standard errors are reported. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.4: Effects of Survey Design on Reporting, Differentiating Rapport Treatments

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
HG Treatment	0.0457***	0.0460***	0.0442***	0.0455***	0.0494***	0.0504***
	(0.0152)	(0.0147)	(0.0121)	(0.0117)	(0.0114)	(0.0110)
Low PII Treatment	0.0186	0.0182	0.0308	0.0325	0.0152	0.0185
	(0.0274)	(0.0267)	(0.0207)	(0.0198)	(0.0199)	(0.0197)
Rapport Treatment (Short)	0.0017	0.0041	-0.0126	-0.0112	0.0064	0.0050
	(0.0225)	(0.0218)	(0.0240)	(0.0231)	(0.0195)	(0.0188)
Rapport Treatment (Long)	0.0270	0.0299	-0.0042	-0.0033	0.0389	0.0406
	(0.0312)	(0.0305)	(0.0271)	(0.0263)	(0.0305)	(0.0293)
Control Group Mean	.0992	.0992	.0153	.0153	.0178	.0178
p(Long − Short Rapport)	[0.460]	[0.443]	[0.793]	[0.800]	[0.326]	[0.263]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes	No	Yes
Observations	2140	2140	2140	2140	2140	2140

Notes: This table reports OLS estimates of treatment effects on workers' reporting, separately estimating the effects of the short- and long-rapport building conditions. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Table A.5: Effects of Survey Design on Reporting of Harassment, Recorded HG Responses (Full Interactions)

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
DE × PII × Rapport 1	-0.0307	-0.0261	-0.0067	-0.0058	0.0156	0.0101
	(0.0254)	(0.0254)	(0.0126)	(0.0134)	(0.0166)	(0.0171)
DE × PII × Rapport 2	-0.0150	-0.0135	0.0092	0.0101	0.0410**	0.0456**
	(0.0272)	(0.0266)	(0.0142)	(0.0145)	(0.0204)	(0.0197)
DE × Low PII × No Rapport	-0.0162	-0.0197	0.0123	0.0092	0.0035	0.0052
	(0.0264)	(0.0255)	(0.0151)	(0.0153)	(0.0150)	(0.0154)
DE × Low PII × Rapport 1	-0.0337	-0.0341	0.0081	0.0046	-0.0022	-0.0063
	(0.0265)	(0.0254)	(0.0146)	(0.0143)	(0.0138)	(0.0142)
HG × PII × No Rapport	-0.0031	-0.0032	0.0488*	0.0476*	0.0450*	0.0431*
	(0.0301)	(0.0291)	(0.0284)	(0.0274)	(0.0248)	(0.0241)
HG × PII × Rapport 1	0.0183	0.0146	0.0045	0.0009	0.0301	0.0310
	(0.0516)	(0.0499)	(0.0515)	(0.0493)	(0.0403)	(0.0391)
HG × PII × Rapport 2	0.0608	0.0639	0.0242	0.0226	0.0784*	0.0758*
	(0.0459)	(0.0444)	(0.0402)	(0.0388)	(0.0474)	(0.0453)
HG × Low PII × Rapport 1	0.0650	0.0699	0.0793**	0.0873***	0.0870**	0.0919***
	(0.0441)	(0.0433)	(0.0347)	(0.0336)	(0.0354)	(0.0342)
Control Group Mean	.0992	.0992	.0153	.0153	.0178	.0178
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Lasso	No	Lasso	No	Lasso
Observations	2140	2140	2140	2140	2140	2140

Notes: This table reports OLS estimates of treatment effects on workers' reporting, separately estimating the effects of each treatment arm (i.e., the effects of the fully interacted treatment conditions). Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the full interactions of treatment variables and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

See Table A.6 on next page for p-values.

Table A.6: Effects of Survey Design on Reporting of Harassment, Recorded HG Responses (Full Interactions, p-values of differences between coefficients)

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
DExPIIxRB1 − DExPIIxRB2	[0.600]	[0.674]	[0.318]	[0.340]	[0.284]	[0.133]
DExPIIxRB1 − DExNoPIIxNoRB	[0.620]	[0.823]	[0.243]	[0.372]	[0.524]	[0.802]
DExPIIxRB1 − DExNoPIIxRB1	[0.921]	[0.780]	[0.354]	[0.520]	[0.325]	[0.383]
DExPIIxRB1 − HGxPIIxNoRB	[0.398]	[0.468]	[0.054]	[0.051]	[0.290]	[0.219]
DExPIIxRB1 − HGxPIIxRB1	[0.358]	[0.434]	[0.828]	[0.896]	[0.731]	[0.610]
DExPIIxRB1 − HGxPIIxRB2	[0.057]	[0.055]	[0.451]	[0.481]	[0.196]	[0.164]
DExPIIxRB1 − HGxNoPIIxRB1	[0.037]	[0.033]	[0.016]	[0.007]	[0.056]	[0.024]
DExPIIxRB2 − DExNoPIIxNoRB	[0.968]	[0.838]	[0.861]	[0.961]	[0.092]	[0.064]
DExPIIxRB2 − DExNoPIIxRB1	[0.551]	[0.496]	[0.951]	[0.755]	[0.050]	[0.018]
DExPIIxRB2 − HGxPIIxNoRB	[0.732]	[0.761]	[0.185]	[0.194]	[0.895]	[0.932]
DExPIIxRB2 − HGxPIIxRB1	[0.538]	[0.593]	[0.929]	[0.854]	[0.805]	[0.728]
DExPIIxRB2 − HGxPIIxRB2	[0.119]	[0.095]	[0.719]	[0.757]	[0.455]	[0.528]
DExPIIxRB2 − HGxNoPIIxRB1	[0.087]	[0.069]	[0.051]	[0.027]	[0.240]	[0.218]
DExNoPIIxNoRB − DExNoPIIxRB1	[0.558]	[0.612]	[0.811]	[0.795]	[0.731]	[0.502]
DExNoPIIxNoRB − HGxPIIxNoRB	[0.696]	[0.612]	[0.223]	[0.183]	[0.121]	[0.149]
DExNoPIIxNoRB − HGxPIIxRB1	[0.518]	[0.506]	[0.882]	[0.870]	[0.523]	[0.523]
DExNoPIIxNoRB − HGxPIIxRB2	[0.110]	[0.072]	[0.776]	[0.742]	[0.123]	[0.134]
DExNoPIIxNoRB − HGxNoPIIxRB1	[0.080]	[0.047]	[0.066]	[0.027]	[0.022]	[0.013]
DExNoPIIxRB1 − HGxPIIxNoRB	[0.362]	[0.337]	[0.173]	[0.133]	[0.075]	[0.057]
DExNoPIIxRB1 − HGxPIIxRB1	[0.334]	[0.346]	[0.946]	[0.941]	[0.432]	[0.355]
DExNoPIIxRB1 − HGxPIIxRB2	[0.049]	[0.035]	[0.696]	[0.653]	[0.092]	[0.074]
DExNoPIIxRB1 − HGxNoPIIxRB1	[0.035]	[0.022]	[0.049]	[0.018]	[0.013]	[0.005]
HGxPIIxNoRB − HGxPIIxRB1	[0.721]	[0.759]	[0.506]	[0.465]	[0.770]	[0.803]
HGxPIIxNoRB − HGxPIIxRB2	[0.250]	[0.210]	[0.640]	[0.624]	[0.563]	[0.555]
HGxPIIxNoRB − HGxNoPIIxRB1	[0.192]	[0.149]	[0.531]	[0.394]	[0.361]	[0.272]
HGxPIIxRB1 − HGxPIIxRB2	[0.547]	[0.469]	[0.780]	[0.749]	[0.452]	[0.466]
HGxPIIxRB1 − HGxNoPIIxRB1	[0.540]	[0.458]	[0.239]	[0.153]	[0.318]	[0.269]
HGxPIIxRB2 − HGxNoPIIxRB1	[0.945]	[0.921]	[0.308]	[0.217]	[0.896]	[0.798]
Complementarity Test	[0.044]	[0.035]	[0.323]	[0.247]	[0.330]	[0.256]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS Lasso Controls	No	Lasso	No	Lasso	No	Lasso

Notes: This table reports p-values of the difference between fully interacted treatment conditions from the OLS regression of treatment effects on workers' reporting. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the full interactions of treatment variables and stratification variables. Even-numbered columns also include controls selected using the PDS lasso.
Complementarity Test: HGxNoPIIxRB1 $\leq$ DExPIIxRB1 + DExNoPIIxNoRB + HGxPIIxNoRB. We test for complementarity because for all outcomes, the point estimate for HGxNoPIIxRB1 is greater than the sum of the point estimates for the other three arms.

Table A.7: Share of teams with k recorded "Yes" responses

k	Threatening	Physical	Sexual
1	0.268	0.411	0.411
2	0.402	0.438	0.366
3	0.277	0.143	0.143
4	0.045	0.009	0.071
5	0.009	0.000	0.009
6	0.000	0.000	0.000
7	0.000	0.000	0.000

Notes: This table reports the share of teams (for workers in HG & PII arms) with $k \in \{1, \cdots, 7\}$ recorded "Yes" responses. The shares are computed directly from the data. The sample includes 112 teams.

Table A.8: Effects of Survey Design on Survey Duration

	Rapport Treatment (Pooled)		Rapport Treatment	
	(1)	(2)	(3)	(4)
HG Treatment	1.6361***	1.5976***	1.7084***	1.6724***
	(0.5328)	(0.5104)	(0.5343)	(0.5115)
Low PII Treatment	-1.7307***	-1.7467***	-1.1749*	-1.1623*
	(0.5870)	(0.5638)	(0.6421)	(0.6132)
Rapport Treatment (Pooled)	6.1307***	6.1805***		
	(0.5402)	(0.5198)		
Rapport Treatment (Short)			5.4945***	5.5072***
			(0.6197)	(0.5946)
Rapport Treatment (Long)			7.1710***	7.2754***
			(0.7865)	(0.7623)
Control Group Mean	42.1471	42.1471	42.1471	42.1471
p(Long − Short Rapport)			[0.056]	[0.038]
Strata FE	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes
Observations	2100	2100	2100	2100

Notes: This table reports OLS estimates of treatment effects on survey duration (in minutes) which is trimmed below and above at 1 and 99 percentiles respectively. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Robust standard errors are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Table A.9: Main treatment effects, estimated with *recorded* response = "no" for confused respondents

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Main effects						
HG Treatment	0.0351**	0.0357**	0.0372***	0.0386***	0.0360***	0.0372***
	(0.0149)	(0.0144)	(0.0117)	(0.0113)	(0.0111)	(0.0108)
Low PII Treatment	0.0120	0.0113	0.0256	0.0272	-0.0014	0.0009
	(0.0242)	(0.0236)	(0.0185)	(0.0178)	(0.0201)	(0.0198)
Rapport Treatment	0.0111	0.0142	-0.0094	-0.0082	0.0192	0.0193
	(0.0200)	(0.0193)	(0.0198)	(0.0190)	(0.0184)	(0.0177)
Control Group Mean	.0992	.0992	.0153	.0153	.0178	.0178
Panel B: Heterogeneity by sex						
HG Treatment × Female	0.0184	0.0191	0.0341***	0.0357***	0.0275**	0.0298**
	(0.0168)	(0.0162)	(0.0127)	(0.0122)	(0.0133)	(0.0130)
HG Treatment × Male	0.1110***	0.1079***	0.0526	0.0516	0.0701**	0.0701**
	(0.0429)	(0.0419)	(0.0339)	(0.0329)	(0.0334)	(0.0329)
Low PII Treatment × Female	0.0127	0.0132	0.0320	0.0335*	0.0043	0.0064
	(0.0260)	(0.0255)	(0.0209)	(0.0201)	(0.0225)	(0.0220)
Low PII Treatment × Male	0.0039	-0.0014	0.0010	0.0027	-0.0304	-0.0284
	(0.0548)	(0.0532)	(0.0462)	(0.0458)	(0.0384)	(0.0376)
Rapport × Female	0.0217	0.0247	-0.0165	-0.0144	0.0312	0.0315
	(0.0223)	(0.0214)	(0.0232)	(0.0223)	(0.0204)	(0.0194)
Rapport × Male	-0.0341	-0.0342	0.0202	0.0161	-0.0386	-0.0368
	(0.0472)	(0.0454)	(0.0373)	(0.0362)	(0.0457)	(0.0453)
Female	-0.0896	-0.1032	-0.0265	-0.0149	0.0680	0.0838
	(0.1060)	(0.1023)	(0.0754)	(0.0753)	(0.0746)	(0.0725)
Control Mean - Female	.08	.08	.0092	.0092	.0185	.0185
Control Mean - Male	.1912	.1912	.0441	.0441	.0147	.0147
p(HGxFemale - HGxMale)	[0.055]	[0.058]	[0.617]	[0.658]	[0.274]	[0.293]
p(NoPIIxFemale - NoPIIxMale)	[0.880]	[0.797]	[0.552]	[0.550]	[0.421]	[0.402]
p(RapportxFemale - RapportxMale)	[0.292]	[0.244]	[0.412]	[0.484]	[0.168]	[0.170]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Lasso	No	Lasso	No	Lasso
Observations	2140	2140	2140	2140	2140	2140

Notes: This table reports OLS estimates of treatment effects on workers' reporting (also heterogeneity by sex). Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the gender interactions of treatment variables and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Table A.10: Do respondents in HG understand that the mechanism has changed? Direct question on supervisor obstruction of leave request

	Supervisor Obstruction of Leave Request			
	(1)	(2)	(3)	(4)
HG Treatment	0.0070	0.0067	0.0058	0.0014
	(0.0156)	(0.0150)	(0.0204)	(0.0197)
Rapport Treatment	0.0093	0.0082	0.0274	0.0288
	(0.0160)	(0.0155)	(0.0210)	(0.0201)
Low PII Treatment	0.0151	0.0153	0.0222	0.0218
	(0.0178)	(0.0172)	(0.0229)	(0.0221)
Control Group Mean	.1298	.1298	.2723	.2723
Variable definition	1	1	2	2
Strata FE	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes
Observations	2143	2143	2143	2143

Notes: This table reports OLS estimates of treatment effects on workers' response to the question, *In the past year, has your line supervisor refused to forward your request for leave to your factory's administration or required you to perform certain tasks in order for him to submit it?* This question is asked in the module on the respondent's experience of harassment immediately after the three harassment questions that use either the DE or HG methods. This question uses a DE method for *all* respondents. In *Variable definition 1*, respondents who report that they did not request leave in the previous year are coded as no. In *Variable definition 2*, respondents who report that they did not request leave in the previous year are coded as yes. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator and stratification variables. Even-numbered columns also include controls selected using the PDS lasso. Robust standard errors are reported in round brackets. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.11: Survey enumerators' assessment of respondent after survey completion

	Comprehension		Comfort		Trust no leakage		Honesty		Patience	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
HG Treatment	-0.0303	-0.0262	-0.0001	0.0012	-0.0199	-0.0180	0.0479*	0.0467*	-0.0191	-0.0172
	(0.0358)	(0.0337)	(0.0396)	(0.0376)	(0.0374)	(0.0358)	(0.0283)	(0.0271)	(0.0406)	(0.0389)
Low PII Treatment	0.0086	0.0097	0.0791*	0.0835**	0.0366	0.0406	-0.0074	-0.0052	-0.0700	-0.0718*
	(0.0404)	(0.0381)	(0.0430)	(0.0412)	(0.0415)	(0.0397)	(0.0317)	(0.0303)	(0.0445)	(0.0425)
Rapport Treatment	-0.0031	-0.0021	-0.0170	-0.0197	-0.0125	-0.0139	0.0609**	0.0618**	0.0981**	0.1016**
	(0.0367)	(0.0348)	(0.0402)	(0.0383)	(0.0370)	(0.0356)	(0.0287)	(0.0275)	(0.0415)	(0.0397)
Control Group Mean	0	0	0	0	0	0	0	0	0	0
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Enumerator FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2143	2143	2143	2143	2143	2143	2143	2143	2143	2143

Notes: This table reports OLS estimates of survey enumerators' assessment of respondents' behavior during the survey. All outcomes are standardised using the control group's mean and standard deviation, with higher values corresponding to more positive outcomes. *Comprehension*: Enumerator's assessment of how well the respondent understood the questions, *Comfort*: Enumerator's assessment of how comfortable the respondent felt answering the questions, *Trust*: Enumerator's assessment on whether the respondent trusts that the research team to not share their responses. *Honesty*: Enumerator's assessment of whether the respondent answered honestly to personal and sensitive questions, *Patience*: Enumerator's assessment of whether the respondent was rushing to finish the survey. Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator, stratification variables, and enumerator fixed effects. Even-numbered columns also include controls selected using the PDS lasso. Robust standard errors are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Table A.12: HTEs by respondents' schooling qualification for supervisor position

	Threatening behavior	Physical harassment	Sexual harassment
	(1)	(2)	(3)
HG Treatment × Female × Min Grade 8	0.0223	0.0430	0.0993***
	(0.0336)	(0.0278)	(0.0291)
HG Treatment × Female × Below Grade 8	0.0321	0.0361	-0.0056
	(0.0281)	(0.0246)	(0.0244)
HG Treatment × Male × Min Grade 8	0.0968*	0.1035*	0.0555
	(0.0573)	(0.0595)	(0.0500)
HG Treatment × Male × Below Grade 8	0.1429**	0.0224	0.1230**
	(0.0604)	(0.0550)	(0.0497)
Rapport Treatment	0.0122	-0.0093	0.0177
	(0.0203)	(0.0200)	(0.0183)
Low PII Treatment	0.0098	0.0275	0.0059
	(0.0245)	(0.0186)	(0.0203)
Control Mean-Female & Above	.0725	.0072	.0145
Control Mean-Female & Below	.0856	.0107	.0214
Control Mean-Male & Above	.2222	.0278	.0278
Control Mean-Male & Below	.1562	.0625	0
p(HGXFemaleXHigh-HGXFemaleXLow)	[0.849]	[0.880]	[0.024]
p(HGXMaleXHigh-HGXMaleXLow)	[0.582]	[0.384]	[0.343]
Strata FE	Yes	Yes	Yes
Observations	2140	2140	2140

Notes: This table reports OLS estimates of heterogeneity in treatment effects on workers' reporting by sex and by whether the respondent has at least 8 years of schooling, an informal cutoff used by garments factories to determine workers' eligibility to become a supervisor. The main effects of sex and schooling are included but not displayed. Rapport pools the short and long rapport conditions. Standard errors clustered by HG batch (HG respondents) or respondent (DE respondents) are reported in round brackets. *p <0.1; **p <0.05; ***p <0.01.

Table A.13: Correlation of team-level reporting rates and response rate to the survey

Correlations	DE	HG	HG-DE
ρ(Threat, Survey Response Rate)	-0.126	-0.139	-0.040
	(0.093)	(0.084)	(0.090)
	[-0.318,0.039]	[-0.294,0.041]	[-0.200,0.152]
ρ(Physical, Survey Response Rate)	-0.096	0.010	0.046
	(0.064)	(0.093)	(0.090)
	[-0.222,0.017]	[-0.181,0.198]	[-0.141,0.220]
ρ(Sexual, Survey Response Rate)	0.070	-0.050	-0.074
	(0.107)	(0.092)	(0.093)
	[-0.124,0.303]	[-0.223,0.137]	[-0.246,0.118]

Notes: This table reports the correlation between the team-level response rate to the survey and the team-level reporting rates of harassment using arms that collect PII. Standard errors (in parenthesis) are computed from 1000 bootstrap replications, drawing samples of reporting rates at the team-level. Confidence intervals [in brackets] are bias corrected and accelerated (BCa) (Efron, 1987, Davison and Hinkley, 1997), implemented using Stata package **bootstrap** (Poi, 2004).

B Reporting harassment & mental health

Sociological research suggests that the act of confiding secrets can improve an individual's well-being through improving one's perceived coping ability and reducing one's mental load associated with the secret (Slepian and Moulton-Tetlock, 2019). To explore this possibility, we resurveyed workers two weeks after the survey experiment to test whether reporting harassment improved workers' mental well-being and job satisfaction. We measure mental health and job satisfaction, respectively, using summary index variables following Anderson (2008). We report the variables comprising each index at the end of this appendix.

As per our PAP, we run a 2SLS model with (6) as our first stage and (B.1) as our second-stage regression:

$$W_{is} = \delta Y_{is} + \rho W_{is}^0 + \theta X_i + \mu_s + \epsilon_{is} \tag{B.1}$$

where W_{is} is worker well-being in the follow-up survey for individual i in stratum s, Y_{is} are reports of threats, physical and/or sexual harassment from the main worker survey, and W_{is}^0 is the baseline worker well-being, measured in the main worker survey. We control for stratum fixed-effects μ_s and individual demographic characteristics X_i. Since there is a possibility that some elements of the survey design directly impact worker well-being (notably, RB),

we also report the reduced form effect in a regression equivalent to (6), with W_{is} as outcome (and controlling for W_{is}^0).

Because we find no main effect of RB and Low PII on reporting, there is not a first stage between these two instruments and reporting. In other words, these are weak instruments, which will bias our 2SLS results towards the OLS results we would get if regressing mental health on reporting.[35] While we pre-specified that we would use (6) as our first stage, because of the weak instruments concern, we also report results only using randomized assignment to HG as the instrument and controlling for assignment to the RB and Low PII arms.

Table B.1 reports the reduced form and 2SLS effects on mental health and job satisfaction, using randomized assignment to HG, RB, and Low PII as instruments, measured in the follow-up survey. Columns (1)-(2) show that the treatments do not directly effect mental health or job satisfaction. Columns (3)-(5) show that reporting harassment improves mental health among those induced to report by the treatment interventions by 8-16% of a standard deviation, although none of the increases is statistically significant. Columns (7)-(10) show that reporting harassment improves job satisfaction among those induced to report by the treatment interventions by 37-68% of a standard deviation on average, although none of the increases is statistically significant.

Table B.2 reports the reduced form and 2SLS effects on mental health and job satisfaction, using randomized assignment to HG as the instrument and including controls for assignment to the RB and Low PII treatment arms. As expected, the estimated coefficients are uniformly more positive than in Table B.1, but they remain imprecise. Column (4) suggests that increasing the reported share of yeses from 0 to 1 improves mental health by 23% of a standard deviation among those induced to report under HG (p=0.291). Column (8) suggests that it improves job satisfaction by 86% of a standard deviation (p=0.170). While imprecise, the large, consistently positive coefficients suggests that the psychological and/or expected social benefits of reporting may be large. We think that more precisely quantifying these benefits, and more broadly exploring the benefits and costs of improved reporting systems for harassment for workers, presents an interesting direction for future research.

[35]The coefficients from the OLS regressions of mental health on reporting are zero or weakly negative (results not reported).

Table B.1: Reduced form & 2SLS effects on mental health & job satisfaction, measured in follow-up survey

	Reduced form		2SLS							
	Mental health index	Job satisfaction index	Mental health index				Job satisfaction index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
HG Treatment	0.0110	0.0387								
	(0.0093)	(0.0250)								
Rapport Treatment	0.0018	0.0117								
	(0.0097)	(0.0258)								
Low PII Treatment	-0.0134	-0.0391								
	(0.0104)	(0.0282)								
Reported threatening behavior			0.1377				0.6976			
			(0.1799)				(0.5787)			
Reported physical harassment				0.0779				0.4131		
				(0.2077)				(0.6076)		
Reported sexual harassment					0.1620				0.7001	
					(0.1792)				(0.5313)	
Share of reports that are yes						0.1480				0.6904
						(0.1924)				(0.5631)
Control Mean	.044	.317	.044	.044	.044	.044	.317	.317	.317	.317
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1987	1987	1984	1984	1984	1984	1984	1984	1984	1984
Kleibergen-Paap Wald F			1.8	1.8	2.4	4.1	1.5	1.8	2.3	3.7

Notes: This table reports reduced form and 2SLS results for respondents' mental health and job satisfaction, measured in the follow-up survey. Columns (1)-(2) report reduced form results, and columns (3)-(10) report 2SLS results using the randomized assignment to the HG, RB, and Low PII treatments as the instrumental variables. All regressions include controls for the baseline value of the dependent variable, gender, age, production section, position type, work experience, tenure, schooling, marital status, and whether the respondent has children. Robust standard errors in round brackets. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table B.2: Reduced form & 2SLS effects on mental health & job satisfaction, measured in follow-up survey, only using HG as an instrument

	Mental health index				Job satisfaction index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reported threatening behavior	0.2308				0.9155			
	(0.2324)				(0.7459)			
Reported physical harassment		0.2625				0.9913		
		(0.2615)				(0.7959)		
Reported sexual harassment			0.2033				0.7694	
			(0.1968)				(0.5735)	
Share of reports that are yes				0.2308				0.8866
				(0.2148)				(0.6238)
Control Mean	.044	.044	.044	.044	.317	.317	.317	.317
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1984	1984	1984	1984	1984	1984	1984	1984
Kleibergen-Paap Wald F	4	4	6.3	10.7	3.5	3.8	6.1	9.9

Notes: This table reports reduced form and 2SLS results for respondents' mental health and job satisfaction, measured in the follow-up survey. All columns report 2SLS results using the randomized assignment to the HG treatment as the instrumental variable. All regressions include controls for the baseline value of the dependent variable, gender, age, production section, position type, work experience, tenure, schooling, marital status, whether the respondent has children, and assignment to the RB and Low PII arms. Robust standard errors in round brackets. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Survey questions used to construct index variables:

1. Mental health:

    - Generalized Anxiety Disorder 7-item (GAD-7) scale): In the past 7 days, how often... (Select one: Not at all or less than 1 day; 1-2 days; 3-4 days; 5-7 days.)
        a have you felt nervous, anxious, or on edge?
        b have you felt depressed?
        c have you felt lonely?
        d have you felt hopeful about the future?
        e have you been so restless that it is hard to sit still?
        f have you become easily annoyed or irritable?
        g have you felt afraid, as if something awful might happen?
    - Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. (Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.)
        a On which step of the ladder would you say you personally feel you stand at this time?
        b On which step do you think you will stand about five years from now?

2. Job satisfaction:

    - How satisfied are you with your job overall? (Select one: Very dissatisfied; Dissatisfied; Neutral; Satisfied; Very satisfied)
    - How satisfied are you with the following aspect of your job: (Select one: Very dissatisfied; Dissatisfied; Neutral; Satisfied; Very satisfied)
        a You are listed to?
        b You are treated with respect?
        c Career opportunities?
        d Job training and support?
        e Pay is fair for your job?

- Which of the following statements best describes your feelings about your job? In my job... (Select one: I only work as hard as I have to; I work hard, but not so that it interferes with the rest of my life; I make a point of doing the best work I can, even if it sometimes does interfere with the rest of my life; I don't know)
- *Main survey experiment only:* For the following statements, please state whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree..
  a For me this is the best of all possible organizations for which to work.
  b I find that my values and the organization's values are very similar.
  c I feel very little loyalty to this organization.
  d Often, I find it difficult to agree with this organization's policies on important matters relating to its employees.
  e I am proud to tell others that I am part of this organization.