Monitoring Harassment in Organizations

Laura Boudreau
Columbia University

Sylvain Chassang
Princeton University

Ada González-Torres
Ben Gurion University

Rachel Heath[*,†,‡]
University of Washington

July 23, 2026

[*]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, by the Israeli Science Foundation, as well as Princeton's Griswold Center for Economic Policy Studies, and Industrial Relations Section. We are grateful to Ferdausi Sumana, Raied Arman, and Krishna Kamepalli for their excellent research assistance.

[†]We are grateful to Nava Ashraf, Florian Englmaier, Yana Gallen, Jorge García Hombrados, Matthew S. Johnson, Emily Nix, Rocco Macchiavello, Andrea Prat, Benjamin Roth, and Jason Sockin for helpful comments. We are grateful to seminar participants at Aix-Marseille School of Economics, Bar-Ilan, Berkeley, Chicago-Booth, Columbia, CUNEF, Haifa, Hebrew University, LMU Munich, Reichman, Rochester, UBC, UCSD, USC, UW, Peking University, Paris School of Economics, Toulouse School of Economics, and the Weatherhead Scholars Program at Harvard, as well as participants at AFE at Chicago, AFEDev, the ASSA Meetings, the Barcelona Summer Forum, the BREAD & MIT Conference, BREAD South Asia, the CEPR/CESifo/Imo-ENT Conference, the Development Workshop at Universitat de Barcelona, the Econometric Society World Congress in South Korea, the Economics of Firms and Labor Conference, the Economics of Violence against Women Conference at Stanford, the Firms and Inclusive Growth workshop at Queens, 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, the Women in Applied Microeconomics Conference, and the World Bank conference on Female Labor Force Participation: Drivers and Policy Options for stimulating comments.

[‡]This research was pre-registered in the AEA RCT registry (AEARCTR-0007103), and received IRB approval from Ben-Gurion University (2001-1; 2021-02-17), Columbia University (IRB-AAAT2938; 2021-02-23), Princeton University (17050; 2024-09-03) and the University of Washington (STUDY00010219; 2020-12-22).

Abstract

We experimentally evaluate secure survey methods for monitoring harassment in organizations. We partner with a Bangladeshi garment manufacturer and survey its workers. "Hard garbling" responses to sensitive questions, i.e., *automatically* coding a random subset as complaints, increases reporting of physical harassment by 290%, sexual harassment by 271%, and threatening behavior by 45%, from rates of 1.5%, 1.8%, and 9.9% under direct elicitation. Rapport-building and removing team identifiers do not increase reporting. We show that garbled reports can be used to consistently estimate policy-relevant statistics of reported harassment. Comparing team-level statistics estimated using HG versus DE, we find that HG reveals substantially more, and more widespread, harassment: because willing reporters are scarce per manager, HG's benefit comes from inducing a single reporter rather than corroborating ones, causing DE to miss most problem managers. Two field replications support the external validity of our findings.

KEYWORDS: harassment, whistleblowing, plausible deniability, randomized response, garbling, secure survey design, gender, garments, Bangladesh

1 Introduction

Organizations' ability to take action against harassment is limited by their ability to elicit information from relevant parties. Both victims and witnesses face possible retaliation and reputational costs if they come forward. As a result, organizations cannot respond to individual complaints or assess the scope of the 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 show how such reporting data can be used to inform policy-relevant questions about the nature of harassment, including: How widespread is it? What share of managers are responsible for what share of the damage? How isolated are victims who report?

Building on principal-agent-monitor models with retaliation (Chassang and Padró i Miquel, 2019, Chassang and Zehnder, 2024), our theoretical framework predicts that

survey designs that (i) increase plausible deniability if reports leak, (ii) build trust in the enumerator, and (iii) reduce the perceived likelihood of leaks should raise reporting. Our field experiment combines three treatments along these lines: hard garbling (HG), which automatically records a random subset of "no" responses as "yes"; rapport building (RB), in which enumerators chat about family and hobbies following a pre-specified script; and Low personally identifying information (PII), which omits questions about the identity of workers' supervisors and production team. Each treatment carries a cost — HG introduces noise that limits the range of feasible organizational responses, RB requires careful planning and lengthens the survey, and Low PII precludes computation of manager-level statistics.

We conducted phone-based surveys of 2,244 workers at two of the apparel producer's plants.[1] We randomly assigned survey respondents to nine different combinations of 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.

In the control group, only 9.9% of respondents report threatening behavior, 1.5% physical harassment, and 1.8% sexual harassment. HG raises these reporting rates by 45%, 290%, and 271%, respectively.[2] Low PII and RB had positive but weak effects. There is suggestive evidence of complementarity between treatment arms: combining the three treatment conditions increases reporting compared to the sum of individual effects. We conduct a pre-specified heterogeneous treatment effect (HTE) analysis by respondents' sex and find that HG increases reporting for both sexes.

Next, we contrast what the apparel producer learns about the scope and nature of

[1]The response rate was 63%. 73% of non-response was due to our inability to reach workers by phone.

[2]Respondents were assured of a strong commitment to confidentiality as part of the IRB-required informed consent process (included in the Supplementary Materials). Even so, control respondents' reporting rates for sensitive harassment are very low.

harassment in its organization under the DE and HG systems.[3] To do so for HG, we extend results from the literature on inference from garbled reports (Warner, 1965, Blair, Imai and Zhou, 2015) to estimate *team-level* statistics of reported harassment.[4] Under DE, 44.6% of teams have at least one worker who reports being threatened, 9.8% have at least one who reports being physically harassed, and 18.8% have at least one who reports being sexually harassed. Under HG, the problem is revealed to be much more widespread: the rates are 71.9%, 25.4%, and 40.2%, respectively. This large gap arises because a typical problem manager has only one worker willing to report at all, so raising each worker's own reporting probability – rather than relying on getting at least one of multiple victims on a team to independently come forward – is what reveals the manager's team. Evidently, a policy of investigating and firing all misbehaving managers is unlikely to be feasible. Instead, the firm needs to take other remedial actions.

Conditional on a type of harassment, workers who do report generally do so alone, and more so for graver forms. Under HG (DE), the probability that a team contains at least two victims willing to report, conditional on having at least one, is 16% (28%) for threatening behavior, 14% (9%) for physical harassment, and 8% (14%) for sexual harassment. We refer to this pattern as *reporting isolation* to distinguish it from true social isolation of victims: a low count of reports per team can reflect either victims being alone in being targeted or correlated under-reporting within the team, and our data cannot separate the two. Either way, requiring multiple reports will miss the majority of cases, so organizations need processes that can be initiated by a single report to address harassment.

Two field replications let us probe external validity along several dimensions: location (other garment factories), moment in time, treatment design (whether plausible deniability is essential or reassurances alone suffice), and, more tentatively,

[3] We provide testable conditions under which HG reduces the bias between relevant statistics of intended reports and statistics of actual harassment. They hold in our sample.

[4] On average, 7 workers per team are sampled in arms that use HG or DE, respectively, and collect PII. We use this sample in the estimation.

short- versus longer-run (by eliciting preferences over mechanisms for ongoing internal monitoring). Quantitative estimates vary across sites, but the replication data support our findings, while also identifying heterogeneity in treatment effects for men and women (in replication 2) that we cannot fully pin down the mechanism for.

This paper contributes to an emerging economic literature on workplace harassment, in particular sexual harassment, and its implications for labor markets (Cheng and Hsiaw, 2022, Dahl and Knepper, 2026, Adams-Prassl et al., 2024, Folke and Rickne, 2022). We contribute evidence that plausible deniability causally improves the 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).[5] There is a dearth of evidence, however, on the effects of sexual harassment in the workplace on workers' labor supply and well-being.[6] We contribute evidence that in the context of garment production in Bangladesh – a low-skill manufacturing sector that is common to many developing countries – harassment against women by managers who are men is common, and harassment by men against subordinate men is also substantial.

This research also contributes to the literature on detecting and deterring misbehavior in organizations. An important body of research in contract theory studies the possibility of side-contracting in principal-agent-monitor settings, with monitors and agents conspiring to limit information transmission (Tirole, 1986, Laffont and

[5]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, Gutiérrez and Villagrán (2021), Kondylis et al. (2025), Borker (2018)).

[6]The poor working conditions (Boudreau, Heath and McCormick, 2024) and extreme gender imbalances between managers and workers (Macchiavello et al., 2020) documented in Bangladesh's garment sector suggest that garment workers have limited outside options, enabling harassment to run rampant.

Martimort, 1997, 2000, Prendergast, 2000, Faure-Grimaud, Laffont and Martimort, 2003, Ortner and Chassang, 2018) or with misbehaving agents disciplining potential informants through threats of punishment (Heyes and Kapur, 2009, Bac, 2009, Makowsky and Wang, 2018). Chassang and Padró i Miquel (2019) show that garbled intervention policies such as HG are needed to discipline agents who can commit to retaliation. We bring HG into a real-world organizational setting and find large effects on information transmission, suggesting that this class of mechanisms deserves further exploration where credible threats or reputational costs suppress reporting.

Finally, this research contributes to a literature on garbled survey designs and on inference from garbled surveys dating back to Warner (1965) on randomized response (RR). A common feature of indirect survey methods – including RR, list experiments (LEs, Raghavarao and Federer (1979)), and crosswise RR methods (Blair, Imai and Zhou, 2015) – is that respondents control the report being sent, making them subject to unraveling in equilibrium (Chassang and Zehnder, 2024) and to non-compliance in practice (Chuang et al., 2021).[7] We provide real-world evidence that HG outperforms DE on policy-relevant statistics and that workers prefer HG for internal monitoring.

The remainder of the paper provides background on Bangladesh's garments sector and the partner firm (Section 2), presents the conceptual framework (Section 3), the research design (Section 4), treatment effects on reporting (Section 5), the partner firm's policy learnings under DE versus HG (Section 6), evidence on external validity (Section 7), and concluding remarks (Section 8).

[7]The literature on survey design for sensitive questions has found that RR performs better than DE in field-based, single shot, large scale surveys (Rosenfeld, Imai and Shapiro, 2016). Blume, Lai and Lim (2019) and Blume, Lai and Lim (2023) study communication via noisy channels in the lab. They find that mechanisms like HG can improve information transmission but may fall well short of perfect reporting.

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.[8] 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, senior management considered setting up systems for ongoing real-time reporting of issues.

Ethnographic evidence and evidence from community-based surveys suggests that harassment is a long-running problem in Bangladesh's garment sector (Siddiqi, 2003, Sumon, Borhan and Shifa, 2018, Kabeer, Huq and Sulaiman, 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 abusive managers (Siddiqi, 2003). While there is reason to believe that harassment is widespread, measuring and constructing informative statistics of harassment is challenging, even in social science research conducted outside of the workplace. For example, in data from Kabeer, Huq and Sulaiman (2020)'s community-based survey of garment workers, 20% (11%) of workers report witnessing physical (sexual) harassment, but 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 add texture and fading to sewn garments. Within these sections, workers are organized into production teams or lines, with stable team assignments over time. The organizational structure is hierarchical: teams of workers are typically overseen by 2 supervisors, followed by

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

line chiefs or team incharges, floor-supervisors and/or assistant production managers, production manager(s), and finally, the managing director. The gender balance varies considerably across production sections: cutting and wet washing sections typically exclusively employ men; sewing and finishing sections mostly employ women; and dry washing sections are mixed. In all sections, more than 90% of managers are men.

In the two plants 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 support sections. 93% of managers are men.

3 Conceptual Framework

This section defines statistics of interest, discusses their relevance for policy, and clarifies how survey design impacts our ability to obtain less biased estimates under the presence of fear of retaliation and reputational costs.

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, there are $n \equiv m \times L$ 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 four statistics that help organizations assess policy options:

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

Statistic S_V measures the share of victimized workers. This allows decision-makers to gauge the magnitude of the harassment problem in their organization, allowing stakeholders to prioritize the issue appropriately.

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 matters for policy. If there exists k large such that $\mathsf{S}_{TV \geq k}$ is small, but $k \times \mathsf{S}_{TV \geq k}$ is large, then a relatively small share of managers is responsible for a large amount of the damage. This means that investigating and firing the worst 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, norm change, 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. 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.

Statistics of intended reports. Statistics S_V, S_{PM} and $\mathsf{S}_{TV\geq k}$ are not identifiable because they depend on actual harassment patterns that are not directly observed by the decision-maker. Instead, the decision-maker must rely on reported information.

Each worker i in team a completes a binary survey on their harassment status, submitting an intended response $r_{i,a} \in \{0, 1\}$, which may differ from the recorded response $\tilde{r}_{i,a}$ under garbled survey design. We approximate statistics of true harassment with statistics of intended reports. Let S^r_V, S^r_{PM}, $\mathsf{S}^r_{TV\geq k}$, and $\mathsf{E}^r_{2V|1V}$ denote analogues of S_V, S_{PM}, $\mathsf{S}_{TV\geq k}$, and $\mathsf{E}_{2V|1V}$ computed using intended reports $r_{i,a} \in \{0, 1\}$ instead of actual harassment status $h_{i,a}$.[9]

Survey design. If workers are concerned about retaliation, statistics of reported harassment will be systematically biased. 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 investigate the role of survey design in obtaining less biased statistics of underlying harassment.

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: $r_{i,a} \in \{0, h_{i,a}\}$.[10] Following Chassang and Padró i Miquel (2019) and Chassang and Zehnder (2024), we consider garbled survey methods that add noise to the report sent by a worker. An intended report

[9] We will show that statistics of intended reports are identified given recorded reports.

[10] We discuss the possibility of false positives, as well as equilibrium responses to garbling by managers, in Online Appendix D.1. As Chassang and Padró i Miquel (2019) observe, getting workers to complain is a necessary first step, even if false positives become an issue.

$r \in \{0,1\}$ 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 $\pi \in (0,1)$. Reports of harassment are always recorded, but reports of no harassment are switched to reports of harassment with 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 believes that their recorded report $\widetilde{r}_{i,a}$ is leaked with subjective probability $p \in [0,1]$. We do not take a stance on whether leaks actually occur or not. In our experimental application, leaks of individual reports would only occur in the event of an unforeseen breach of confidentiality. 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. For instance, the manager may be sent to a training seminar. Leaks are inevitable even under ideal governance.

Bias and inference. If individuals fear retaliation, statistics of harassment computed using direct elicitation are likely biased. The value of survey design offering plausible deniability, is to reduce this bias. Online Appendix B formalizes this idea in a model of reporting under fear of retaliation, where incentives to hold back infor-

mation are proportional to the likelihood of a leak p and the public posterior belief that the worker submitted a complaint $r_{i,a} = 1$. The following holds.

Claim 1 (bias)**.** *Biases* $|\mathsf{S}_V - \mathsf{S}^r_V|$, $|\mathsf{S}_{PM} - \mathsf{S}^r_{PM}|$, *and* $|\mathsf{S}_{TV\geq k} - \mathsf{S}^r_{TV\geq k}|$ *are decreasing in garbling rate* π *and increasing in perceived leakage probability* p*.*

This suggests using reporting data obtained under higher garbling rates π, and surveys with lower subjective p, to make inferences about statistics of interest S_V, S_{PM}, and $\mathsf{S}_{TV\geq k}$, as they exhibit lower bias. We evaluate two strategies to achieve lower subjective p: varying whether team identifiers are elicited or not; varying the extent of rapport built prior to the sensitive module. However, we acknowledge that all instances of harassment still may not be reported.

Because statistics $\mathsf{E}_{2V|1V}$ and $\mathsf{E}^r_{2V|1V}$ are not necessarily monotonic in $h_{i,a}$ (both the numerator and denominator are increasing functions of harassment and reported harassment), their case requires more nuance. We note that from a policy perspective, $\mathsf{E}^r_{2V|1V}$ may be at least as relevant as $\mathsf{E}_{2V|1V}$. It captures the likelihood of two or more occurrences of harassment being reported conditional on at least one being reported. This number is useful to evaluate burden of proof policies: if $\mathsf{E}^r_{2V|1V}$ is low, then governance processes that only escalate cases when multiple victims come forward end up ignoring a majority of the victims that choose to come forward.

Some care is needed if $\mathsf{E}_{2V|1V}$ is the statistic of interest, since it is not necessarily true that increasing the garbling rate π reduces the bias $|\mathsf{E}_{2V|1V} - \mathsf{E}^r_{2V|1V}|$. Under additional assumptions, Online Appendix B provides a data driven condition under which it is the case, and shows that it holds in our setting.

Identification. Importantly, statistics of intended reports S^r_V, S^r_{PM}, $\mathsf{S}^r_{TV\geq k}$, and $\mathsf{E}^r_{2V|1V}$ are identified from garbled reports. 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:

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

Claim 2 (identification)**.** *Under the garbled survey design used in the experiment, there exists a fixed* $(L+1, L+1)$ *matrix* M *such that* $M\widetilde{\mu}$ *is:*

(i) an unbiased estimator of the distribution of intended reports μ*;*

(ii) a consistent estimator of μ *as the number of teams* m *grows large.*

See Appendix B for a formal statement and a proof.

This generalization of Warner (1965) yields consistent estimators of statistics S^r_V, S^r_{PM}, and $\mathsf{S}^r_{TV\geq k}$ since they are functions of distribution μ. Without Claim 2, Claim 1 would have no operational value.

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, conducted all data collection. We prepared a pre-analysis plan (PAP) 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. The first call included the informed consent process,[11] in which we provided strong general assurances around the confidentiality of workers' responses and informed them that participation in the survey was

[11] See section 1.1 of the Supplementary Materials for the script.

completely voluntary and could be stopped at any time. We informed them that only aggregated results would be shared with senior management and would inform HR policy. 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. All survey enumerators for the study were women.[12] The Supplementary Materials include all scripts and instruments.

4.1 Harassment Outcomes

We were interested in measuring workers' experience of three types of harassment: threatening behavior, physical harassment, and sexual harassment. For each type of 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 reporting threatening behavior (sexual harassment) would be least (most) sensitive.

4.2 Treatment Conditions

We randomly assigned survey participants to different combinations of treatment conditions. 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.

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

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.

- 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.[13] 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

[13]We implemented blocked garbling ex ante. It is not conditional on the intended response.

explaining a 1-in-5 flipping rate. Online Appendix C further explains the reasons for our HG design choices. Supplementary Materials provide implementation details and scripts. 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."[14]

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. Instead, respondents may trust the assertion that "Our system is set up so that it's safe to report an issue" without understanding the garbling mechanism. We address this challenge by replicating this experiment using a design that disentangles these two mechanisms, and find evidence consistent with the former (see Section 7).

[14] We thank an anonymous referee for noting that the sentence *"The researchers are only interested in the total number of yes/no responses from all surveys"* is incomplete: we also estimate team-level statistics, which are aggregate statistics that fall within the scope of the informed consent process. We amended the language for the second replication (see Section 7).

Rapport-building (RB):

- Status quo approach: survey enumerators follow a typical research introduction script before beginning the survey and then ask the survey questions.
- RB approach: survey enumerators allocate survey time to build rapport with the participant. RB entails chatting about family and hobbies in a natural but pre-specified manner, beyond the small talk typical in social science surveys.[15]

We developed our RB treatment modules building on practitioner guidelines for 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, Vasquez and Reichert, 2020) and from research focused on protocols for criminal investigations of sexual abuse allegations (e.g. Cowles, 1988, Vallano and Compo, 2011, Hershkowitz, Lamb and Katz, 2014). For details on the development of our RB approach and our RB modules, see the paper's Supplementary Materials.

The control is the status quo; the treatment is rapport building (RB), administered in a shorter and a longer version to test whether excessive rapport building backfires by eliciting personal disclosures respondents find uncomfortable. In the shorter version, RB1, the enumerator signals care through emotional mirroring and acknowledgment. The longer version, RB2, adds reciprocity: the respondent can ask the enumerator questions, and the enumerator shares a related experience.

Removing personally-identifying information (Low-PII):

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

[15]During training, survey enumerators developed and practiced the RB approach using role plays. A senior research associate tested each enumerator before 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 the team-level statistics of harassment. This represents an unavoidable trade-off.[16]

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)	2244 (2620)

Table 2 summarizes the treatment arms. Arm 1 represents the status quo survey approach. Ex ante, we identified arm 7 as the most protective. Figure A.1 displays the survey modules and locations of treatment interventions, which are after all non-harassment related modules.

4.3 Sampling and Assignment to Treatment Arms

Sampling. We conducted a stratified random selection of workers by team and sex, selecting a minimum of one worker per stratum per treatment arm and sampling larger strata in proportion to their share of the population.

Based on power calculations, we targeted a sample size of 2,620. We attempted

[16]We sampled workers from teams using the employee lists, so respondent in the low PII condition may have inferred that we possessed some team-level identifying information; that said, we did not know workers' supervisors, who are the subject of the harassment questions. For the rest of the paper, consistent with the intention of this treatment condition, we drop respondents in the low PII condition when estimating team-level statistics.

to recruit 3,582 workers and 2,244 participated (63%). The main reason for non-response was that we were not able to reach workers by phone (73% of cases); of workers whom we reached, we surveyed 87%. Workers were not aware of their treatment assignment when deciding whether to participate, and *no* workers dropped out during the survey. We did not achieve our target sample size because some strata ran out of candidate replacement workers. The median team in the data has 39% of its workers surveyed and a response rate of 65%.

Online Appendix C.4 provides details on the sampling, and Online Appendix C.5 presents detailed evidence that non-response is uncorrelated with harassment experience and treatment status.[17]

Assignment to Treatment Arms. The unit of randomization is a worker, stratified by team and sex. To improve balance, we followed Banerjee et al. (2020), conducting 10 randomizations and selecting the one that performed best in terms of balance on available covariates. Online Appendix C.4 provides details on the randomization.

Integrity of the experiment and baseline balance. Table A.1 presents worker-level summary statistics. As explained in Online Appendix C.2, we drop surveys of one enumerator who did not adhere to the protocol for HG, which reduces the sample size to 2,143. Table A.2 presents team-level summary statistics. Table A.3 shows balance tests for workers' characteristics across the main treatment conditions, and Table A.4 does the same across no rapport, short rapport, and long rapport conditions. There are no statistically significant differences across treatment conditions.

[17] One concern is that supervisors who engage in more harassment may have pressured workers not to participate. This would not affect internal validity, as respondents did not know their treatment status when deciding their participation. However this would bias downward the statistics of intended reports in Section 6. Online Appendix C.5 presents evidence that this was not the case.

5 The Impact of Survey Design

We seek to estimate coefficients in the following regression:

$$r_{is} = \alpha\, HG_i + \beta\, Rapport_i + \gamma\, LowPII_i + \mu_s + \theta X_i + \epsilon_{is} \tag{1}$$

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 (PDS) lasso (Belloni, Chernozhukov and Hansen, 2014). Eligible control variables include the worker characteristics reported in Table A.1.

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

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

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

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

with $\pi = .2$ for the HG group and $\pi = 0$ for the DE group.[18] Normalized report $\hat{r}_i$ is equal to the intended report plus a heteroskedastic error term ξ_i. If r_i satisfies (1), then $\hat{r}_i$ satisfies a similar regression (1b) with heteroskedastic, mean-zero errors

[18]The garbling was blocked within team but not within treatment arm, which resulted in small variations across arms in the realized garbling rate. We use the realized garbling rates.

(conditional on covariates). Consequently, OLS is consistent, and robust standard errors are correct. With blocked-garbling (see Online Appendix B), error terms ψ_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 block (HG respondents) or respondent (DE respondents):[19]

$$\hat{r}_{is} = \alpha\, HG_i + \beta\, Rapport_i + \gamma\, LowPII_i + \mu_s + \theta X_i + \xi_{is} \tag{1b}$$

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

$$\begin{aligned}\hat{r}_{is} = {} & \alpha_f\, HG_i{*}Female_i + \alpha_m\, HG_i{*}Male_i + \beta_f\, Rapport_i{*}Female_i + \beta_m\, Rapport_i{*}Male_i \\ & + \gamma_f\, LowPII_i * Female_i + \gamma_m\, LowPII_i * Male_i + \lambda\, Female_i + \mu_s + \theta X_i + \xi_{is}\end{aligned} \tag{2}$$

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.

$$\begin{aligned}\hat{r}_{is} = {} & \alpha_1\, \mathbb{1}(\text{DE} \times \text{PII} \times \text{RB 1})_i + \alpha_2\, \mathbb{1}(\text{DE} \times \text{PII} \times \text{RB 2})_i + \alpha_3\, \mathbb{1}(\text{DE} \times \text{Low PII} \times \text{RB 1})_i \\ & + \alpha_4\, \mathbb{1}(\text{DE} \times \text{Low PII} \times \text{No RB})_i + \beta_1\, \mathbb{1}(\text{HG} \times \text{PII} \times \text{No RB})_i + \beta_2\, \mathbb{1}(\text{HG} \times \text{PII} \times \text{RB 1})_i \\ & + \beta_3\, \mathbb{1}(\text{HG} \times \text{PII} \times \text{RB 2})_i + \beta_4\, \mathbb{1}(\text{HG} \times \text{Low PII} \times \text{RB 1})_i + \mu_s + \theta X_i + \xi_{is}\end{aligned} \tag{3}$$

[19] We pre-specified reporting of 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. For regressions that do not use intended reports, we report robust standard errors as per our PAP.

5.1 Results

Main effects of survey design on reporting. Table 3 reports the main treatment effects.[20] In this and other regression tables, odd-numbered columns display the results from the baseline specification, while even-numbered columns display the results with PDS lasso-selected controls.

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. In percentage points (ppts), HG increases the reporting of threatening behavior, physical harassment, and sexual harassment, respectively, by 4.5 ppts (45.4%, p <0.01), 4.4 ppts (290%, p <0.01), and 4.8 ppts (271%, p <0.01).[21] Removing questions about respondents' supervisor (Low PII) increases the reporting of physical harassment by a marginally statistically significant 2.9 ppts (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, a 11.6% increase) and sexual harassment (1.90 ppts, a 107% increase), but it is not statistically significant. Rapport has no detectable effect on the reporting of physical harassment.

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 less clear. Low PII appears to increase the reporting of physical harassment, but not of threatening behavior or sexual harassment, and it comes at the cost of not being able to calculate manager-level statistics. That low PII generated weaker effects than HG may be due to its more limited guarantee of plausible deniability in case of a leak; HG reduces the informativeness of any leak about whether a worker intentionally reported, while low PII leaves the report's content intact, so workers who are the

[20] Table A.5 reports the main results with separate indicators for short- and long-RB conditions.

[21] The average treatment effects of HG are similar across types of harassment, but this is not the case conditional on gender (see Table 4), or in replications.

Table 3: 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.

sole possible source of a complaint remain identifiable to a well-informed supervisor even absent explicit identifying fields.

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.

Online Appendix D conducts further tests on the robustness of our findings. We show that lack of comprehension among respondents increasing false positives under HG is not the main driver. We also show that treatment status does not affect reporting behavior associated with a placebo question using DE for all arms, which suggests that respondents' behavior reflects a real understanding of the benefits of plausible deniability, as opposed to their generally trusting the surveyor more. We also discuss how to interpret our findings if there are concerns over false reporting by workers. Section 7 discusses the external validity of these findings along several

dimensions, with additional evidence from two replication surveys.

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 are different for men and women, we estimate the main effects separately by sex in Table 4. 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 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.

HG continues to increase reporting across the board. Interestingly, the point estimates 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 fact that the effects of HG are, if anything, more positive for men echoes recent results on intimate partner violence (IPV) in Kenya. Using list experiments, Castilla and Murphy (2023) find that roughly equal numbers of men and women are victims of IPV. In contrast, using DE, Haushofer et al. (2020) find that women report being the victim of IPV at higher rates than men.[22]

The impact of removing PII appears to be weakly more positive for women, although we cannot reject that the effects for both groups are zero or the same. Table 4 suggests that the impact of rapport on men and women may be different. For 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 small talk with an unknown woman may have raised

[22]Nontrivial numbers of men are also found to be victims of sexual harassment in other contexts. For instance, in a 2019 survey from the United States, 81% of women and 43% of men reported experiencing some form of sexual harassment and/or assault in their lifetime, with 49% and 18% reporting having been sexually touched in an unwelcome way (UC San Diego Center on Gender Equity and Health, 2019). In a study conducted in seven cities in India, 19% of men reported being victims of sexual harassment (ET Bureau, Aug 22, 2010).

men's suspicion regarding the survey. Our main takeaway is that RB may be effective at increasing reporting, but the RB process needs to be tightly tailored to target subjects. Further experimentation probing how trusting relationships can be built with heterogeneous survey respondents is needed.

Table 4: 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	Yes	No	Yes	No	Yes
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.

Interactions among treatment conditions. Figure 1 examines the possibility that the treatment conditions substitute or complement each other using equation (3) (see also Table A.6). It shows that, to a first order, removing PII and building

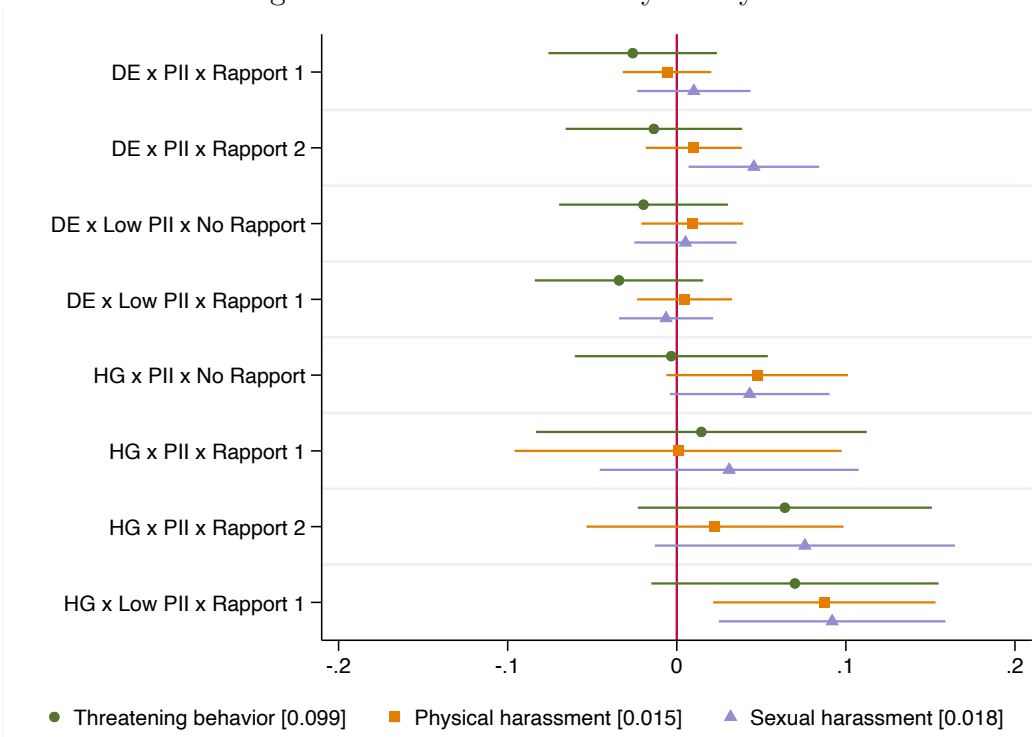
rapport do not seem to have a large impact when using DE. There may be an impact of extended rapport on the reporting of sexual harassment. While this individual coefficient is significant, the overall picture invites caution. Figure A.2 shows that the negative effect of RB on reporting of threatening behavior is driven by men, while its positive effect on reporting of sexual harassment is driven by women. This suggests that: (1) RB does impact reporting, but its effect is subtle and varies across participants; and (2) that RB needs to be carefully tailored.

Returning to Figure 1, it shows that when using HG, the 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 effects. This suggests that there exist complementarities between HG and other steps that foster trust in the survey protocol, which is intuitive since HG can only be effective if respondents trust that the protocol will be followed. The bottom of Table A.7 reports the p-values from the statistical test of the null hypothesis of no complementarity among HG, removing team-level identifying information, and RB (see table notes for details). The test is rejected for threatening behavior (p=0.035), but not for physical and sexual harassment (p=0.247 and p=0.256, respectively). We thus interpret this as suggestive evidence of complementarity among the design features.

6 Understanding Harassment

We now report team-level statistics of harassment that inform policy-making. Consistent with Claims 1 and 2, we believe responses from garbled treatment arms provide the least biased estimates of statistics of true harassment. We contrast them to statistics computed from DE responses. We emphasize that the policy insights we draw are organization specific. Our policy diagnostics should be replicated to tailor policy responses in other organizations.

Figure 1: Treatment effects by survey arm

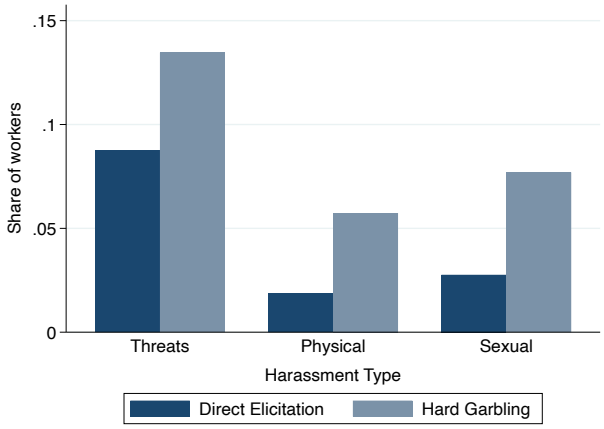


Notes: This figure reports coefficients from separate regressions of the outcome variable on the treatment arm indicators, strata fixed effects, and controls selected using the PDS lasso. The regression specification is equation (3). The whiskers are 95% confidence intervals estimated using robust standard 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.

Based on intended reports from garbled surveys, we find that harassment is prevalent and that victims willing to report tend to be *effectively* isolated. This refutes (for this organization) the hypothesis that a few bad actors are responsible for the majority of the misconduct and cautions against policies that exclusively target small numbers of offenders with multiple known victims. Further, our evidence cautions against relying on policies that require multiple victims to initiate action. In our setting, they would fail to resolve the reports of nearly all victims willing to report.

Scale of the issue and potential gains. Figure 2 illustrates the estimated share of workers reporting harassment S_V^r, computed using Claim 2 under DE and HG.

Figure 2: Share of workers who report harassment (S^r_V) by survey method



Notes: This figure illustrates reported harassment rates estimated using reporting with DE and HG, respectively. We compute the harassment rates using (B.5) 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.

Since this statistic does not require PII, we pool data across low and high PII arms.

HG considerably increases workers' propensity to report harassment relative to DE: 13.5% vs. 8.7% for threatening behavior, 5.7% vs. 1.9% for physical harassment, and 7.7% vs. 2.8% for sexual harassment. For context, the firm's own internal reporting channels capture even less — 3.7%, 1.0%, and 1.9%, respectively, based on DE respondents' reports of whether they had disclosed to the firm.

The primary takeaway is that harassment is meaningfully more widespread than standard surveys – or the firm's internal reporting channels – 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.

Scope and nature of the issue. We compute the team-level statistics of interest, $\mathsf{S}^r_{TV\geq k}$ and $\mathsf{E}^r_{2V|1V}$, by pooling reporting data from treatment arms that collect PII.

When reports are garbled, we estimate the distribution of number of intended reports in the sample population of teams using the unbiased consistent estimator of Claim 2. The number of workers surveyed per team in treatment arms that collect PII varies from 3 to 12, with a median and mean of 7 for both HG and DE arms.

Table 5: Estimated team-level statistics $\mathsf{S}^r_{TV\geq k}$ for $k \in \{1,2\}$ and $\mathsf{E}^r_{2V|1V}$

	Threatening Behavior		Physical Harassment		Sexual Harassment	
Statistic	HG	DE	HG	DE	HG	DE
$\mathsf{S}^r_{TV\geq 1}$	0.719	0.446	0.254	0.098	0.402	0.188
	(0.068)	(0.046)	(0.045)	(0.029)	(0.053)	(0.039)
$\mathsf{S}^r_{TV\geq 2}$	0.112	0.125	0.035	0.009	0.033	0.027
	(0.038)	(0.032)	(0.019)	(0.011)	(0.022)	(0.015)
$\mathsf{E}^r_{2V\|1V}$	0.156	0.280	0.138	0.091	0.083	0.143
	(0.055)	(0.063)	(0.076)	(0.118)	(0.056)	(0.074)

Notes: This table reports team-level statistics $\mathsf{S}^r_{TV\geq k}$ for $k \in 1,2$ and $\mathsf{E}^r_{2V|1V}$, estimated using pooled data from treatment arms that collect PII under either HG and DE. The sample includes 112 teams. Bootstrapped standard errors are reported in parentheses.

Table 5 reports $\mathsf{S}^r_{TV\geq k}$ for $k \in \{1,2\}$ and $\mathsf{E}^r_{2V|1V}$. Under DE, the estimated share $\mathsf{S}^r_{TV\geq 1}$ of teams in which at least one worker reports is 44.6% for threatening behavior, 9.8% for physical harassment, and 18.8% for sexual harassment. Under HG, the problem is revealed to be much more widespread: the rates are 71.9%, 25.4%, and 40.2%, respectively. Under-reporting at the individual level makes it likely to miss harassment even when aggregated at the team level.

The share of teams with at least 2 workers reporting harassment under DE (HG) is 12.5% (11%) for threats, 0.9% (3.5%) for physical harassment, and 2.7% (3.3%) for sexual harassment. Notably, while $\mathsf{S}^r_{TV\geq 1}$ is substantially higher under HG, $\mathsf{S}^r_{TV\geq 2}$ is similar between HG and DE for threatening behavior and sexual harassment, indicating that HG is primarily changing the team-level statistics relative to DE by inducing a first person to report on teams where no one else comes forward, as opposed to adding corroborating reports to teams already captured by DE. Further, these pat-

terns indicate that the organization needs to have policies to address widespread manager engagement in harassment with limited numbers of victims reporting per manager: S^r_{PM} is high, while $k\mathsf{S}^r_{TV\geq k}$ is not high for k large.

Across all types of harassment, even under HG, victims willing to report are *effectively isolated*: the $\mathsf{E}^r_{2V|1V}$ estimates show that 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.[23] Our estimate of intended reports is a lower bound to the real amount of harassment, and we cannot separate true isolation of victims from sparse reporting. Under additional assumptions, Online Appendix B argues that when the rate at which harassment is reported is small, $\mathsf{E}^r_{2V|1V} < \mathsf{E}_{2V|1V}$, i.e. isolation based on reports is worse than isolation based on victim status. In addition, the bias shrinks as the garbling rate increases.

This isolation is also why individual-level underreporting causes the scope to be underestimated. If a typical problem manager had several willing reporters among their victims, even a low per-victim reporting probability, such as 10%, would compound across independent reporters into a high chance that at least one comes forward, and DE might suffice to reveal most problem managers. It is because willing reporters are so scarce per manager – often just one – that team-level detection depends on that single individual's reporting probability. Consequently, many more problem managers are identified under HG than DE.

Policy implications. Our findings suggest four policy takeaways. First, harassment is much more prevalent than complaints under DE would suggest, and it affects both men and women. Second, harassment is widespread across teams. This means that firing a few bad apples cannot be the sole policy option. Instead, the behavior of many existing managers must be changed. Third, HG raises team-level detection primarily by inducing first, as opposed to corroborating, reporters: the share of teams with at least two reporters is broadly similar under HG and DE for threatening

[23] Victims willing to report are similarly *effectively* isolated under DE.

behavior and sexual harassment, even as the share with at least one is substantially higher under HG. Fourth, even under HG, victims willing to report are *effectively* isolated across all harassment types — the probability of a second reporter given a first ranges from only 8% to 17%.

Together, these last two points imply that raising individual-level reporting mainly expands how many problem managers can be identified, since getting one person to report rather than gaining corroboration surfaces cases DE would have missed. It does not make any single case easier to resolve: even under HG, most identified managers remain linked to only one report. Consequently, requiring multiple victims to come forward would therefore miss the majority of cases regardless of harassment type. Remedying harassment requires processes that can be initiated by a single report.

7 External Validity

Three challenges bear on the external validity of our findings:

(i) Do the findings replicate across treatment site and across time? Neither is guaranteed (Cook, Shadish and Wong, 2008, Rosenzweig and Udry, 2020, Vivalt, 2020, Dehejia, Pop-Eleches and Samii, 2021).

(ii) What aspect of the treatment yields the effect? Is plausible deniability instrumental, or would making safety assurances in DE as close as possible to those in HG suffice?

(iii) Would the impact remain over the long-term, or would garbled reporting encourage misbehaving managers to increase retaliation, reducing reporting and generating conflict within the organization?

Two replication efforts allow us to make progress on (*i*), and (*ii*), and partially

address (*iii*).[24,25] We next briefly describe the context of each replication; Online Appendix E gives additional details on these replications' backgrounds, including response rates, summary statistics, and balance tests (Tables E.1, E.4, E.5).

7.1 Validity across sites and time

Our primary experiment was conducted with two large factories from September to November of 2021, making the findings potentially sensitive to those factories and to a particular macroeconomic, political, and social moment (Rosenzweig and Udry, 2020). We used the endline survey of Boudreau, Chassang and González-Torres (2022), which studies the effectiveness of information escrows proposed by Ayres and Unkovic (2012), to evaluate the impact of HG in a different context.

Survey participants for this replication were recruited through the community, allowing us to reach 200 factories (2,951 workers, 55% women), with interviews staggered from September 2023 to November of 2024. Workers' experience of harassment were elicited using the same DE and HG scripts as our primary experiment, adding shouting as a non-sensitive placebo question.

Columns 1-4 of Table 6 report our findings. As in our main experiment, sensitive types of harassment are unlikely to be reported under DE, and HG has a large, significant positive impact. The null for shouting is consistent with its high DE reporting rate. Panel B presents HTEs by gender: effects appear to larger for men for threatening behavior, but are similar magnitudes for men and women for physical and sexual harassment.

[24]Online Appendix D presents further empirical evidence supporting the robustness of our primary experiment's findings.

[25]The ARCED Foundation, a non-profit research organization in Bangladesh that conducts research on various social issues, implemented both replications.

7.2 Validity across mechanisms and time-horizons

In a second replication effort from December 2025 to March 2026, we surveyed 5,490 workers (50% women) from 4 randomly-selected garment suppliers to a large international buyer that aimed to improve its supplier guidelines on resolving worker grievances. The survey is highly similar to our primary survey and was preregistered with an analogous analysis plan. The key design changes target mechanisms:

1. We made safety assurances associated with DE as close as possible (while preserving honesty) to safety assurances associated with HG.

2. We introduced variation in the garbling rate (which was known to survey respondents): a third of respondents were surveyed using a garbling rate of 20% (HG20), and a third were surveyed using a garbling rate of 1% (HG01). The latter treatment provides a materially lower level of plausible deniability compared to HG20.

3. We measured participants' understanding of fractions and used more difficult comprehension questions for HG.

4. We surveyed participants on their preferences over reporting mechanisms for ongoing use within the factory.

The first three changes clarify the importance of plausible deniability in facilitating the transmission of sensitive information.[26] The fourth indicates whether survey respondents anticipate adverse effects of garbling on welfare in the longer run.

[26] See Supplementary Materials Section 2 for scripts and comprehension questions.

7.2.1 Is plausible deniability an important component of treatment?

Columns 5-6 of Table 6 summarize the main findings.[27] Even with increased assurances in DE, and no general assurances in HG, HG20 has large, positive impacts on the reporting of physical harassment (by 1.3 ppt, a 186% increase; p=0.005), from a baseline of 0.7% under DE. HG01 has a similar effect as HG20 on the reporting of physical harassment (by 1.2 ppt, a 171% increase; p=0.009). We find no impacts on the reporting of sexual harassment in this replication. Panel B shows that only HG20, not HG01, increases men's reporting of physical harassment (col. 5, p=0.033). For women, there are effects of both HG20 and HG01 on reporting of physical harassment (col. 5), with larger estimated effects of HG01 compared to HG20 (p=0.095). Evidently, men value HG20's additional plausible deniability and respond to it, while the effects for women are less clear. We return to this below.

An internal validity test provides evidence that these results are not simply explained by increased confusion under HG generating more false positives: Table E.7 reports effects on a verifiable outcome – a worker fatality in the prior 12 months due to an accident – for which the known rate is zero. Neither HG20 nor HG01 increases false positives for men or women.

Two tests provide further evidence that respondents who understand fractions and HG's plausible deniability feature are reporting more under HG due to them. First, Table E.8 shows that controlling for understanding of fractions and its interaction with HG, or dropping the respondents who incorrectly answered the fractions question, grows HG20's effect while shrinking HG01's – consistent with individuals who understand probabilities responding to their assigned garbling rates.[28] Second, as our more difficult HG comprehension questions reveal somewhat lower understanding of HG: 39% (12%) of respondents answer 1 (2) or more of 3 questions wrong,

[27] For conciseness, Table 6 only presents results with PDS lasso-selected controls (results are unchanged without them, see Table E.6). We preregistered that we would report the p-value for the 1-sided test that the effect of HG20 is larger than that of HG01; the test result is reported at the bottom of the panel.

[28] 91% of respondents answered the probability question correctly.

compared to 9% (5%) of 2 questions in the primary experiment; Table E.9 imputes zeros for the *recorded* responses of respondents who answered less than 2 questions correctly and shows that HG's effects on reporting of physical harassment survive.

Finally, Online Appendix E also explores mechanisms for the similar effects of HG20 and HG01 on women's reporting of physical harassment, and in particular, whether lack of comprehension of the HG mechanism or of fractions contributes. The main insight is that the results are inconsistent with either of these being the sole driver; that said, the evidence is consistent with women's higher level of reporting under HG01 (and HG20) being due to their perceiving greater security under it. Beyond this, we cannot identify what explains it and leave this to future research.

Beyond this open question about women's response, as detailed above, our comprehension questions show that a meaningful share of respondents do not fully grasp HG's mechanics in the short run. Yet even less numerate respondents report more harassment under HG than DE, suggesting they anticipate some protection without a precise understanding of the mechanism. As such, in ongoing monitoring settings, it is important to actually provide plausible deniability, and continue to earn respondents' trust as they learn whether the protocol is real.[29]

7.2.2 Is HG helpful in equilibrium?

A concern with HG is that it may lead to increased retaliation by misbehaving managers. In fact, the magnitude of retaliation may increase to induce workers to refrain from reporting. If so, adopting HG would increase organizational strife, without generating much useful information, thereby decreasing worker welfare. We

[29]Chassang and Zehnder (2024) study reporting using DE, HG, and randomized response (RR), which turns out to be strategically equivalent to DE with an HG framing. They show in a laboratory setting that while RR performs well in environments where players do not learn, RR unravels in environments where participants have the opportunity to learn from others' experiences. This implies that in an organizational setting, where participants have enough time to learn to optimize, it is not sufficient to make safety assurances, they need to be supported with actual mechanisms, even if participants only partially understand them in the short run.

Table 6: Main Results: Replications 1 and 2

	Replication 1				Replication 2	
	(1)	(2)	(3)	(4)	(5)	(6)
	Shouting	Threats	Physical	Sexual	Physical	Sexual
Panel A: Main effects						
HG-20%	0.012	0.031*	0.026**	0.031**	0.013***	0.002
	(0.020)	(0.017)	(0.012)	(0.012)	(0.005)	(0.004)
HG-1%					0.012***	0.002
					(0.004)	(0.005)
Control Mean	0.285	0.119	0.013	0.010	0.007	0.013
p(HG20 = HG01)					[0.867]	[0.881]
p(HG20 > HG01)					[0.433]	[0.441]
Panel B: Heterogeneous effects by gender						
HG-20% × Female	0.029	0.005	0.026	0.030*	0.011*	0.006
	(0.027)	(0.023)	(0.017)	(0.017)	(0.006)	(0.007)
HG-20% × Male	-0.009	0.064***	0.028	0.032*	0.014**	-0.001
	(0.030)	(0.025)	(0.019)	(0.018)	(0.007)	(0.006)
HG-1% × Female					0.025***	-0.003
					(0.007)	(0.007)
HG-1% × Male					-0.002	0.006
					(0.005)	(0.007)
Control Mean (Female)	0.278	0.132	0.009	0.013	0.007	0.013
Control Mean (Male)	0.294	0.104	0.019	0.007	0.008	0.012
p(HG20xF = HG20xM)	[0.341]	[0.077]	[0.934]	[0.951]	[0.758]	[0.423]
p(HG01xF = HG01xM)					[0.003]	[0.317]
p(HG20xF = HG01xF)					[0.095]	[0.224]
p(HG20xM = HG01xM)					[0.035]	[0.308]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2935	2935	2935	2935	5421	5421

Notes: This table reports OLS estimates of treatment effects on workers' reporting in Replications 1 and 2. Each column in the table reports the estimated coefficient from a separate regression. Columns 1–4 use the Replication 1 sample (shouting, threats, physical and sexual harassment in the past 10 months); columns 5–6 use the Replication 2 sample (physical and sexual harassment in the past year). Panel A effects pooled across gender; Panel B shows effects interacted with a dummy variable for each gender, and additionally includes a Female indicator as a control (a separate regressor in Replication 1; absorbed by the strata fixed effects, which include sex, in Replication 2). The garbled outcomes are adjusted for the constant flipping rate in Replication 1 (i.i.d. garbling and empirical flipping rate not tracked), and for the empirical flipping rate per treatment-arm in Replication 2 (blocked garbling rate and flipping rate tracked). Replication 1 has a single HG arm with a 20% garbling rate; Replication 2 has an HG arm with a 20% garbling rate, and an HG arm with a 1% garbling rate. All regressions include escrow-treatment (Replication 1) or strata fixed effects defined by factory, production team or section and sex (Replication 2), and controls selected by the post-double-selection (PDS) lasso. The PDS candidate-control vector is the same in both replications and comprises age, schooling and tenure (standardized and squared), ever-married, having children, as well as department indicators (sewing, finishing, cutting, washing) and position (helper, iron/folding, operator, quality). PDS lasso-selected controls: 12 controls in Replication 1; none in Replication 2. Number of factories (strata): Replication 1 = 200; Replication 2 = 4 (121). Standard errors clustered by factory in Replication 1; by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents) in Replication 2; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 7: Replication 2 | Impacts on preferences for different garbling rates

	HG Preferred		HG-20% Best (uncond.)		HG-20% Best (cond.)	
	(1)	(2)	(3)	(4)	(5)	(6)
HG-20%	-0.008	-0.007	0.003	0.003	-0.001	-0.001
	(0.022)	(0.022)	(0.014)	(0.014)	(0.016)	(0.016)
HG-1%	-0.032	-0.033	-0.165***	-0.165***	-0.171***	-0.171***
	(0.022)	(0.022)	(0.015)	(0.015)	(0.018)	(0.018)
Control Mean (DE)	0.632	0.632	0.920	0.920	0.941	0.941
p(HG20 = HG01)	[0.102]	[0.095]	[0.000]	[0.000]	[0.000]	[0.000]
p(HG20 > HG01)	[0.051]	[0.048]	[0.000]	[0.000]	[0.000]	[0.000]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes	No	Yes
Observations	4135	4135	4135	4135	2551	2551

Notes: This table reports OLS estimates of treatment effects on preferences for HG and HG rates in Replication 2. Each column in the table reports the estimated coefficient from a separate regression. Outcomes: HG Preferred indicates preferring the HG system over DE; HG-20% Best (uncond.) indicates preferring the 20% garbling rate over a 5% or a 1% rate, among all respondents who are asked this question (those under HG or a subset of DE who are explained HG after the own harassment section); HG-20% Best (cond.) restricts to respondents who prefer HG. Each regression includes strata fixed effects; odd columns omit, even columns include PDS lasso-selected controls (as described in notes to Table 6). Lasso-selected controls (union across PDS columns): finishing department. Standard errors are clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

probe this issue by asking workers their preferences over reporting systems.

We survey respondents assigned to DE, HG01, and HG20 arms on their preferences over reporting systems twice (this required explaining HG to some of the respondents assigned to the DE arm after eliciting their harassment experience).[30] We first ask them whether they think workers would be better served by DE or HG, without specifying the garbling rate. We then ask respondents about their preferences over garbling rate conditional on HG being implemented (1%, 5%, 20%).

Table 7 reports preferences by treatment arm. Participants exposed to either

[30]We did not explain HG to the entire DE arm because we thought there might be effects of having learned about HG on job satisfaction or mental health on our follow-up study, and indeed, table E.13 does show evidence that having been explained HG – but not given the opportunity to report harassment with it – does lower job satisfaction.

DE or HG20 have very similar preferences: 63% prefer HG to DE (cols 1-2), and more than 90% prefer HG20 to HG01 and HG05, both unconditionally (cols 3-4) and conditional on preferring HG to DE (cols 5-6). Respondents in the HG01 arm are less likely to prefer HG over DE and much less likely to prefer HG20. We interpret this as evidence that exposure to HG01 increases confusion about the benefits of HG. Altogether, this suggests that workers do not perceive adverse welfare effects of HG survey designs in the context of ongoing, longer-run implementation. That most workers prefer higher garbling rates also suggests limited concern about organizational strife.

Online Appendix D considers a second concern about responses to HG in the longer-run: strategic misreporting by workers. It shows strategic misreporting is unlikely to drive the primary experiment's results, and presents theoretical guidance on HG's effectiveness in equilibrium with strategic responses by managers and workers.

8 Conclusion

Building on principal-agent-monitor models with retaliation (Chassang and Padró i Miquel, 2019, Chassang and Zehnder, 2024), we experimentally evaluate survey designs that vary plausible deniability, trust-building, and the perceived likelihood of leaks. HG considerably raises reporting of harassment by both women and men; rapport-building and reducing PII have positive but weak effects. Extending inference results from the garbled-survey literature (Warner, 1965, Blair, Imai and Zhou, 2015) to *team-level* statistics, we show that HG reveals substantially more widespread harassment inside the firm than DE. Two field replications support the external validity of our HG findings.

The greater share of victims reporting is HG's main benefit. Appendix C.3 weighs this against its costs – participants' concerns over leaked garbled reports, the added noise that limits how severely organizations can act on reports, and implementation costs – and argues each is manageable in our context: workers reveal a preference for

HG despite leak concerns (Section 7.2.2), and set-up costs were modest, though they could be larger for organizations with less capacity to explain HG to respondents.

Three directions for future research strike us as particularly valuable. The first concerns bringing HG into firms' ongoing governance systems. On the information-gathering side, internal surveyors may face weaker trust than the third-party enumerators in our setting. The increasing effects of HG with our trust-enhancing steps (Figure 1) suggests this margin matters, and we need evidence on what steps are required in internal settings. On the intervention side, we need evidence on what meaningful actions organizations can take based on noisy signals that do not invite strategic responses by managers or workers – issues we discuss in Online Appendix C.

The second concerns the welfare effects of reporting for workers, managers, and firms (Adams-Prassl et al., 2024, Folke and Rickne, 2022, Dahl and Knepper, 2026). We resurveyed workers 1–2 weeks after the primary and second replication experiments to assess effects on mental well-being and job satisfaction; using HG assignment as an instrument for reporting, we find suggestive evidence of improvements on both margins (Online Appendix Table E.12).

The third concerns the generalizability of our findings on HG, RB, and Low PII. Section 7 provides external validity evidence for HG, but evidence is still needed across other countries, sectors, organization types, and sensitive issues. Sample-size constraints kept us from including RB and Low PII in the replications. For RB, future work could vary scripts, match enumerator and respondent genders, and explore other approaches to building trust with heterogeneous respondents. For Low PII, it remains open whether greater *de jure* anonymity raises reporting in other settings, or whether the fact that actions on reports may *de facto* reveal identifying information limits the approach's value.

References

Adams-Prassl, Abi, Kristiina Huttunen, Emily Nix, and Ning Zhang.

2024. "Violence Against Women at Work." *The Quarterly Journal of Economics*, 139(2): 937–991.

Aguilar, Arturo, Emilio Gutiérrez, and Paula Soto 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(4): 1379–1410.

Anderson, Michael 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(484): 1481 –1495.

Ayres, Ian, and Cait Unkovic. 2012. "Information escrows." *Mich. L. Rev.*, 111: 145.

Bac, Mehmet. 2009. "An economic rationale for firing whistleblowers." *European Journal of Law and Economics*, 27(3): 233–256.

Banerjee, Abhijit V., Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. "A Theory of Experimenters: Robustness, Randomization, and Balance." *American Economic Review*, 110(4): 1206–30.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Contr." *Review of Economic Studies*, 81(2): 608–650.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. "Design and analysis of the randomized response technique." *Journal of the American Statistical Association*, 110(511): 1304–1319.

Blume, Andreas, Ernest K Lai, and Wooyoung Lim. 2019. "Eliciting private information with noise: the case of randomized response." *Games and Economic Behavior*, 113: 356–380.

Blume, Andreas, Ernest K Lai, and Wooyoung Lim. 2023. "Mediated talk: An experiment." *Journal of Economic Theory*, 208: 105593.

Borker, Girija. 2018. "Safety First: Perceived Risk of Street Harassment and Educational Choices of Women." mimeo.

Boudreau, Laura, Rachel Heath, and Tyler H McCormick. 2024. "Migrants, Experience, and Working Conditions in Bangladeshi Garment Factories." *Journal of Economic Behavior & Organization*, 219: 196–213.

Boudreau, Laura, Sylvain Chassang, and Ada González-Torres. 2022. "Whistleblowing and Worker Well-being: Evidence from Bangladesh's Garments Sector." *working paper*.

Carril, Alvaro. 2017. "Dealing with misfits in random treatment assignment." *Stata Journal*, 17(3): 652–667.

Castilla, Carolina, and David MA Murphy. 2023. "Bidirectional intimate partner violence: Evidence from a list experiment in Kenya." *Health economics*, 32(1): 175–193.

Chassang, Sylvain, and Christian Zehnder. 2024. "Secure Survey Design in Organizations: Theory and Experiments." *American Economic Journal: Microeconomics*.

Chassang, Sylvain, and Gerard Padró i Miquel. 2019. "Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports." *Review of Economic Studies*, 86: 2530–2553.

Cheng, Ing-Haw, and Alice Hsiaw. 2022. "Reporting Sexual Misconduct in the #MeToo Era." *American Economic Journal: Microeconomics*, 14(4): 761–803.

Chuang, Erica, Pascaline Dupas, Elise Huillery, and Juliette Seban. 2021. "Sex, lies, and measurement: Consistency tests for indirect response survey methods." *Journal of Development Economics*, 148: 102582.

Cook, Thomas D, William R Shadish, and Vivian C Wong. 2008. "Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 27(4): 724–750.

Cowles, Kathleen V. 1988. "Issues in qualitative research on sensitive topics." *Western Journal of Nursing Research*, 10(2): 163–179.

Dahl, Gordon B, and Matthew Knepper. 2026. "Why Is Workplace Sexual Harassment Underreported? The Value of Outside Options amid the Threat of Retaliation." *American Economic Review*, 116(3): 897–933.

Davison, Anthony Christopher, and David Victor Hinkley. 1997. *Bootstrap methods and their application.* Cambridge university press.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2021. "From local to global: External validity in a fertility natural experiment." *Journal of Business & Economic Statistics*, 39(1): 217–243.

Efron, Bradley. 1987. "Better bootstrap confidence intervals." *Journal of the American statistical Association*, 82(397): 171–185.

ET Bureau. Aug 22, 2010. "Even men aren't safe from sexual harassment at workplace: Survey." *The Economic Times*.

Faure-Grimaud, Antoine, Jean-Jacques Laffont, and David Martimort. 2003. "Collusion, delegation and supervision with soft information." *The Review of Economic Studies*, 70(2): 253–279.

Folke, Olle, and Johanna Rickne. 2022. "Sexual Harassment and Gender Inequality in the Labor Market." *Quarterly Journal of Economics*, 1–50.

Greenberg, Bernard G, Abdel-Latif A Abul-Ela, Walt R Simmons, and Daniel G Horvitz. 1969. "The unrelated question randomized response model: Theoretical framework." *Journal of the American Statistical Association*, 64(326): 520–539.

Haushofer, Johannes, Charlotte Ringdal, Jeremy Shapiro, and Xiao Yu Wang. 2020. "Spousal disagreement in reporting of intimate partner violence in Kenya." Vol. 110, 620–624, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Hershkowitz, Irit, Michael E Lamb, and Carmit Katz. 2014. "Allegation rates in forensic child abuse investigations: Comparing the revised and standard NICHD protocols." *Psychology, Public Policy, and Law*, 20(3): 336.

Heyes, Anthony, and Sandeep Kapur. 2009. "An economic model of whistle-blower policy." *The Journal of Law, Economics, & Organization*, 25(1): 157–182.

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

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

Kondylis, Florence, Arianna Legovini, Kate Vyborny, Astrid Zwager, and Luiza Andrade. 2025. "Demand for "Safe Space": Avoiding harassment and complying with norms." *Journal of Development Economics*, 174.

Laffont, Jean-Jacques, and David Martimort. 1997. "Collusion under asymmetric information." *Econometrica*, 65: 875–911.

Laffont, Jean-Jacques, and David Martimort. 2000. "Mechanism design with collusion and correlation." *Econometrica*, 68(2): 309–342.

Ljungqvist, Lars. 1993. "A unified approach to measures of privacy in randomized response models: A utilitarian perspective." *Journal of the American Statistical Association*, 88(421): 97–103.

Macchiavello, Rocco, Andreas Menzel, Atonu Rabbani, and Christopher Woodruff. 2020. "Challenges of Change: An Experiment Promoting Women to Managerial Roles in the Bangladeshi Garment Sector." mimeo.

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

Muraglia, Sammantha, Amanda Vasquez, and Jessica Reichert. 2020. "Conducting research interviews on sensitive topics." *Illinois Criminal Justice Information Authority (ICJIA)*.

Ortner, Juan, and Sylvain Chassang. 2018. "Making corruption harder: Asymmetric information, collusion, and crime." *Journal of Political Economy*, 126(5): 2108–2133.

Poi, Brian P. 2004. "From the help desk: Some bootstrapping techniques." *The Stata Journal*, 4(3): 312–328.

Prendergast, Canice. 2000. "Investigating Corruption." Working Paper, World Bank Development Group.

Raghavarao, Damaraju, and W.T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society*, B(41): 40–45.

Rosenfeld, Bryan, Kosuke Imai, and Jacob N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science*, 60(3): 783–802.

Rosenzweig, Mark R, and Christopher Udry. 2020. "External validity in a stochastic world: Evidence from low-income countries." *The Review of Economic Studies*, 87(1): 343–381.

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

Slepian, Michael L., and Edythe Moulton-Tetlock. 2019. "Confiding Secrets and Well-Being." *Social Psychology and Personality Science*, 10(4): 472–484.

Sumon, Mahmudul H, Ahmed Borhan, and Nazneen Shifa. 2018. "Garment Workers' Rights: Situation analysis in Dhaka, Gazipur, Narayanganj, and Chittagong." Manusher Jonno Foundation, Dhaka, Bangladesh.

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

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive questions in surveys." *Psychological bulletin*, 133(5): 859.

UC San Diego Center on Gender Equity and Health. 2019. ""Measuring MeToo: A National Study on Sexual Harassment and Assault." mimeo.

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, Jonathan P, and Nadja Schreiber 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(6): 960–970.

Vivalt, Eva. 2020. "How much can we generalize from impact evaluations?" *Journal of the European Economic Association*, 18(6): 3045–3089.

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

Online Appendix – Not for Publication

A Figures & Tables

Figure A.1: Survey Modules & Treatment Interventions

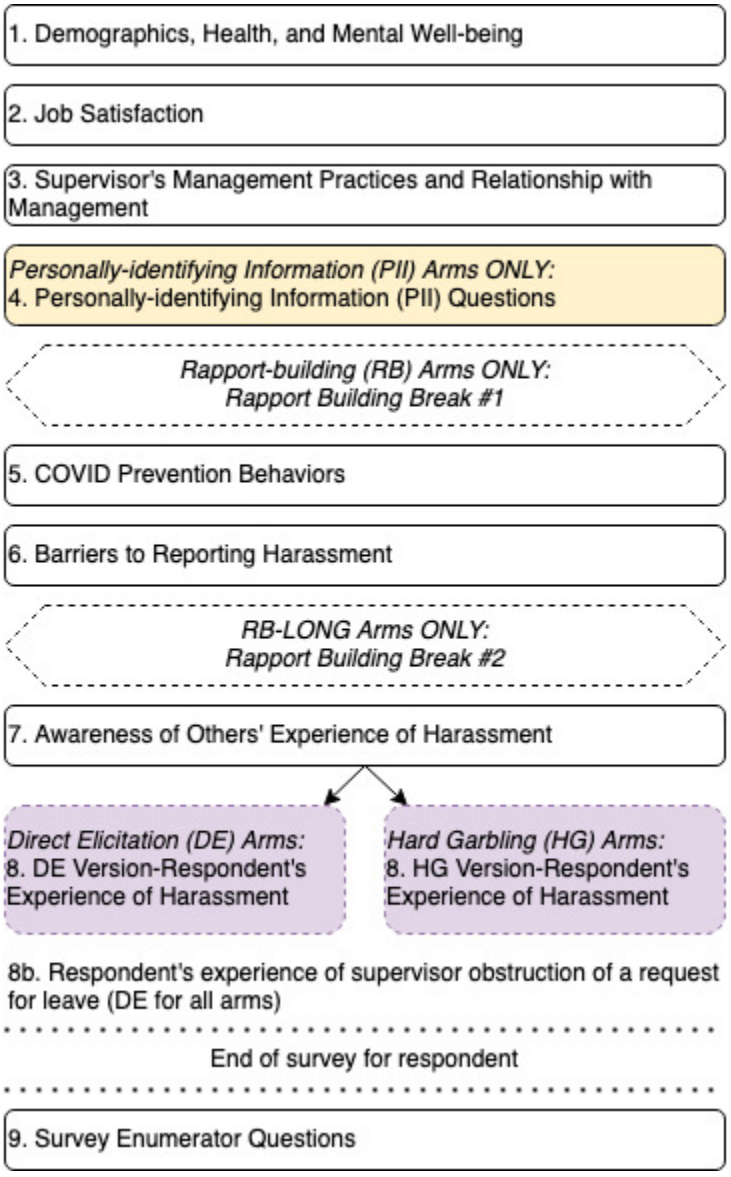
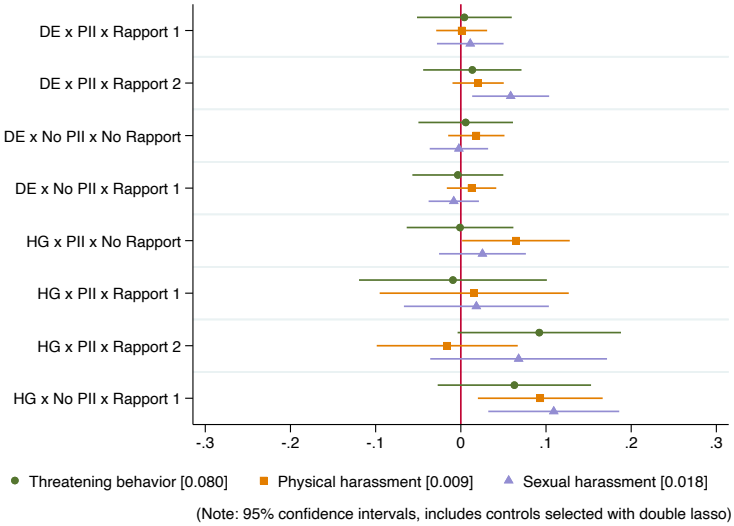
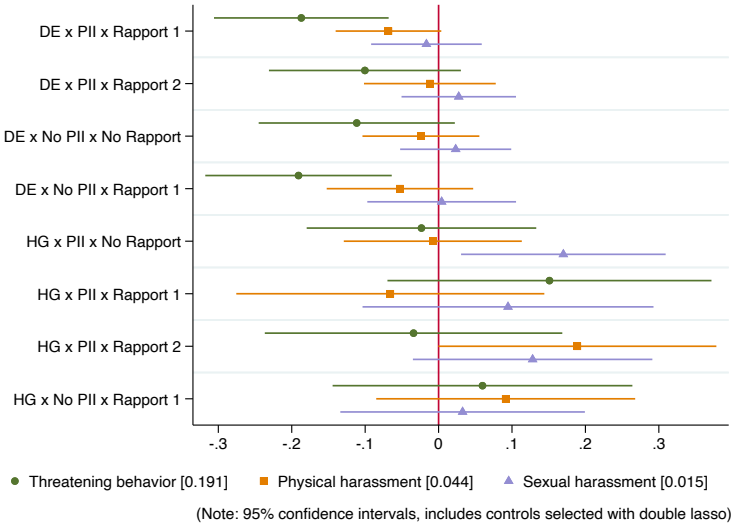


Figure A.2: Treatment effects by survey arm, separately by sex

(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 (3), 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.

Table A.1: 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.

Table A.2: 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.3: Balance tests: main treatment conditions

	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]
Team Size	57.244	57.428	58.227	56.580	57.395	57.180	0.051	0.091	0.090
	(20.550)	(20.331)	(20.421)	(20.437)	(20.546)	(20.203)	[0.670]	[0.419]	[0.524]
Team's Female Share	0.807	0.813	0.810	0.810	0.810	0.811	0.001	0.001	0.002
	(0.276)	(0.271)	(0.280)	(0.268)	(0.275)	(0.271)	[0.567]	[0.192]	[0.182]
Sewing Section	0.480	0.498	0.508	0.472	0.490	0.484	0.002	0.001	0.002
	(0.500)	(0.500)	(0.500)	(0.499)	(0.500)	(0.500)	[0.312]	[0.596]	[0.450]
Finishing Section	0.345	0.329	0.323	0.349	0.335	0.344	-0.001	-0.000	-0.000
	(0.476)	(0.470)	(0.468)	(0.477)	(0.472)	(0.475)	[0.642]	[1.000]	[1.000]
Washing Section	0.175	0.173	0.169	0.179	0.175	0.172	-0.001	-0.001	-0.002
	(0.380)	(0.379)	(0.375)	(0.383)	(0.380)	(0.378)	[0.286]	[0.309]	[0.271]
Position: Helper	0.180	0.167	0.175	0.173	0.167	0.191	-0.004	-0.004	0.024
	(0.384)	(0.374)	(0.380)	(0.379)	(0.373)	(0.393)	[0.777]	[0.816]	[0.155]
Position: Ironing/Folding	0.087	0.084	0.076	0.094	0.084	0.091	0.001	0.012	0.003
	(0.282)	(0.278)	(0.265)	(0.293)	(0.277)	(0.288)	[0.955]	[0.291]	[0.795]
Position: Operator	0.591	0.608	0.604	0.595	0.609	0.576	0.002	0.001	-0.026
	(0.492)	(0.488)	(0.489)	(0.491)	(0.488)	(0.495)	[0.899]	[0.951]	[0.183]
Position: Packer	0.040	0.048	0.041	0.046	0.044	0.045	0.007	0.005	0.001
	(0.196)	(0.214)	(0.198)	(0.210)	(0.204)	(0.207)	[0.347]	[0.499]	[0.952]
Position: Quality	0.102	0.092	0.104	0.091	0.097	0.097	-0.006	-0.015	-0.002
	(0.302)	(0.289)	(0.306)	(0.288)	(0.296)	(0.296)	[0.626]	[0.227]	[0.889]
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.4: Balance tests: no rapport, short rapport, and long rapport arms

	Mean / (SD)			Difference in means / [p-value]		
Variable	(0) No Rapport	(1) Short Rapport	(2) Long Rapport	(1) - (0)	(2) - (0)	(2) - (1)
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]
Team Size	58.227	56.673	56.377	0.087	0.091	0.037
	(20.421)	(20.280)	(20.802)	[0.511]	[0.535]	[0.823]
Team's Female Share	0.810	0.813	0.806	0.001	0.000	-0.001
	(0.280)	(0.267)	(0.271)	[0.191]	[0.490]	[0.498]
Sewing Section	0.508	0.476	0.464	0.000	0.002	0.002
	(0.500)	(0.500)	(0.499)	[0.794]	[0.318]	[0.475]
Finishing Section	0.323	0.349	0.350	0.001	-0.002	-0.003
	(0.468)	(0.477)	(0.478)	[0.670]	[0.318]	[0.184]
Washing Section	0.169	0.175	0.186	-0.001	-0.000	0.001
	(0.375)	(0.380)	(0.389)	[0.293]	[]	[0.401]
Position: Helper	0.175	0.179	0.161	0.002	-0.011	-0.026
	(0.380)	(0.384)	(0.368)	[0.904]	[0.590]	[0.236]
Position: Ironing/Folding	0.076	0.093	0.098	0.011	0.015	0.004
	(0.265)	(0.290)	(0.298)	[0.383]	[0.369]	[0.809]
Position: Operator	0.604	0.588	0.609	-0.006	0.012	0.024
	(0.489)	(0.492)	(0.489)	[0.766]	[0.631]	[0.371]
Position: Packer	0.041	0.054	0.030	0.012	-0.012	-0.021**
	(0.198)	(0.226)	(0.171)	[0.167]	[0.178]	[0.037]
Position: Quality	0.104	0.086	0.101	-0.019	-0.003	0.019
	(0.306)	(0.281)	(0.302)	[0.144]	[0.843]	[0.278]
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.5: 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.6: 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.7 on next page for p-values.

Table A.7: 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.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: Correlation of team-level reporting rates with witnessed harassment

Panel A: Witness Metric: Share of workers in team having ever witnessed harassment

	Witnessed Sexual Harassment		*Witnessed Physical Harassment*	
Correlations	DE	HG	DE	HG
ρ(Threat, Witness Metric)	0.002	0.031	0.011	0.156
	(0.083)	(0.088)	(0.078)	(0.095)
	[-0.139,0.186]	[-0.144,0.192]	[-0.131,0.179]	[-0.033,0.338]
ρ(Physical, Witness Metric)	-0.009	-0.045	-0.079	0.081
	(0.108)	(0.095)	(0.073)	(0.113)
	[-0.199,0.229]	[-0.238,0.129]	[-0.228,0.061]	[-0.146,0.303]
ρ(Sexual, Witness Metric)	0.059	0.039	0.044	0.005
	(0.093)	(0.093)	(0.097)	(0.092)
	[-0.094,0.267]	[-0.155,0.225]	[-0.143,0.241]	[-0.178,0.184]

Panel B: Witness Metric: Share of workers in team having witnessed at least 1 harassed victim

	Witnessed Sexual Harassment		*Witnessed Physical Harassment*	
Correlations	DE	HG	DE	HG
ρ(Threat, Witness Metric)	0.013	0.104	-0.032	0.194
	(0.084)	(0.095)	(0.079)	(0.097)
	[-0.139,0.193]	[-0.090,0.285]	[-0.168,0.147]	[0.001,0.361]
ρ(Physical, Witness Metric)	0.018	0.019	-0.016	0.125
	(0.095)	(0.098)	(0.075)	(0.108)
	[-0.180,0.199]	[-0.193,0.198]	[-0.150,0.149]	[-0.107,0.321]
ρ(Sexual, Witness Metric)	0.015	0.076	0.063	0.013
	(0.085)	(0.092)	(0.103)	(0.098)
	[-0.127,0.206]	[-0.129,0.243]	[-0.124,0.288]	[-0.191,0.184]

Notes: This table reports the correlation between team-level measures of witnessed harassment and team-level reporting rates of harassment using arms that collect PII. Panel A uses witness reports based on frequency of observed harassment. Panel B uses witness reports based on witnessing at least 1 victim being harassed. 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), following (Efron, 1987, Davison and Hinkley, 1997), implemented using Stata package **bootstrap** (Poi, 2004).

B Formal framework

This appendix formalizes the framework described in Section 3, including statements and proofs of the claims that statistics of intended reports from garbled surveys are identified from recorded data, and provide a less biased estimate of true underlying harassment statistics.

We note that it is not necessary to know a worker's team to compute S^r_V. In contrast, S^r_{PM}, $\mathsf{S}^r_{TV\geq k}$, and $\mathsf{E}^r_{2V|1V}$ 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^r_V.

Also, while harassment may be observed by workers other than the victim, this paper focuses on the reporting of one's own harassment status. We leave the question of witnesses' role in detecting and counter-acting harassment to future research. Getting victims to come forward strikes us as the main bottleneck, and we cannot confidently take a stand on how informed potential witnesses are. In our data (Table A.9), team-level measures of witnessed harassment are uncorrelated to team-level reports of own harassment. This means that information from third parties is not a good substitute for victims' reports.

B.1 Modeling a worker's reporting decision

Consistent with the model of Ljungqvist (1993), 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 or retaliation costs

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

Term 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 $\mathsf{SB}(0|1) = \mathsf{SB}(0|0) = 0$.[31] Recorded complaints only yield social benefits if they are associated with a misbehaving manager.

$\mathsf{RC}(\widetilde{r})$ is a reputational 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 normalized so that $K(0) = 0$; $\text{prob}(r_{i,a} = 1|\widetilde{r}_{i,a})$ is the posterior belief that worker i intended to submit a complaint $r_{i,a} = 1$ about manager a, conditional on recorded report $\widetilde{r}_{i,a}$.[32]

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

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

In turn, a harassed worker's payoffs are

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

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

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

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

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

[31] The assumption that $\mathsf{SB}(1|0) < 0$ implies that arbitrarily high garbling rates are not desirable.
[32] This formulation captures concerns over ex post retaliation and reputational damages.

Taking as given the share of null reports prob($r_{i,a} = 0$), it follows from (B.1) and (B.2) that increasing π increases a worker's propensity to send report $r_{i,a} = 1$. Equation (B.2) captures the fact that increasing π reduces the impact of a positive recorded report $\widetilde{r}_{i,a}$ on posterior beliefs that a worker intentionally complained. In addition, coefficient $(1 - \pi)$ in (B.1) captures the fact that increasing π shrinks the reputational or retaliation cost savings associated with sending a null report $r_{i,a} = 0$.[33] As a result, the left-hand side of (B.1) exhibits single crossing in the garbling rate π: whenever its value is negative, it is increasing in π.

Proposition B.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.

We note that the impact of garbling π on reporting is an equilibrium effect. For HG to be effective in the long run, respondents as well as managers must understand that reported complaints are not necessarily associated with intended complaints.

This implies Claim 1 as a corollary.

More generally, for statistics that are monotonic in $h_{i,a}$ (resp. $r_{i,a}$), increasing garbling rate π reduces the bias between statistics based on intended reports and statistics based on actual harassment. This motivates the use of intended reports obtained for higher garbling rates π to make inferences about statistics of interest S_V, S_{PM}, and $\mathsf{S}_{TV \geq k}$: intended reports from garbled surveys exhibit lower bias.

Because statistics $\mathsf{E}_{2V|1V}$ and $\mathsf{E}^r_{2V|1V}$ need not be monotonic, their case requires more nuance.

[33] Indeed, HG increases the welfare of workers who make a report, but can decrease the welfare of workers who don't by exposing them to retaliation with some probability. This is a design trade-off inherent to HG. This issue can be addressed by letting workers as a group weigh in on whether they wish a particular reporting scheme to be implemented. In our specific experimental setting, attrition patterns across HG and DE treatments do not suggest that HG caused a reduction in welfare among workers who chose not to complain. See Online Appendix E.3 for further discussion.

B.2 Proxying for $\mathsf{E}_{2V|1V}$

Because $\mathsf{S}^r_{TV\geq k}$ is increasing in reporting, it always underestimates statistics of true harassment $\mathsf{S}_{TV\geq k}$, and the bias is reduced by survey design steps that increase reporting, such as increasing garbling rate π.

$\mathsf{E}^r_{2V|1V}$ is not monotonic in reporting, and assessing the impact of survey design on bias requires a more involved analysis. We work under the following assumption.

Assumption B.1 (independent reporting)**.** *Conditional on harassment taking place, the rate q at which harassment is reported is independent of other team outcomes:*

$$prob(r_{i,a} = 1 | h_{i,a} = 1, (r_{j,a}, h_{j,a})_{j\neq i}) = prob(r_{i,a} = 1 | h_{i,a} = 1) = q$$

Recall that the rate q at which harassment is reported increases with π and decreases with p. Under Assumption B.1, the following corollary holds.

Lemma B.1. *(i) For q small, $\mathsf{E}^r_{2V|1V} < \mathsf{E}_{2V|1V}$ and $|\mathsf{E}_{2V|1V} - \mathsf{E}^r_{2V|1V}|$ is decreasing in q.*

(ii) For q approaching 1, $\mathsf{E}^r_{2V|1V}$ approaches $\mathsf{E}_{2V|1V}$.

Proof. Let $\gamma_k \equiv \text{prob}_\mu \left(\sum_{i=1}^L h_{i,a} = k \right)$. We have that

$$\text{prob}\left(\sum_{i=1}^L r_{i,a} \geq 2 \right) = \sum_{k=2}^L \sum_{m\geq 2}^k \gamma_k \binom{k}{m} q^m (1-q)^{k-m}$$
$$\text{prob}\left(\sum_{i=1}^L r_{i,a} \geq 1 \right) = \sum_{k=1}^L \sum_{m\geq 1}^k \gamma_k \binom{k}{m} q^m (1-q)^{k-m}.$$

For q small,

$$\text{prob}\left(\sum_{i=1}^{L} r_{i,a} \geq 2\right) \sim q^2 \sum_{k=2}^{L} \gamma_k \binom{k}{2} + o(q^2)$$
$$\text{prob}\left(\sum_{i=1}^{L} r_{i,a} \geq 1\right) \sim q \sum_{k=1}^{L} \gamma_k \binom{k}{1} + o(q),$$

so that $\mathsf{E}^r_{2V|1V} = O(q)$. This establishes point (i)
For q approaching 1, observe that

$$\text{prob}\left(\sum_{i=1}^{L} r_{i,a} \geq 2\right) \sim \sum_{k=2}^{L} \gamma_k + o(1)$$
$$\text{prob}\left(\sum_{i=1}^{L} r_{i,a} \geq 1\right) \sim \sum_{k=1}^{L} \gamma_k + o(1),$$

so that $\mathsf{E}^r_{2V|1V} \sim \mathsf{E}_{2V|1V}$. ■

Under Assumption B.1, Corollary B.1 shows that when the reporting rate conditional on harassment q is close to 0 then increasing q reduces measurement bias $|\mathsf{E}_{2V|1V} - \mathsf{E}^r_{2V|1V}|$. In addition, measurement bias vanishes for q approaching 1. The first case is empirically relevant and suggests that in our context, increasing reporting rates improves the validity of $\mathsf{E}^r_{2V|1V}$ as a proxy for $\mathsf{E}_{2V|1V}$. The second point implies that if $\mathsf{E}^r_{2V|1V}$ is monotonically increasing in q then it cannot overshoot $\mathsf{E}_{2V|1V}$ so that measurement bias diminishes as reporting rate q increases.

In principle, there exist harassment patterns for which $\mathsf{E}^r_{2V|1V}$ is not monotonically increasing in q. Whether $\mathsf{E}^r_{2V|1V}$ is monotonic in q or not turns out to depend on the distribution of harassment counts $(\gamma_k)_{k\in\{0,\cdots,L\}}$, with $\gamma_k \equiv \text{prob}_\mu(\sum_{i=1}^{L} h_{i,a} = k)$. As an illustration, we provide sufficient conditions under which $\mathsf{E}^r_{2V|1V}$ is decreasing in q as q approaches 1 below. In addition, we provide a data driven condition ensuring

that $\mathsf{E}^r_{2V|1V}$ is increasing in q.

Non-monotonicity for q approaching 1. We have that

$$\begin{aligned}
\text{prob}_\mu\left(\sum_{i=1}^{L} r_{i,a} \geq 1\right) &= 1 - \sum_{k=0}^{L} \gamma_k (1-q)^k \\
\text{prob}_\mu\left(\sum_{i=1}^{L} r_{i,a} \geq 2\right) &= \text{prob}_\mu\left(\sum_{i=1}^{L} r_{i,a} \geq 1\right) - \sum_{k=1}^{L} \gamma_k k q (1-q)^{k-1}.
\end{aligned}$$

For q approaching 1, we have

$$\begin{aligned}
1 - \sum_{k=0}^{L} \gamma_k (1-q)^k &= 1 - \gamma_0 - \gamma_1 (1-q) + o(1-q) \\
\sum_{k=1}^{L} \gamma_k k q (1-q)^{k-1} &= \gamma_1 - \gamma_1 (1-q) + 2\gamma_2 (1-q) + o(1-q),
\end{aligned}$$

so that

$$\begin{aligned}
\mathsf{E}^r_{2V|1V} &= \frac{\text{prob}_\mu\left(\sum_{i=1}^{L} r_{i,a} \geq 2\right)}{\text{prob}_\mu\left(\sum_{i=1}^{L} r_{i,a} \geq 1\right)} = \frac{1 - \gamma_0 - \gamma_1 - 2\gamma_2 (1-q) + o(1-q)}{1 - \gamma_0 - \gamma_1 (1-q) + o(q)} \\
&= \frac{1}{1-\gamma_0} \left(1 - \gamma_0 - \gamma_1 - 2\gamma_2 (1-q)\right) \left(1 + \frac{\gamma_1}{1-\gamma_0} (1-q)\right) + o(1-q) \\
&= 1 - \frac{\gamma_1}{1-\gamma_0} - \frac{1}{1-\gamma_0} \left(2\gamma_2 - \gamma_1 + \frac{\gamma_1^2}{1-\gamma_0}\right) (1-q) + o(1-q).
\end{aligned}$$

This implies that whenever $2\gamma_2 - \gamma_1 + \frac{\gamma_1^2}{1-\gamma_0} < 0$ then $\mathsf{E}^r_{2V|1V}$ is decreasing in q for q approaching 1, so that $\mathsf{E}^r_{2V|1V}$ is non-monotonic in q. This condition is non-empty for $L \geq 3$. For instance, it holds for $L = 3$ and $\gamma_0 = .05, \gamma_1 = .2, \gamma_2 = .05, \gamma_3 = .7$.

A data-driven test. In order to evaluate whether the increase in reporting due to hard garbling reduces measurement bias $|\mathsf{E}_{2V|1V} - \mathsf{E}^r_{2V|1V}|$, we need more information on $(\gamma_k)_{k \in 0, \cdots, L}$. We can use estimates of $p_k \equiv \text{prob}_\mu(\sum_{i \in I} r_i = k)$ (obtained from

Claim 2)) to impose restrictions on $(\gamma_k)_{k\in 0,\cdots,L}$. Specifically, $(q, p_k, \gamma_k)_{k\in 0,\cdots,L}$ must satisfy the system of equations

$$\forall k \in \{0, \cdots, L\}, \quad p_k = \sum_{k'=k}^{L} \gamma_{k'} \binom{k'}{k} q^k (1-q)^{k'-k} \tag{B.3}$$

For any given q, this is a triangular linear system of equations with unknowns $(\gamma_k)_{k\in 0,\cdots,L}$, which admits a unique solution. This allows us to evaluate the monotonicity of $\mathsf{E}^r_{2V|1V}$ as a function of q as follows:

1. Pick $q \in [0, 1]$, and solve (B.3) for $(\gamma_k)_{k\in\{0,\cdots,L\}}$.
2. If the inferred $(\gamma_k)_{k\in\{0,\cdots,L\}}$ form a probability distribution, evaluate whether the resulting function $\mathsf{E}^r_{2V|1V}$ is monotonic in q over the range $[0, 1]$.

In our data for all types of harassment, we estimate that $p_k = 0$ for $k > 3$. We impose the following regularizing constraint on estimates.

Assumption B.2.

$$\frac{p_3}{p_2} \leq \frac{p_2}{p_1}. \tag{B.4}$$

In our data, Assumption B.2 is non-binding for estimates of $(p_k)_{k\in\{0,\cdots,L\}}$ related to reports of threats and physical harassment, but is binding in the case of sexual harassment. Under Assumption B.2, our estimates of $(p_k)_{k\in\{0,\cdots,L\}}$ are reported in Table B.1.

These estimated values are such that $\mathsf{E}^r_{2V|1V}$ is increasing in q over the range $[0, 1]$ for all possible values of q. This implies that the increase in reporting rate q associated with hard garbling decreased measurement error $|\mathsf{E}_{2V|1V} - \mathsf{E}^r_{2V|1V}|$.

B.3 Identification from garbled reports

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.

Table B.1: Distribution of team-level count of intended reports

	Threats	Physical	Sexual
p_0	0.281	0.746	0.598
p_1	0.607	0.219	0.368
p_2	0.112	0.035	0.031
p_3	0	0	0.003
p_4	0	0	0
p_5	0	0	0

Estimates of $(p_k)_{k\in\{0,\cdots,L\}}$ obtained by applying Proposition B.2 and projecting the estimate onto the set of distributions satisfying (B.4).

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

$$\mathsf{S}_V^{\widetilde{r}} \equiv \frac{\frac{1}{n}\sum_{a,i\in M\times I}\widetilde{r}_{i,a} - \pi}{1-\pi}. \tag{B.5}$$

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 of 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

$$\widetilde{r}_{i,a} = r_{i,a} + (1 - r_{i,a})\eta_{i,a} \tag{B.6}$$

where $\eta_{i,a} \in \{0, 1\}$ is a Bernoulli random variable equal to 1 with probability π.

We are interested in recovering μ from $\widetilde{\mu}$ under three garbling protocols:

- 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 realized 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 $\sum_{i\in I} \eta_{i,a} = \pi L$.[34]

Proposition B.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.*

Proof. 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 param-

[34] 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.

eters $(p_k)_{k\in\{1,\cdots,L\}}$ 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

$$\begin{aligned}
&\forall k \in \{0,\cdots,g-1\}, \quad \widetilde{p}_k = 0 \\
&\forall k \in \{0,\cdots,L-g\}, \quad \widetilde{p}_{g+k} = \sum_{h=0}^{g} p_{k+h}\binom{k+h}{h}\binom{L-k-h}{g-h}
\end{aligned} \tag{B.7}$$

In general the system of equations (B.7) 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 (B.7) 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. ■

Note that the appropriate matrix M depends on the garbling scheme. 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.

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. This is illustrated in Table B.2. Even though the modal sampled team size is 7, no team in our sample has more than 5 recorded "Yes" responses across different harassment types.

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

Table B.2: 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.

precision of the estimator $\mathsf{S}_V^{\widetilde{r}}$, defined in (B.5), 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 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

$$\mathsf{Var}\left(\sum_{j=1}^{n} \widetilde{r}_j \middle| \mathbf{r}\right) = \mathsf{Var}\left(\sum_{j=1}^{n} r_j \eta_j \middle| \mathbf{r}\right) = \overline{r}\left(1 - \frac{\overline{r}n - 1}{n-1}\right)\pi(1-\pi)n.^{35} \tag{B.8}$$

Indeed,

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

Whenever mean reporting rate $\overline{r}$ is small, as is the case in our setting, equation (B.8) induces a significant reduction in the variance of the estimator $\mathsf{S}_V^{\widetilde{r}}$.

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 below. 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. $\pi = 2/10$).[36] Because team sample 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 malicious channels) are an especially high concern, then i.i.d. garbling may be preferred to (team-)blocked garbling. In our context, where leaks would only occur in

[36] See Online Appendix C.2 for details on HG implementation in our application.

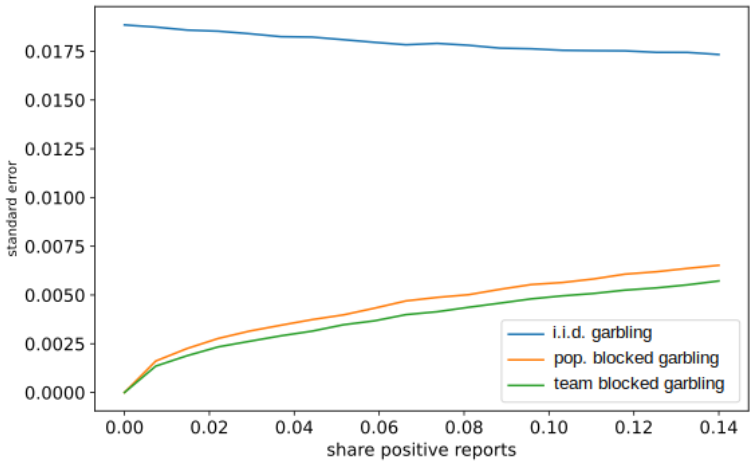
the case of an unforeseen breach of confidentiality, and we are especially interested in learning about team level statistics, the benefits of team-blocked garbling outweighed its costs.

Simulations. Figure B.1 reports findings from a simulated inference exercise. We consider data generating processes for harassment in which a fraction ρ of managers does not harass, a fraction $.6\times(1-\rho)$ harasses each worker under their span of control independently with probability .1, and a fraction $.4 \times (1-\rho)$ harasses each worker under their span of control independently with probability .2. For each value $\rho \in \{0, .05, .1, .15, \cdots, 1\}$ we simulate 1000 draws of a dataset of harassment outcomes for 100 teams of size 7. For each harassment dataset we associate three garbled reporting datasets by applying i.i.d. garbling, population-blocked garbling, and team-blocked garbling with garbling rate .2. For each such dataset of garbled responses we compute estimates of the share S_V of victims in the population of workers, as well as estimates of $\mathsf{S}_{TV\geq 2}$, the share of teams with at least 2 victims.

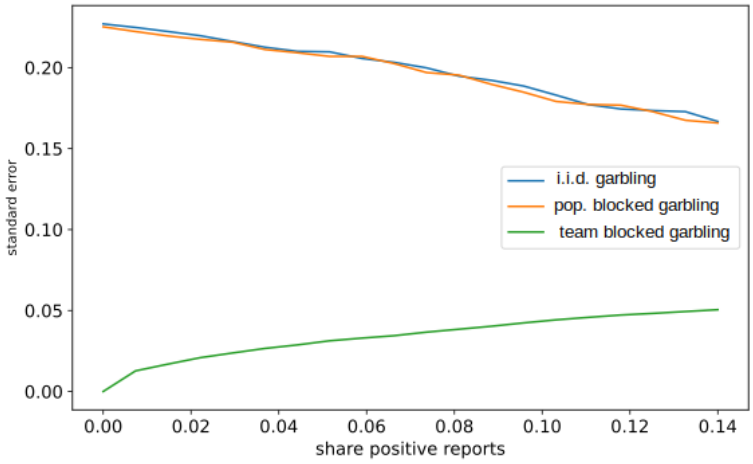
Figure B.1 reports the standard error of estimates as a function of the true underlying share of intended reports of harassment. It show that for estimates of S_V both population and team-blocked garbling reduce standard-errors, while only team blocked garbling reduces standard errors for team-level statistic $\mathsf{S}_{TV\geq 2}$.

Figure B.1: Estimated statistic & standard error

(a) Share of positive reports



(b) Share of teams with 2 or more victims



Notes: This figure plots the results of simulations of the standard error of estimators for a range of values of the estimated statistic. In our simulations, the sample size is 100 teams of size 7, which approximately corresponds to the sample size in our survey experiment's HG condition.

C Information on survey design & implementation

C.1 Design Choices

This section discusses key design choices in our field implementation of HG and how they might differ in other settings.

C.1.1 Alternative indirect survey response methods

The first design decision was to use HG, where noise is externally added to responses before any recording, rather than other garbling strategies. Many alternatives exist, including randomized response (Warner, 1965), the unrelated question approach (Greenberg et al., 1969), list experiments (Raghavarao and Federer, 1979), and crosswise RR (Blair, Imai and Zhou, 2015); we use HG over these for several reasons.

A common feature of these alternatives, including RR and LE, is that the respondent controls the report being sent — what we call *soft garbling* — so plausible deniability holds only if respondents comply with the instructions. Tourangeau and Yan (2007) and Chuang et al. (2021) highlight that respondents frequently do not comply and selectively deviate when instructed to submit a more sensitive answer; Chuang et al. (2021) develop tests of non-compliance and show it is large and problematic in both RR and LEs. Non-compliance has two consequences. First, it breaks standard inference, since the number of randomly induced sensitive responses is unknown, and recorded responses cannot be normalized to recover intended-response rates as in (B.5). Second, it undermines plausible deniability itself: as Chassang and Zehnder (2024) show in the lab, deniability unravels in equilibrium. This is especially salient in organizational settings, where participants repeatedly interact with the survey method — for instance, under LEs, if a higher number incriminates one's manager, workers may be told to systematically agree with the smallest plausible number of statements.

HG addresses both concerns. First, the survey software performs the garbling, so the nature of the noise is exactly 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. (2021), 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.

That said, HG does have some 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 with repeated interaction or with trusted surveyors (as in our application) but harder in one-shot, large-scale surveys, where RR's respondent-controlled noise may be preferable despite the compliance issue.

Relatedly, because the survey tool — rather than the respondent — performs the garbling, in enumerator-administered surveys, the enumerator typically learns the respondent's intended answer.[37] Under RR or list experiments, the enumerator only learns the garbled output. Three features of our setting mitigate this concern. First, the survey was phone-based, so enumerators were not physically present in respondents' workplaces or communities. Second, enumerators were employees of BIGD, a third-party research organization independent of both the firm and the respondents' communities. Third, our encoding scheme ensured that neither enumerator nor respondent knew which recorded responses had been garbled (see Section C.2). In settings without these features, this shortcoming may weigh more heavily and may favor soft garbling or HG implementations that eliminate enumerator observation, such as respondent-entered codes (see Section C.2,"Possible improvements").

[37]This is not required. For example, in in-person surveys, the enumerator could explain the question and then hand a survey tablet to the respondent to select their response and move to the next question. Survey logic could then prevent enumerators from returning to the completed question.

C.1.2 Garbling scheme

Garbling rate. A first key decision is the garbling rate. The garbling rate trades off plausible deniability against the information content of individual responses: 100% garbling offers complete deniability but makes recorded reports uninformative. As a compromise, we chose 20% of "No" responses to randomly be switched to "Yes." This choice is important in practice. If the garbling rate is set too low, then agents may respond by increasing rather than decreasing threats, and retaliation may end up happening more often in equilibrium. In our application, prior survey reporting rates, focus groups, and piloting suggested 20% was perceived as providing material protection, and a 1-in-5 rate is easy to explain.

Replication 2 tested two rates with different levels of plausible deniability and elicited workers' preferences over rates (see Section 7 and Online Appendix E).

Blocked or i.i.d. garbling. A second important design choice is whether to use i.i.d. or blocked garbling. As discussed in Section B.3, blocked garbling can considerably improve the precision of estimates by shrinking the sampling error associated with garbling. A drawback is that blocked garbling, especially at a small unit, diminishes the scope of plausible deniability. For instance, if we block at the level of teams with 10 members, ensuring that 2 "No" answers are mechanically switched to "Yes", then if more than 2 answers are recorded as "Yes", it is possible to infer that someone on the team must have reported harassment. Where group punishments are feasible, i.i.d. garbling or blocked garbling at a larger unit (e.g., sufficiently large teams, or population-level blocks) is preferable.

C.1.3 Follow-up actions

In our field implementation, workers' reports are used to provide statistics to the firm's senior management. The data was not used to target interventions toward individual managers. In effect, this corresponds to a setting with very little consequences following a recorded report of harassment. This also supports our assumption that there is no malicious reporting of good managers (see also Online Appendix

D.1).

Further applications may be interested in associating recorder reports of harassment with more meaningful follow-up actions; a suitable choice should satisfy:

- In equilibrium, plausible deniability is only effective if agents are aware that reports are garbled. For this reason, it is important to ensure that the reporting system's design is common knowledge.
- Follow-up actions need to be proportionate to noisy evidence. Lower-stakes responses – additional manager review, one-on-one HR training, rotating a worker to a different team– are appropriate; firing, demotion, or denial of promotion are not.
- Follow-up actions should not raise the attractiveness of malicious reporting. Again, this suggests actions that are not detrimental to good managers.

C.2 Hard-Garbling (HG) Implementation

In this section, we describe how we implemented HG. This was our and the BIGD's first attempt at implementing HG in a field-based survey, and the research team improved our implementation in our replications (see Online Appendix E). Below, we indicate ways that the primary experiment's implementation could be improved.

1. We wrote a Python script that generates encoded HG response codes corresponding to "Yes" and "No" intended responses and that decodes the encoded HG response codes as binary "Yes" and "No" garbled responses. One purpose of the encoding was to keep the enumerator and the respondent uncertain about whether their particular response had been garbled or not. Maintaining this uncertainty was intended to reduce participant stress.

    (a) To generate the encoded HG response codes, we set a seed and used the following parameters: 1500 codes, a 20% flip rate, 1-2 digit numeric encoded values, and blocking within sets of 10 consecutive surveys. Panel

A of Table C.1 displays an example of the resulting encoded response codes.

2. The research team ordered prospective subjects assigned to HG by production team and gender and assigned a unique HG subject id to each subject.

3. The research team provided the code assignments to the senior RA running the field implementation. If a prospective subject was unable to be reached or declined to participate, the senior RA reassigned their HG subject id to the next prospective subject from a randomly-ordered list of workers on the same production team and with the same gender.

4. The senior RA output spreadsheets for each enumerator with the HG codes assigned to their assigned sample. These spreadsheets were not retained after being distributed. After the survey's completion, the spreadsheet of encoded HG response codes used by the senior RA was destroyed. Enumerators were also instructed to destroy their spreadsheet after the survey's completion.

5. When conducting the survey, enumerators were instructed to input the numeric response code corresponding to the question number and the participant's intended response. Enumerators were required to input the response code twice, and the values were required to match to move to the next question.

6. The raw survey data provided to the research team had the encoded numeric values input for HG respondents' answers in the harassment module. We decoded the values and added the decoded, garbled HG responses to the survey data. Panel B of Table C.1 displays an example of how the decoding works. Encoded intended responses of "Yes" are always decoded as "Yes," but 2-of-10 encoded intended responses of "No" are decoded as "Yes" responses.

Table C.1: Examples of encoded HG response codes and decoding

Panel A: Encoded HG response codes

questions	Q1		Q2		Q3	
answers	Yes	No	Yes	No	Yes	No
subject_id						
0	77	90	30	57	28	15
1	3	35	79	14	39	22
2	83	38	92	54	94	44
3	65	99	1	36	49	54
4	31	5	42	40	0	91
5	32	59	83	64	34	94
6	59	95	4	28	59	63
7	51	82	65	59	30	64
8	10	46	94	28	97	34
9	76	11	23	99	54	71

Panel B: Example of coded and decoded HG response values

	coded resp.			decoded resp.		
subject id	Q1	Q2	Q3	Q1	Q2	Q3
0	90	30	15	0	1	1
1	35	79	22	0	1	0
2	38	92	44	0	1	0
3	99	1	54	0	1	1
4	5	42	91	0	1	0
5	59	83	94	1	1	0
6	95	4	63	0	1	0
7	82	65	64	1	1	0
8	46	94	34	0	1	0
9	11	23	71	0	1	0

Notes: The values displayed were generated using a different seed than in the original research implementation. In Panel B's example, Q1 and Q3 intended responses are set to "No" while Q2 intended responses are set to "Yes."

Possible improvements to this implementation:

- Prior to the survey, the BIGD did not believe that it was possible to program the garbling into the survey using SurveyCTO. This led us to develop the numeric code system. We have since determined that it is possible to program i.i.d. and blocked garbling directly into the survey using SurveyCTO. In Replication 1, we programmed i.i.d. garbling, and in Replication 2, we programmed blocked garbling into the survey.

- Programming the garbling directly into the survey offers sizable advantages: it is more secure, and it eliminates the possibility of enumerator data entry errors and non-compliance with the protocol, both of which occurred to a small extent in our implementation.[38] The primary issue that we faced was non-compliance with the numeric code system by one enumerator, which required us to drop all surveys completed by this enumerator. Specifically, the enumerator shirked by inputting the same response code for all questions for a given respondent. This deviation does not imply that these respondents' data were not hard garbled, as the garbling was encoded in the intended response codes, but these respondents' recorded answers could not be interpreted.

- There are some security weaknesses in our implementation. While we worked with the BIGD to uphold a strong commitment to data security, and the BIGD team took this very seriously and complied with all of our requests, a malicious team member could have tracked intended responses at the time of surveying and maintained their own database of intended answers. We would like to offer ways to mitigate this risk in future implementations. For instance, if respondents have access to a smart phone or tablet, they could directly provide responses to sensitive questions without being observed by an enumerator.

C.3 Costs of treatments

The potential costs of HG can broadly be categorized into three types: potential mental stress for participants who do not want to report harassment but worry that

[38] We programmed the garbling directly into the survey in our replications.

their (garbled) positive reports will be leaked; increased false positives in reporting data; and increased implementation costs. Starting with stress to participants about leaked reports, we first point out that, as emphasized in Section 3, garbling helps workers who want to report by providing plausible deniability, and may also benefit workers as a group by increasing reports of actual harassment.[39] Section 7.2.2 provides evidence that the net welfare effect is positive for the majority of workers: in replication 2, most workers in both DE and HG arms report a preference for HG over DE. Further suggestive evidence of the welfare benefits of HG comes from Table E.13, which shows a negative effect on job satisfaction among the DE group that was shown the HG system, consistent with disappointment that their organization is not actually using HG. Thus, in our setting, we view worries over leaked garbled reports as being counteracted by the welfare benefits to workers of increased ability to report with plausible deniability. However, this may not be true in all settings, and we encourage organizations using HG to assess respondents' preferences over HG versus DE.

Turning to increased false positives in reporting data, there are two main associated costs. First, as mentioned in the introduction and discussed in more detail in section C.1.3, the noise introduced into garbled reports of harassment will keep organizations from implementing severe responses, such as firing perpetrators. Instead, we believe that responses such as increased monitoring or training are appropriate. This fits with our collaborating apparel producer's goal for the data collection; see Appendix C.1.3.

Second, false positives increase noise in statistical estimation, as demonstrated in Section 3. We also demonstrate in Appendix B that iid garbling increases noise relative to blocked garbling. Accordingly, Table 5 displays higher standard errors for $\mathsf{S}^r_{TV\geq k}$ for $k \in \{1,2\}$ under HG, but not drastically so; standard errors for $\mathsf{S}^r_{TV\geq k}$ for $k \in \{1,2\}$ range from 19% to 73% higher under HG than under DE. Meanwhile, standard errors for $\mathsf{E}^r_{2V|1V}$ were actually smaller under HG, but we attribute this to

[39] One additional potential cost, which is outside the scope of the model in section 3, is that HG harms both reporters and non-reporters who want their organization to take direct, strong action against perpetrators.

noise from the bootstrap rather than to a conceptual reason why the HR estimates would actually be more precise. We emphasize that blocking is crucial in delivering the precision of these team-level estimates, but comes at the expense of lower privacy protection if answers are leaked or otherwise revealed indirectly.

As far as implementation costs, HG requires a survey team capable of understanding and explaining the garbling module to survey respondents, additional time in training and piloting, and additional survey time to explain the garbling module and ask comprehension questions. In our replications, we were able to program garbling directly into the survey software (iid in replication 1; blocked in replication 2), which eliminated the burden on the survey team to maintain and use codes and reduced the risk of enumerator shirking through improper recording of codes.

Still, implementation costs remain for HG compared to DE. On survey-team capacity, we had great confidence in BIGD's senior leadership to understand HG and explain it to enumerators and other staff, but this may not be true for every implementation partner. BIGD only considered enumerators among its top 10 to 15 percent of performers as potential recruits. BIGD's minimum educational qualification requirement for enumerators is a Higher Secondary School Certificate (HSC), and the senior RA reported to us that this level of education meant that enumerators could fully comprehend the concepts quickly. We also allowed for approximately 1 additional day of training and 1.5 days of piloting focused on HG modules, compared to a typical survey in this context that would use DE to ask about workplace harassment. While these costs were lower in replication 2, where garbling was programmed into the survey software, we still allowed approximately 1 additional day of training and 1.5 days of piloting for extra practice.

During the survey, Table A.8 indicates that the HG survey took 1.7 minutes longer than DE, or 4% of the total survey time of 41 minutes. In replication 2, where enumerators did not need to input codes but where we increased the number of comprehension questions from 2 to 3 and increased their difficulty (using multiple choice rather than true/false), the lengthening effect from the comprehension questions appeared to dominate: HG took 3.4 minutes longer than the version of DE where we did not explain HG afterwards in order to elicit preferences, or 11% of the

total survey time of 30 minutes in DE. Note that we worked with a different partner (Aureolin Research, Consultancy, and Expertise Development; ARCED) in the replications, which provides some reassurance that organizations other than BIGD have the capacity to implement HG as well.

A final implementation cost is the cost to the survey team of programming HG. Programming HG is straightforward for iid garbling but more complex for blocked garbling. To lower costs for other researchers who want to implement HG in a blocked design, we have developed software in partnership with Princeton's Data Driven Social Science Lab: the Safely Report software, https://github.com/princeton-ddss/safely-report.

Rapport building also comes with additional costs, both in terms of script design and in terms of training, piloting, and survey time. In designing the script, a senior Bangladesh-based RA led the design, in consultation with the authors, following guidance from the literature on GBV/IPV measurement (e.g. this guide) and drawing on insights from qualitative interviews with 20 workers lasting one to two hours each. We estimate that she spent approximately 140 hours developing the script at an hourly rate of $12.80 USD. We include the RB scripts in the Supplementary Materials, as well as guidelines on how to develop a RB script, with the goal of lowering the development costs for other researchers who want to use RB in their surveys. Enumerator training included approximately 0.5 days devoted to explaining the RB components of the survey and 0.75 days of piloting. In the survey, Table A.8 shows that the short RB treatment took an additional 5.5 minutes (13% of total survey time) and the long RB treatment took an additional 7.2 minutes (17% of total survey time).

Finally, as discussed throughout, the main cost of low PII is the loss of ability to calculate team-level statistics such as those reported in Table 5. Indeed, there is some evidence consistent with the premise that when asked about harassment, workers value the organization's ability to gather useful information at the team level: assignment to the low-PII treatment decreases job satisfaction at follow-up by 0.08 standard deviations (columns 5–8 of Table E.13).

C.4 Sampling & Randomization

Sampling. We conducted a stratified random selection of workers. We sampled workers from 4 types of production teams using the plants' employee lists: sewing lines, finishing teams, dry washing teams, and wet washing 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 5,948 eligible workers out of a workforce of 7,727 workers (77% of workers). The average size of teams in our survey is 53 workers (Table A.2). Within teams, we stratified workers by sex, which we identified based on name (male, female, uncertain).[40] In a small number of teams with few members of one sex, we aggregated members of this sex across teams to the smallest level that yielded a suitable group size. We selected 9 workers per stratum, 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. Because we had the plants' employee lists, we were able to replace workers who were unreachable or who declined to participate. We attempted to recruit 3,582 workers and 2,244 participated (63%). The main reason for non-response was that we were not able to reach workers by phone (73% of cases); of workers whom we reached, we surveyed 87%.[41] Conditional on survey start, the completion rate was 100%. Workers were not aware of their treatment assignment when deciding whether to participate.

Assignment to Treatment Arms. The unit of randomization is a worker, stratified by plant-production team and sex whenever possible. 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

[40] We categorized names as male or female for 99.7% of eligible workers. Among surveyed workers, the classification error rate was 2.76% for men (11 respondents) and 1% for women (17 respondents).

[41] We attempted to call workers 9 times to recruit them.

(2017) to address misfits across strata. To improve balance, we followed Banerjee et al. (2020), conducting 10 randomizations and selecting the one that performed best in terms of balance on two covariates available to the research team: tenure, and high versus low skill status.

C.5 Non-response

This section investigates whether non-response patterns threaten our inference. We obtained workers' phone numbers from the apparel manufacturer's HR department, so the main cause of our inability to reach workers (76% of non-response, Section C.4) was likely outdated phone numbers. Our checks of non-response patterns are consistent with this possibility.

We conduct several checks of whether non-response might be correlated with workers' experience of harassment or otherwise impact our estimates. Table C.2 predicts survey response with variables available in the producer's HR data, with team-level harassment rates estimated using data from the survey, and with measures of when we attempted to survey the respondent in the roll-out. Consistent with outdated phone numbers contributing to non-response, tenure is a positive, statistically significant predictor of non-response in all specifications. Consistent with the research team's experience conducting surveys of garment workers in Bangladesh, women are more difficult to reach, possibly because they were less likely to have their own phone upon joining the factory. Importantly, columns (6)-(9) show that team-level harassment rates are not significant predictors of survey response.

One may worry that workers communicated with each other, and that as workers learned about the survey, more workers who experienced harassment opted out. The research team assigned the survey roll-out order across teams to ensure that workers on the same floor were surveyed around the same time and that the survey rolled out across floors in an ordered manner. Column (10) shows that individuals on teams ordered to be surveyed later were no less likely to respond than those ordered earlier. Column (11) directly uses the first attempted call date at the worker-level to check for increasing non-response; again, it shows no difference in response rate across

workers attempted to be surveyed later versus earlier.

Assignment to HG, which increases willingness to report harassment in the main survey, does *not* reduce willingness to participate in the follow-up survey (Table C.3).

Table C.4 reports the correlations between the team-level response rate and the team-level reporting rates for harassment with DE, HG, and their team-level differences. Most correlations are small or zero and not statistically significant. For threatening behavior, the correlations are weakly negative, suggesting teams with higher rates of threatening behavior have slightly lower response rates.

Figure C.1 shows that response rates by treatment arm are very similar, which is unsurprising as workers were not aware of their treatment assignment when deciding whether to participate, and that response rates by treatment arm are stable over the survey's rollout.

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%; Figure C.2 displays the distribution of response rates across teams. Encouragingly, the distribution displays a short left tail, showing that there are no teams with very low response rates.

Table C.2: Predicting Participation in the Main Survey

	Responds to Main Survey										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Female	-0.0917***	-0.0989***	-0.0527**	-0.0575**	-0.0620**	-0.0575**	-0.0585**	-0.0574**	-0.0581**	-0.0632**	-0.0620**
	(0.0246)	(0.0219)	(0.0259)	(0.0268)	(0.0280)	(0.0273)	(0.0275)	(0.0273)	(0.0276)	(0.0281)	(0.0281)
Tenure	-0.0185***	-0.0127***	-0.0138***	-0.0153***	-0.0141***	-0.0153***	-0.0153***	-0.0152***	-0.0152***	-0.0139***	-0.0142***
	(0.0042)	(0.0046)	(0.0044)	(0.0044)	(0.0051)	(0.0044)	(0.0045)	(0.0045)	(0.0045)	(0.0052)	(0.0052)
Sewing Section			-0.0020	-0.0261	-0.0683**	-0.0263	-0.0265	-0.0268	-0.0262	-0.0666**	-0.0678**
			(0.0187)	(0.0234)	(0.0265)	(0.0238)	(0.0233)	(0.0234)	(0.0237)	(0.0268)	(0.0267)
Washing Section			0.0819***	0.0644*	0.0669*	0.0643*	0.0630*	0.0645**	0.0637*	0.0576	0.0651
			(0.0296)	(0.0337)	(0.0400)	(0.0329)	(0.0332)	(0.0328)	(0.0338)	(0.0450)	(0.0400)
Position: Helper				-0.0008	-0.0067	-0.0009	-0.0011	-0.0013	-0.0011	-0.0055	-0.0068
				(0.0294)	(0.0296)	(0.0285)	(0.0286)	(0.0287)	(0.0286)	(0.0297)	(0.0296)
Position: Operator				0.0499*	0.0516*	0.0499*	0.0499*	0.0499*	0.0499*	0.0520*	0.0507*
				(0.0300)	(0.0303)	(0.0293)	(0.0293)	(0.0293)	(0.0293)	(0.0303)	(0.0302)
Position: Packer				0.0500	0.0277	0.0498	0.0499	0.0494	0.0498	0.0290	0.0255
				(0.0511)	(0.0505)	(0.0530)	(0.0532)	(0.0531)	(0.0529)	(0.0507)	(0.0501)
Position: Quality				-0.0093	-0.0135	-0.0093	-0.0096	-0.0090	-0.0093	-0.0132	-0.0136
				(0.0352)	(0.0352)	(0.0353)	(0.0353)	(0.0354)	(0.0354)	(0.0354)	(0.0352)
Threatening Behavior Rate†						-0.0028			0.0099		
						(0.0474)			(0.0499)		
Physical Harassment Rate†							-0.0504		-0.0368		
							(0.0706)		(0.0723)		
Sexual Harassment Rate†								-0.0416	-0.0364		
								(0.0588)	(0.0647)		
Team order										-0.0002	
										(0.0006)	
First call date											-0.0001
											(0.0004)
Response rate	.6267	.6267	.6267	.6267	.6267	.6267	.6267	.6267	.6267	.6268	.6258
Factory FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Factory x Floor FE	No	No	No	No	Yes	No	No	No	No	Yes	Yes
Observations	3563	3563	3563	3563	3563	3563	3563	3563	3563	3561	3552

Notes: This table reports OLS estimates of predictors of response to the main survey. The dependent variable is an indicator for whether the worker participates in the main survey. For production section, the omitted section is Finishing. For position type, the omitted category is ironer/folder. Standard errors clustered by team are reported in round brackets in columns (1)-(5), (10)-(11), and bootstrapped standard errors by team are reported in columns (6)-(9). † This variable measures the team-level harassment rate estimated using survey responses from treatment arms that use HG and collect PII. *p <0.1; **p <0.05; ***p <0.01.

Table C.3: Effects of Survey Design on Attrition from Follow-up Survey

	Attrites from follow-up survey	
	(1)	(2)
HG Treatment	0.0015	0.0013
	(0.0114)	(0.0110)
Rapport Treatment	-0.0051	-0.0048
	(0.0120)	(0.0115)
Low PII Treatment	0.0095	0.0099
	(0.0131)	(0.0126)
Control Group Mean	.0611	.0611
Strata FE	Yes	Yes
PDS lasso controls	No	Yes
Observations	2143	2143

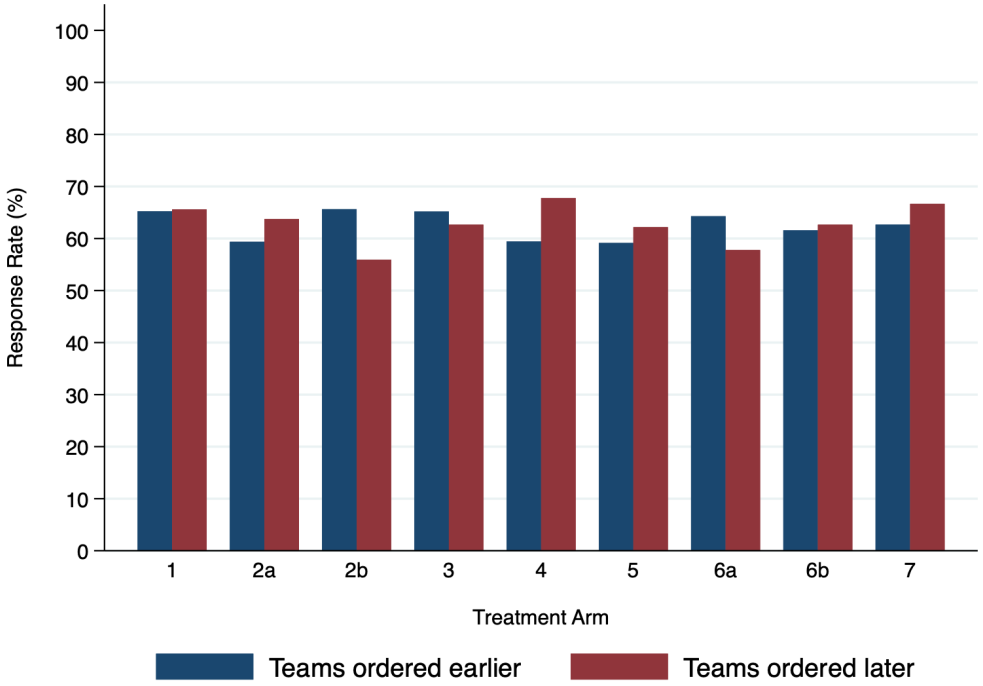
Notes: This table reports OLS estimates of treatment effects on workers' attrition in the follow-up survey (conducted two weeks after the main survey). Each column in the table reports the estimated coefficient from a separate regression. The dependent variable in each column is regressed on the treatment indicator 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 C.4: 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]

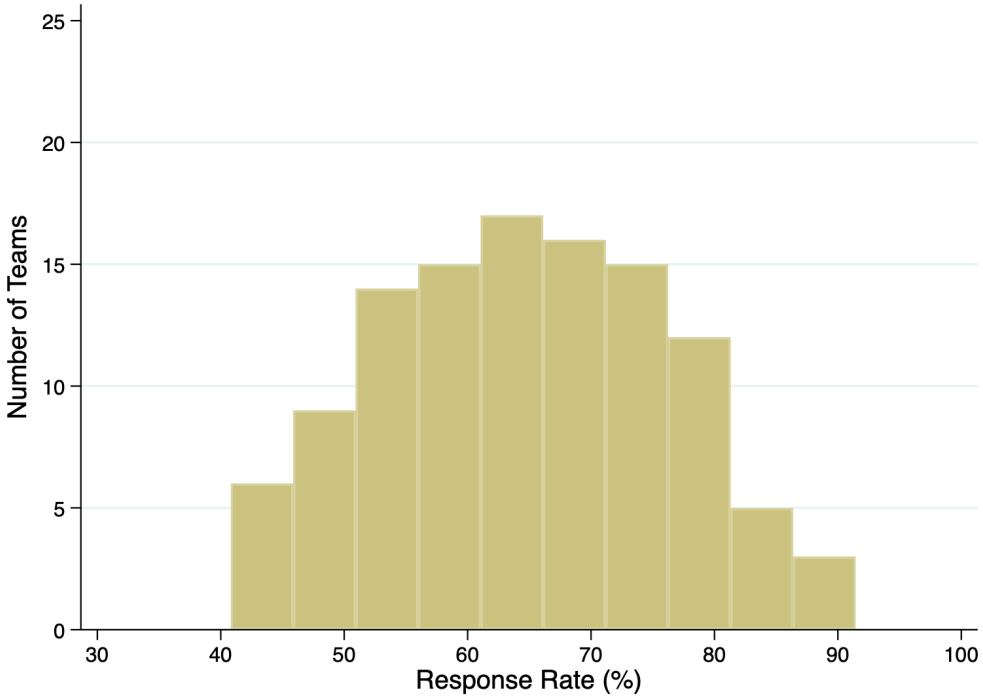
Notes: This table reports the correlation between the team-level response rate to the survey and the team-level reporting rates of harassment using arms that collect PII. Standard errors (in parenthesis) are computed from 1000 bootstrap replications, drawing samples of reporting rates at the team-level.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

Figure C.1: Main Survey Response Rate by Treatment Arm and Team Order



Notes: This figure presents the response rate to the main survey by treatment arm and "team order", which refers to the order in which a team was assigned by the research team to be surveyed first. Teams ordered earlier (later) are teams with team order below (above) median team order. Arm 1 is the control condition, DE × PII × No RB. Arm 2a is DE × PII × RB1. Arm 2b is DE × PII × RB2. Arm 3 is DE × No PII × No RB. Arm 4 is DE × No PII × RB1. Arm 5 is HG × PII × No RB. Arm 6a is HG × PII × RB1. Arm 6b is HG × PII × RB2. Arm 7 is HG × No PII × RB1.

Figure C.2: Distribution of Team-level Response Rates



Notes: This figure is a histogram of the number of teams with given response rates to the main survey.

D Robustness

D.1 Strategic misreporting by workers

False positives are a priori unlikely in our context given the low baseline reporting rates and the stigma that victims face. We test whether career concerns motivate false reporting by exploiting an informal eligibility cutoff for supervisor promotion (8+ years of schooling).[42,43] As men are much more likely to be promoted, we split the sample by sex. Table D.1 shows no consistent evidence that respondents above

[42]This analysis was not pre-specified and should be interpreted as exploratory.

[43]In a separate survey of supervisors employed by the apparel producer, 87% had at least 8 years of schooling, with a sharp jump at exactly 8 years.

this cutoff report more harassment under HG.

Theoretical guidance. Chassang and Padró i Miquel (2019) 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, 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.

D.2 Lack of comprehension in the HG condition

HG is a more complicated mechanism than DE, raising the concern that respondents may not understand the protective mechanism – that the 20% garbling rate reduces retaliation risk – especially given an average of 6.7 years of schooling in our sample. To assess this, we included two comprehension questions in the HG module.[44] Respondents answered these prior to being asked the questions about harassment, and survey enumerators explained the correct answers.

8.8% of HG respondents answered at least 1 comprehension question incorrectly, while 4.8% answered 2 incorrectly. Men and women answer incorrectly at somewhat similar rates: 9.6% (8.6%) of men (women) in HG answer at least 1 question incorrectly ($p = 0.685$), while 6.9% (4.3%) of men (women) answer 2 incorrectly ($p = 0.133$). Since reporting rates are weakly higher among respondents who answered at least one question incorrectly, we conservatively re-estimate our main results after recoding to "No" the *recorded* harassment answers from respondents who

[44]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.' "

failed any comprehension question.[45] Even under this conservative treatment, HG effects remain positive, large, and statistically significant for all harassment types (Table D.2, Panel A), with no differential attenuation by sex (Panel B).

D.3 General assurances in the HG condition

The HG script states that "our system is set up so that it's safe to report an issue" before explaining that garbling ensures that at least 1 in 5 answers is recorded as a complaint. This raises two alternative mechanisms for the results, which are that respondents trust this assurance directly without understanding plausible deniability, or that they interpret the script as a signal of the research team's general trustworthiness.

We note that while HG does admit stronger security assurances, they correspond to the real-life need of explaining the point and legitimacy of an unfamiliar survey methodology. In addition, even if respondents respond to assurances rather than to a precise understanding of plausible deniability, the mechanism is still necessary: in any ongoing monitoring setting, reassurances must be backed by an actual protective mechanism, both for ethical reasons and to maintain credibility as workers learn whether the assurances are real (Chassang and Zehnder, 2024).

In addition, we think it is unlikely that these assurances drive our results. First, all respondents were assured of a strong commitment to confidentiality as part of the 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.[46] Second, comprehension rates were high (discussed above), suggesting respondents comprehended the mechanics of how HG works.

Third, a placebo question asked under DE to all respondents – whether their supervisor obstructed a leave request – directly followed the harassment questions. If respondents under HG understood that plausible deniability applied only to the harassment module, this question should show no HG effect. Table D.3 confirms

[45] 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.

[46] The Supplementary Materials collect all scripts, survey instruments, and consent forms.

that HG and DE shares are indistinguishable.

Table D.1: 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 D.2: Main treatment effects, estimated with *recorded* response = "no" for respondents who answer less than 2 questions correctly

	Threatening behavior		Physical harassment		Sexual harassment	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Main effects						
HG Treatment	0.0343**	0.0347**	0.0373***	0.0386***	0.0351***	0.0360***
	(0.0145)	(0.0140)	(0.0122)	(0.0118)	(0.0118)	(0.0114)
Low PII Treatment	0.0119	0.0107	0.0264	0.0276*	0.0000	0.0021
	(0.0229)	(0.0223)	(0.0172)	(0.0166)	(0.0192)	(0.0189)
Rapport Treatment	0.0108	0.0136	-0.0098	-0.0081	0.0166	0.0168
	(0.0194)	(0.0188)	(0.0181)	(0.0174)	(0.0174)	(0.0168)
Control Group Mean	.0992	.0992	.0153	.0153	.0178	.0178
Panel B: Heterogeneity by sex						
HG Treatment × Female	0.0145	0.0147	0.0342***	0.0359***	0.0248*	0.0267**
	(0.0162)	(0.0156)	(0.0132)	(0.0127)	(0.0139)	(0.0136)
HG Treatment × Male	0.1237***	0.1217***	0.0519	0.0506	0.0771**	0.0769**
	(0.0416)	(0.0406)	(0.0341)	(0.0325)	(0.0341)	(0.0334)
Low PII Treatment × Female	0.0138	0.0137	0.0293	0.0308	0.0074	0.0094
	(0.0249)	(0.0244)	(0.0199)	(0.0191)	(0.0214)	(0.0209)
Low PII Treatment × Male	-0.0022	-0.0075	0.0161	0.0160	-0.0365	-0.0348
	(0.0496)	(0.0476)	(0.0423)	(0.0419)	(0.0380)	(0.0367)
Rapport × Female	0.0235	0.0263	-0.0151	-0.0124	0.0287	0.0290
	(0.0215)	(0.0207)	(0.0213)	(0.0205)	(0.0193)	(0.0185)
Rapport × Male	-0.0447	-0.0445	0.0127	0.0091	-0.0420	-0.0398
	(0.0461)	(0.0443)	(0.0332)	(0.0322)	(0.0414)	(0.0411)
Female	-0.0838	-0.0895	-0.0131	-0.0042	0.0582	0.0674
	(0.1069)	(0.1026)	(0.0752)	(0.0747)	(0.0754)	(0.0733)
Control Mean - Female	.08	.08	.0092	.0092	.0185	.0185
Control Mean - Male	.1912	.1912	.0441	.0441	.0147	.0147
p(HGxFemale - HGxMale)	[0.019]	[0.018]	[0.632]	[0.678]	[0.185]	[0.194]
p(NoPIIxFemale - NoPIIxMale)	[0.764]	[0.681]	[0.786]	[0.757]	[0.297]	[0.273]
p(RapportxFemale - RapportxMale)	[0.183]	[0.148]	[0.489]	[0.583]	[0.126]	[0.130]
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Lasso	No	Lasso	No	Lasso
Observations	2141	2141	2141	2141	2141	2141

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 D.3: 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)	(5)	(6)
HG Treatment	0.0088	0.0090	0.0070	0.0067	0.0058	0.0014
	(0.0184)	(0.0176)	(0.0156)	(0.0150)	(0.0204)	(0.0197)
Rapport Treatment	0.0095	0.0085	0.0093	0.0082	0.0274	0.0288
	(0.0191)	(0.0182)	(0.0160)	(0.0155)	(0.0210)	(0.0201)
Low PII Treatment	0.0255	0.0267	0.0151	0.0153	0.0222	0.0218
	(0.0212)	(0.0203)	(0.0178)	(0.0172)	(0.0229)	(0.0221)
Control Group Mean	.1513	.1513	.1298	.1298	.2723	.2723
Variable definition	N/A	N/A	1	1	2	2
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	Yes	No	Yes	No	Yes
Observations	1788	1788	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. Columns 1-2 drop workers who did not request leave (16.6% in DE and HG, respectively). Columns 3-4 use *Variable definition 1*, which codes respondents who report that they did not request leave in the previous year as no. Columns 5-6 use *Variable definition 2*, which codes respondents who report that they did not request leave in the previous year 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. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

E Evidence from Two Replications

This section presents evidence from two replication experiments in related settings.

Replication 1 Context & Design. Replication 1 uses the endline survey of Boudreau, Chassang and González-Torres (2022), which evaluates information escrows (Ayres and Unkovic (2012)) via a worker helpline serving around 768 Bangladeshi garment factories. The study is preregistered on the Open Science Foundation Registry. It randomly assigned an escrow reporting option to a subset of factories. With this option, workers who called the helpline could report a labor issue, such as a harassment case, and have the helpline hold their report and act on it only after another worker reported a similar issue. The study's primary outcome measures entail reports to the helpline.

An endline survey was conducted with workers from study factories, eliciting workers' experience of threats, physical and sexual harassment, and shouting as a non-sensitive question. The survey used the same scripts for DE and HG as our primary experiment, with a 20% garbling rate for HG (HG20). Workers from study factories were recruited through the community using a representative sampling method and seeking additional contacts through referrals. A total of 4,891 workers were approached, resulting in a sample of 2,951 completed surveys from 200 factories (55% women), with interviews staggered from September 2023 to November of 2024 (Table E.1). Two-thirds of respondents were randomly assigned to HG20; the HG randomization was not stratified, and the garbling was iid.

Replication 2 Context & Design. From December 2025 to March 2026, we approached a total of 8,076 workers and surveyed 5,490 (50% women) from 4 randomly-selected garment suppliers to a large international buyer that aimed to improve its supplier guidelines on resolving worker grievances (Table E.1). The survey is highly similar to our primary survey and was preregistered with an analogous analysis plan. As discussed in Section 7, the survey was designed with the aim of providing additional evidence on mechanisms through which HG impacts reports. The experiment tests for differences in reporting between higher and lower HG rates, assesses workers' preferences over different HG rates, as well as the welfare effects of the ability to report harassment using HG. Within each stratum (defined by factory, production team or section and sex), we randomly assigned respondents with equal probability

to one of 3 treatment conditions: DE, HG with a 20% garbling rate, and HG with a 1% garbling rate.

To the extent that workers value plausible deniability, and have a good understanding of the relative garbling rates, we expect larger effects for HG20 than HG01. On the other hand, while a 1% garbling rate provides materially less plausible deniability than 20%, given the very low reporting rate for physical and sexual harassment under DE (1-1.4%; Section 6), it is possible that it provides sufficient plausible deniability to induce reporting. To address concerns that HG is sensitive to security assurances, in this replication, we made the language of DE and HG as comparable as possible, including the same (truthful) security assurances in both arms (see Supplementary Materials Section 2 for scripts). We estimate effects on the two most sensitive primary outcome variables: physical and sexual harassment.[47]

We designed open-ended comprehension questions to capture comprehension more accurately. The survey also included an internal validity test: it asked about whether there was at least 1 fatal accident in the factory in the past 12 months. This is a verifiable outcome using data from the international buyer's internal records and media reports. We use this to assess whether HG leads to false positive reports, which we hypothesized would be due to higher rates of confusion under HG, but could in principle also be due to strategic misreporting. In the HG arm, we further asked all individuals what their preferred elicitation system is. We also asked preferences and comprehension questions to a random one-third of the DE-arm, to whom we explained the HG-mechanism after the own harassment module. This is to understand the extent to which experience with the HG mechanism affects preferences for it. We only asked it to one-third of DE respondents to ensure there was an uncontaminated control group for the follow-up survey measuring mental health and job satisfaction.

As part of the baseline demographic questions module, we elicited all respondents' understanding of fractions.[48]

[47] This was to keep the survey's run time within our budget.

[48] We ask whether a factory with 20% male workers has a greater number of male workers compared to a factory with 5% or 1% male workers and the same number of total workers.

E.1 Response rates, samples, and internal validity: Primary experiment & Replication 1 and 2

Table E.1 provides the response rates across all three experiments; Table E.2 provides reasons for not reaching workers. Table E.3 provides comprehension rates across experiments. Table E.4 provides summary statistics; Table E.5 shows that baseline characteristics are balanced in both replications.

E.2 Results: Replications 1 and 2

The regression model is

$$r_{is} = \alpha\, HG20_i + \beta\, HG01_i + \mu_s + \theta X_i + \epsilon_{is} \tag{E.1}$$

where r_{is} is the intended reporting outcome of interest for individual i in stratum s. $HG20_i$ and $HG01_i$ are hard-garbling treatment arms with different garbling rates respectively; $HG01_i$ is only included in regressions using Replication 2 data. We control for stratum fixed-effects μ_s; in replication 1, this is a factory-level stratification-by-treatment variable from the reporting escrow experiment (assignment to HG was not stratified); in 2, it is factory, production team or section, and sex. We present results including individuals' characteristics X_i selected using the PDS lasso. For individuals in the HG arm, we identify intended reports using garbled reports; intended reports are identified with a mean-zero heteroskedastic error term. For replication 1, we address this by reporting robust standard errors, which we cluster at the factory-level. For replication 2, we address this by reporting robust standard errors clustered by HG-block and treatment arm for HG20, with garbling blocked in groups of 5 workers within a stratum, and by individual for DE and HG01, which used iid garbling. We also present treatment effects broken down by gender of the respondent.

Main results. Section 7 discusses the main results, which are reported in Table 6. Table E.6 presents them without PDS lasso controls.

Does increased confusion under HG explain the results? Table E.7 reports the results of the internal validity test, reporting a worker fatality at the plant in the past year. No plants had a fatality, so there should not be a treatment effect of HG on the reporting of one. Consistent with this, HG does not affect reporting of fatalities, overall or by either sex. This provides additional evidence that increased reporting under HG is not simply due to greater confusion causing false positives, reinforcing HG's internal validity in Replication 2. It also supports the interpretation that respondents do not use HG to strategically make false accusations.

Are respondents who understand fractions and HG's plausible deniability feature reporting more due to them? Table E.8 replicates columns 5-6 of Table 6 for Replication 2. Columns (1) and (4) replicate the main results; (2) and (5) control for an indicator for answering the probability question incorrectly and its interaction with HG20 and HG01, respectively;[49] (3) and (6) drop respondents who answered the probability question incorrectly. The results for physical harassment show that the estimated difference between HG20 and HG01 is larger than for the average respondent, and a one-sided test of the difference between the coefficients is almost statistically significant. For women, the difference between the (smaller) HG20 and (larger) HG01 coefficients shrinks and is no longer statistically significant. For men, HG20's larger effect compared to HG01 grows and increases in statistical significance. In sum, the results point toward respondents who understand fractions being more responsive to HG20's additional protection compared to HG01.

Table E.9 replicates both replications' results when replacing the *recorded* responses of workers answering less than 2 questions correctly with 0. In replication 1 (2), this is 4% (12%) of respondents. Focusing on replication 2, for which the more challenging comprehension questions identify more respondents who do not fully comprehend HG, Panel A column (5) shows that the main results for physical

[49]Our PAP pre-specified that we would test for heterogeneity by comprehension of the HG mechanism. As 91% of respondents answered the probability question correctly (3 ppt higher for men than for women), we focus on the robustness of the main results to (1) controlling for an indicator for answering the probability question incorrectly and its interaction with HG20 and HG01, respectively, and (2) dropping respondents who answered incorrectly.

harassment survive. In Panel B, column (5) shows that the patterns of effects for men are unchanged. For women, the coefficients for HG20 and HG01 attenuate by similar amounts, and the effect for HG20 loses statistical significance. For HG01, while the effect for women is slightly attenuated, it remains highly statistically significant; evidently, women's increased reporting under HG01 is not solely due to lack of comprehension of how HG's protective mechanism operates.

Further exploration of HG's effects for women in Replication 2. To probe the mechanisms behind the similar effects of HG20 and HG01 on women's reporting of physical harassment, Table E.10 presents HTEs by education and Table E.11 presents them adding a control for gender and its interactions with the HG arms. The analyses compare individuals with above/below 8 years of schooling – the sample median in Replication 2 – and used in the primary experiment (Panel A), as well as above/below 9 years schooling – just above the sample mean in Replication 2 – (Panel B). These results should be interpreted as exploratory, as they were not prespecified. HG20's effect on reporting physical harassment is driven by more educated respondents, who do *not* report more under HG01, and the gap between HG20 and HG01 remains stable or grows once gender controls are added (Table E.11, column (5)). This is consistent with more educated respondents valuing plausible deniability and responding to it. For HG01, while its effect is driven by less educated respondents (Table E.10, column (5)), its effect for this group does not survive controlling for gender and its interaction with HG (Table E.11, column (5)). This indicates that it is not an education effect per se, but rather a women's effect.

Combined with the findings reported above that women's HG01 effect survives imputing zeros for *recorded* responses of respondents who do not fully comprehend HG (Table E.9) and controlling for understanding of fractions and its interaction with HG (Table E.8), and that there is no effect of HG on women's reporting of fatalities (under either rate, Table E.7), we interpret the totality of evidence as indicating that these mechanisms cannot be the sole reason(s) for women's increased reporting under HG01. Controlling for understanding of fractions (Table E.8), women respond similarly to HG20 and HG01, and we cannot identify what explains this. That said,

the aforementioned evidence is consistent with women's higher level of reporting under HG being due to their perceiving greater security under it.

E.3 Reporting & mental health and job satisfaction

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 (one) weeks after the primary (replication 2) experiment to test whether reporting improved workers' mental well-being and job satisfaction, measured using summary index variables following Anderson (2008). We report the variables comprising each index at the end of this section. We pool both survey samples to increase statistical power.

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

$$W_{is} = \delta Y_{is} + \rho W_{is}^0 + \theta X_i + \mu_s + \epsilon_{is} \tag{E.2}$$

where W_{is} is worker well-being in the follow-up survey for individual i in stratum s, Y_{is} are reports of physical or sexual harassment from the main survey, and W_{is}^0 is the baseline worker well-being, measured in the main survey. We control for data-source (primary experiment or replication 2) and stratum fixed-effects μ_s, individual demographic characteristics X_i selected using post-double lasso, and assignment to the RB and low PII treatments (set to zero for replication 2 respondents).[50] Since there is a possibility that HG directly impacts worker well-being (e.g., if workers prefer not to be automatically associated with a garbled report), we also report the reduced form effect in a regression equivalent to (1), with W_{is} as outcome (and controlling for W_{is}^0).

The results in columns 1-2 of Table E.12 show that assignment to HG has positive effects on mental health (although not statistically significant across the whole

[50]For our primary experiment, we pre-specified that we would use eqn. (1) as our first stage. But because RB and low PII are weak instruments in that experiment, we report results only using randomized assignment to HG as the instrument.

sample). HG20 has weakly statistically significant effects on mental health for men, consistent with HG assignment *not* harming welfare, and possibly slightly improving it. Columns 2-4 show that reporting harassment is not associated with differential mental health in the follow-up (controlling for mental health as reported during the main experiment). Columns 5-6 report the effects of reporting on mental health, estimated using 2SLS; the point estimates are large: 55% (52%) of a standard deviation for reporting physical (sexual) harassment; but they are imprecisely estimated and not statistically significant.

Table E.13 presents analogous results for job satisfaction. Mirroring the mental health results, assignment to HG has on average positive, although statistically insignificant, effects on job satisfaction. The 2SLS estimates show that reporting physical (sexual) harassment is associated with an increase of 1.2 (1.0) standard deviations in an index of job satisfaction, though again neither effect is statistically significant at conventional levels. 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.

Survey questions used to construct index variables: The survey instruments are included in the paper's Supplementary Materials. In the main survey, the questions that comprise each index are listed below; unless otherwise noted, the corresponding question in the follow-up survey is used.

1. Mental health: Section 1, questions 13 and 14.

2. Job satisfaction: Section 2, questions 1, 2, and 3 and Section 3, question 14 (main survey round only).

Table E.1: Response rates across all 3 experiments

	(1) Primary	(2) Replication 1	(3) Replication 2
Counts			
Workers attempted	3,582	4,891	8,076
Workers reached	2,575	3,967	5,646
Workers consented	2,501	3,245	5,542
Workers completed	2,244	2,951	5,490
Workers completed (101 excluded ex post, enum. checks)	2,143		
Answered harass. module, error in codes	3		5
Consented, not completed	257	294	52
Started survey, did not answer harass. module	0	22	2
Dropped just before reaching harass. module	0	22	0
Dropped after answering one own-harass q.	0	0	0
Answered some, not all harassment questions	0	7	1
Did not reach harass. module due to error in codes		0	13
Follow-up completed	1,987		5,186
Follow-up attrited	156		304
Rates			
% reached / attempted	71.9	81.1	69.9
% consented / reached	97.1	81.8	98.2
% consented / attempted	69.8	66.3	68.6
% completed / consented	89.7	90.9	99.1
% completed / reached	87.1	74.4	97.2
% completed / attempted	62.6	60.3	68.0
% follow-up completed / completed at baseline	88.5		94.5
Number of factories	2	200	4

Notes: This table reports the number of workers sampled and their response rates at different stages of the survey, for the Primary experiment, Replication 1 and Replication 2, followed by the number of distinct factories the workers who completed the survey were drawn from in each experiment. We report the number of workers we attempted to recruit; those we reached via phone (as opposed to not answering or someone else answering), those who consented to participate in the survey; those who completed all sections of the survey; and for the Primary experiment, those who completed all sections of the survey except for 101 excluded ex post (one enumerator did not comply with the garbling codes protocol) and 3 individuals we excluded due to errors in the codes. This is followed by the number of individuals who agreed to participate but did not complete the survey (consented, not completed), which we further break down to understand at which point they skipped the survey. We also report the number of workers who completed the follow-up survey, and the number of workers who did not, among all who completed the survey and we did not drop ex-post.

Table E.2: Call status (Original, R1, R2)

	Primary		Replication 1		Replication 2	
	N	%	N	%	N	%
Number of workers attempted	3,582	100.0	4,891	100.0	8,076	100.0
Completed	2,244	62.6	2,951	60.3	5,490	68.0
Completed partially			154	3.1	49	0.6
Not eligible	5	0.1	139	2.8	82	1.0
Not completed	1,333	37.2	1,647	33.7	2,455	30.4
Among those not completed:						
No answer	635	47.6	645	39.2	514	20.9
Answered, not by respondent	232	17.4	305	18.5	469	19.1
Number does not work	102	7.7			842	34.3
Answered, not completed	364	27.3	619	37.6	558	22.7
Refusal	0	0.0	78	4.7	72	2.9

Notes: This table reports the call status of workers sampled for the Primary experiment, Replication 1 and Replication 2. In the top panel, percentages are shares of all attempted workers; under , percentages are shares of workers who did not complete the survey. Partially completed surveys are identified as those for which the respondent answered some survey questions, but at least in one of the sections, no question was answered.

Table E.3: Comprehension and Enumerator Assessments across all 3 surveys

	Primary			Replication 1			Replication 2		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
Probability question correct	.	.	0	.	.	0	0.91	0.28	5465
HG rate correct	0.93	0.25	1021	0.97	0.17	1971	0.75	0.43	4250
HG no-leak question correct	0.93	0.25	1021	0.98	0.15	1971	0.91	0.28	4250
HG probability question correct	.	.	0	.	.	0	0.82	0.39	4250
≥ 1 correct questions	0.95	0.21	1021	0.99	0.11	1971	0.99	0.11	4250
≥ 2 correct questions	0.91	0.28	1021	0.96	0.20	1971	0.88	0.32	4250
≥ 3 correct questions	.	.	0	.	.	0	0.61	0.49	4250
Has questions about HG	0.03	0.17	1021	.	.	0	0.09	0.28	4250
Prefers HG	.	.	0	.	.	0	0.61	0.49	4198
Prefers HG-20%	.	.	0	.	.	0	0.85	0.36	4179
Prefers HG-1%	.	.	0	.	.	0	0.11	0.31	4179
Prefers HG-20% (cond. on pref. HG)	.	.	0	.	.	0	0.87	0.34	2558
Trusts answers are private	.	.	0	.	.	0	0.91	0.29	5465
Enum. asses. - Good Comprehension	0.92	0.26	2143	.	.	0	.	.	0
Enum. asses. - Excellent Comprehension	0.48	0.50	2143	.	.	0	.	.	0
Enum. asses. - Comfortable	0.96	0.19	2143	0.97	0.17	3198	0.95	0.23	1216
Enum. asses. - Completely Comfortable	0.79	0.41	2143	0.77	0.42	3198	0.77	0.42	1216
Enum. asses. - Trusting	1.00	0.07	2143	1.00	0.06	3198	0.98	0.13	1216
Enum. asses. - Completely Trusting	0.83	0.37	2143	0.91	0.29	3198	0.84	0.37	1216
Enum. asses. - Patient	0.76	0.43	2143	0.93	0.26	3198	0.89	0.31	1216
Enum. asses. - Completely Patient	0.76	0.43	2143	0.83	0.37	3198	0.80	0.40	1216
Enum. asses. - Honest	0.92	0.27	1971	0.96	0.20	3130	0.97	0.18	1168

Notes: Summary statistics of respondents' HG comprehension and enumerators' end-of-survey assessment for all three survey experiments (Primary experiment, Replication 1, and Replication 2). Each set of three columns reports the mean, standard deviation, and number of observations for each sample. All variables are 0/1 indicators. Blank cells indicate that the variable is not collected in that sample. The first variable aims to proxy for respondents' understanding of probabilities at baseline, and is asked to all respondents in Replication 2. It is an indicator for correctly identifying a factory with 20% male workers as having a greater number of male workers compared to a factory with 5% or 1% male workers and the same number of total workers. The following 11 rows measure respondents' answers to comprehension and preference questions, and they are asked after explaining HG to those randomly assigned to the HG treatment arm, and in Replication 2 also to a random subset of resondents under the DE treatment arm, who are explained HG after completing the own harassment survey section. In Primary and Replication 1 the respondent chooses between different options the enumerator reads out loud; while in Replication 2 the questions are open-ended and are dichotomized here for presentation. The comprehension indicators include: whether the respondent recalls correctly the assigned garbling rate; whether they correctly identified that someone accessing garbled data does not know who actually answered yes; whether they identify correctly 20% as being the rate that flips a larger amount of responses among 20%, 5% and 1%; a dummy for confused respondent, based on at least 1, 2 or 3 correct HG comprehension questions, respectively; whether the individual had follow-up questions about HG. The preference questions asked in Replication 2 identify whether an individual prefers HG to DE, which garbling rate they prefer among 20%, 5% and 1%, and we also present an indicator for preferring 20%, conditional on stating that they prefer HG over DE. We proxy their trust at endline with an indicator for whether they are very reassured that their answers are private. The last 9 rows are asked to enumerators after completing the survey, regardless of treatment assignment (in Replication 2 we dropped enumerator assessment questions after 1216 surveys to cut survey length). They indicate whether the enumerator asseses the respondent to have good/excellent comprehension, feel comfortable/completely comfortable, trust/complete trust the enumerator, patient/completely patient as opposed to rushing through the survey, and honest in their answers.

Table E.4: Replications 1 and 2: Summary statistics

	Replication 1					**Replication 2**				
	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Age (years)	28.25	6.59	18.00	63.00	3239	27.64	5.97	18.00	60.00	5423
Education (years)	7.05	3.77	0.00	16.00	3248	8.04	3.39	0.00	16.00	5423
Tenure (years)	3.83	3.52	0.00	25.00	2972	3.32	3.08	-0.01	19.23	5423
Married (or prior marriage)	0.85	0.36	0.00	1.00	3248	0.76	0.43	0.00	1.00	5423
Has children	0.84	0.37	0.00	1.00	2746	0.82	0.38	0.00	1.00	860
Department - Sewing	0.48	0.50	0.00	1.00	2970	0.53	0.50	0.00	1.00	5423
Department - Finishing	0.14	0.34	0.00	1.00	2970	0.12	0.33	0.00	1.00	5423
Department - Cutting	0.07	0.25	0.00	1.00	2970	0.08	0.28	0.00	1.00	5423
Department - Washing	0.02	0.15	0.00	1.00	2970	0.01	0.08	0.00	1.00	5423
Department - Knitting	0.01	0.11	0.00	1.00	2970	0.03	0.18	0.00	1.00	5423
Department - Quality	0.04	0.20	0.00	1.00	2970	0.06	0.23	0.00	1.00	5423
Position: Helper	0.15	0.36	0.00	1.00	2972	0.13	0.34	0.00	1.00	1194
Position: Ironing/Folding	0.05	0.22	0.00	1.00	2972	0.09	0.28	0.00	1.00	1194
Position: Operator	0.61	0.49	0.00	1.00	2972	0.60	0.49	0.00	1.00	1194
Position: Quality	0.08	0.28	0.00	1.00	2972	0.10	0.29	0.00	1.00	1194

Notes: Summary statistics for workers. The left columns report the Replication 1 sample; the right columns report the Replication 2 sample. Position questions were later dropped in the Replication 2 survey to reduce survey length.

Table E.5: Replications 1 and 2: Balance tests

	Replication 1			Replication 2					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variable	HG-20%	DE	HG–DE	HG-20%	HG-1%	DE	HG20–HG01	HG–DE	DE expl.–base
Age (years)	28.20	28.39	-0.12	27.76	27.57	27.58	0.20	0.06	-0.00
	(6.52)	(6.79)	[0.67]	(6.15)	(5.90)	(5.87)	[0.23]	[0.67]	[0.99]
Education (years)	7.10	7.03	0.11	8.03	8.03	8.07	0.01	-0.03	0.03
	(3.81)	(3.70)	[0.45]	(3.33)	(3.39)	(3.44)	[0.86]	[0.72]	[0.85]
Tenure (years)	3.90	3.69	0.18	3.34	3.39	3.24	-0.06	0.09	0.05
	(3.58)	(3.39)	[0.18]	(3.09)	(3.12)	(3.01)	[0.38]	[0.17]	[0.66]
Ever married	0.85	0.84	0.00	0.81	0.82	0.81	-0.01	0.00	-0.00
	(0.36)	(0.36)	[0.88]	(0.39)	(0.39)	(0.39)	[0.26]	[0.85]	[0.91]
Has children	0.84	0.84	-0.01	0.85	0.84	0.77	0.03	0.07**	0.13**
	(0.37)	(0.37)	[0.41]	(0.35)	(0.37)	(0.42)	[0.33]	[0.02]	[0.02]
Prob. q. correct	–	–	–	0.90	0.92	0.92	-0.02**	-0.01	0.02
				(0.30)	(0.27)	(0.27)	[0.011]	[0.256]	[0.245]
Dept.: Sewing	0.48	0.49	-0.02	0.53	0.53	0.53	-0.00	-0.00*	-0.00
	(0.50)	(0.50)	[0.36]	(0.50)	(0.50)	(0.50)	[0.60]	[0.06]	[0.48]
Dept.: Finishing	0.13	0.15	-0.02	0.12	0.13	0.12	-0.00**	-0.00	-0.00
	(0.34)	(0.35)	[0.20]	(0.33)	(0.33)	(0.33)	[0.05]	[0.83]	[0.29]
Dept.: Cutting	0.07	0.07	0.00	0.08	0.08	0.08	-0.00	0.00	0.00
	(0.25)	(0.25)	[0.98]	(0.28)	(0.28)	(0.28)	[]	[]	[]
Dept.: Washing	0.02	0.02	0.01	0.01	0.01	0.01	-0.00	0.00	-0.00
	(0.15)	(0.14)	[0.23]	(0.08)	(0.09)	(0.08)	[]	[]	[]
Dept.: Knitting	0.01	0.01	0.00	0.03	0.03	0.03	-0.00	0.00	-0.00
	(0.12)	(0.11)	[0.52]	(0.18)	(0.18)	(0.18)	[]	[]	[]
Dept.: Quality	0.04	0.04	-0.00	0.06	0.06	0.06	0.00	-0.00	-0.00
	(0.20)	(0.20)	[0.82]	(0.23)	(0.23)	(0.23)	[]	[]	[]
Observations	1,992	981	3,248	1,805	1,809	1,809	3,614	5,423	1,809

Notes: Balance table for baseline covariates, for Replication 1 (columns 1–3) and Replication 2 (columns 4–9). Columns 1-3 and 4-6 report the means and standard deviations of each variable separately by treatment condition. Columns 3 and 7-9 report the differences in means between each treamtent condition, estimated from a regression of the covariate on the treatment indicator and stratification variables. Standard errors clustered by factory in Replication 1; by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents) in Replication 2; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.6: Replications 1 and 2: Main effects, without controls

	Replication 1				Replication 2	
	(1) Shouting	(2) Threats	(3) Physical	(4) Sexual	(5) Physical	(6) Sexual
Panel A: Main effects						
HG-20%	0.014	0.031*	0.026**	0.031**	0.013***	0.002
	(0.020)	(0.017)	(0.012)	(0.012)	(0.005)	(0.004)
HG-1%					0.012***	0.002
					(0.004)	(0.005)
Control Mean	0.285	0.119	0.013	0.010	0.007	0.013
p(HG20 = HG01)					[0.867]	[0.881]
p(HG20 > HG01)					[0.433]	[0.441]
Panel B: Heterogeneous effects by gender						
HG-20% × Female	0.035	0.006	0.026	0.029*	0.011*	0.006
	(0.027)	(0.023)	(0.017)	(0.017)	(0.006)	(0.007)
HG-20% × Male	-0.012	0.061**	0.026	0.032*	0.014**	-0.001
	(0.030)	(0.025)	(0.019)	(0.018)	(0.007)	(0.006)
HG-1% × Female					0.025***	-0.003
					(0.007)	(0.007)
HG-1% × Male					-0.002	0.006
					(0.005)	(0.007)
Control Mean (Female)	0.278	0.132	0.009	0.013	0.007	0.013
Control Mean (Male)	0.294	0.104	0.019	0.007	0.008	0.012
p(HG20xF = HG20xM)	[0.243]	[0.105]	[0.988]	[0.902]	[0.758]	[0.423]
p(HG01xF = HG01xM)					[0.003]	[0.317]
p(HG20xF = HG01xF)					[0.095]	[0.224]
p(HG20xM = HG01xM)					[0.035]	[0.308]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	No	No	No	No	No	No
Observations	2935	2935	2935	2935	5421	5421

Notes: This table reproduces the main results (Table 6) with no additional controls. It estimates of treatment effects on workers' reporting in Replications 1 and 2. Each column in the table reports the estimated coefficient from a separate regression. Columns 1–4 use the Replication 1 sample (shouting, threats, physical and sexual harassment in the past 10 months); columns 5–6 use the Replication 2 sample (physical and sexual harassment in the past year). Panel A effects pooled across gender; Panel B shows effects interacted with a dummy variable for each gender, and additionally includes a Female indicator as a control (a separate regressor in Replication 1; absorbed by the strata fixed effects, which include sex, in Replication 2). All regressions include escrow-treatment (Replication 1) or strata fixed effects defined by factory, production team or section and sex (Replication 2) and are adjusted as described in notes to Table 6. Standard errors clustered by factory in Replication 1; by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents) in Replication 2; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.7: Replication 2: Internal validity test using verifiable fatality outcome

	Fatal accidents	
	(1)	(2)
Panel A: Main effects		
HG-20%	0.005	0.005
	(0.006)	(0.006)
HG-1%	-0.000	-0.000
	(0.006)	(0.006)
Control Mean	0.032	0.032
p(HG20 = HG01)	[0.464]	[0.464]
p(HG20 > HG01)	[0.232]	[0.232]
Panel B: Heterogeneous effects by gender		
HG-20% × Female	0.009	0.009
	(0.010)	(0.010)
HG-1% × Female	0.000	0.000
	(0.010)	(0.010)
HG-20% × Male	0.001	0.001
	(0.008)	(0.008)
HG-1% × Male	-0.001	-0.001
	(0.008)	(0.008)
Control Mean (Female)	0.038	0.038
Control Mean (Male)	0.025	0.025
p(HG20: F = M)	[0.546]	[0.546]
p(HG01: F = M)	[0.930]	[0.930]
p(HG20 = HG01: female)	[0.424]	[0.424]
p(HG20 = HG01: male)	[0.842]	[0.842]
Fixed Effects	Yes	Yes
PDS lasso controls	No	Yes
Observations	5421	5421

Notes: This table reports OLS estimates of treatment effects on workers' reporting of having witnessed a fatal accident (placebo outcome) in Replication 2. Each column in the table reports the estimated coefficient from a separate regression. Panel A reports main effects; Panel B reports heterogeneous effects by gender (the female main effect is included but omitted from display). Column 1 omits and column 2 includes PDS lasso-selected controls (as described in notes to Table 6). All regressions include factory-strata fixed effects (factory, production team or section and sex). Standard errors are clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.8: Replication 2: Main effects, controlling for understanding of probabilities at baseline

	Replication 2					
	Physical			Sexual		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Main effects						
HG-20%	0.013***	0.017***	0.017***	0.002	-0.008	-0.008
	(0.005)	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)
HG-1%	0.012***	0.009**	0.010**	0.002	0.002	0.002
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Control Mean	0.007	0.007	0.007	0.013	0.013	0.012
p(HG20 = HG01)	[0.867]	[0.291]	[0.274]	[0.881]	[0.163]	[0.146]
p(HG20 > HG01)	[0.433]	[0.146]	[0.137]	[0.441]	[0.919]	[0.927]
Panel B: Heterogeneous effects by gender						
HG-20% × Female	0.011*	0.011	0.011	0.006	-0.012	-0.012
	(0.006)	(0.009)	(0.008)	(0.007)	(0.010)	(0.009)
HG-1% × Female	0.025***	0.021***	0.021***	-0.003	-0.004	-0.005
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
HG-20% × Male	0.014**	0.022***	0.022***	-0.001	-0.003	-0.003
	(0.007)	(0.008)	(0.008)	(0.006)	(0.008)	(0.007)
HG-1% × Male	-0.002	-0.001	-0.001	0.006	0.007	0.008
	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)
Control Mean (Female)	0.007	0.007	0.007	0.013	0.013	0.012
Control Mean (Male)	0.008	0.008	0.007	0.012	0.012	0.012
p(HG20xF = HG20xM)	[0.758]	[0.390]	[0.286]	[0.423]	[0.463]	[0.468]
p(HG01xF = HG01xM)	[0.003]	[0.015]	[0.018]	[0.317]	[0.244]	[0.205]
p(HG20xF = HG01xF)	[0.095]	[0.357]	[0.295]	[0.224]	[0.424]	[0.455]
p(HG20xM = HG01xM)	[0.035]	[0.011]	[0.007]	[0.308]	[0.233]	[0.182]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Controlling for interaction with incorrect prob.		Yes			Yes	
Dropped indiv. incorrect prob.			Yes			Yes
Observations	5421	5421	4961	5421	5421	4961

Notes: This table reports OLS estimates of treatment effects on workers' reporting in Replication 2. Each column in the table reports the estimated coefficient from a separate regression. Columns 1–3 report physical harassment and columns 4–6 sexual harassment. Panel A reports main effects; Panel B reports heterogeneous effects by gender. Within each outcome the first column (Columns 1 and 4) reproduces the Replication 2 specification of the main replication results table (Columns 5-6 of Table 6). The second column (Columns 2 and 5) controls for an indicator for answering the probability-comprehension question incorrectly and its interaction with each HG arm (and, in Panel B, also their interaction with gender), so the displayed HG effect is the effect for workers who answered correctly. The third column (Columns 3 and 6) drops workers who answered the probability-comprehension question incorrectly. The sample is held fixed across columns to workers with a non-missing probability-comprehension answer. All regressions include factory-strata fixed effects (factory, production team or section and sex) and PDS lasso-selected controls; the probability-comprehension controls are omitted from display. Standard errors are clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.9: Replications 1 and 2, Main effects, estimated with *recorded* response = "no" for respondents who answered less than 2 questions correctly

	Replication 1				Replication 2	
	(1) Shouting	(2) Threats	(3) Physical	(4) Sexual	(5) Physical	(6) Sexual
Panel A: Main effects						
HG-20%	-0.002	0.024	0.021*	0.031**	0.010**	0.002
	(0.020)	(0.017)	(0.012)	(0.012)	(0.005)	(0.005)
HG-1%					0.010**	-0.000
					(0.004)	(0.005)
Control Mean	0.285	0.119	0.013	0.010	0.007	0.013
p(HG20 = HG01)					[0.983]	[0.743]
p(HG20 > HG01)					[0.509]	[0.371]
Panel B: Heterogeneous effects by gender						
HG-20% × Female	0.017	-0.002	0.022	0.031*	0.006	0.007
	(0.027)	(0.023)	(0.016)	(0.017)	(0.007)	(0.007)
HG-20% × Male	-0.027	0.057**	0.021	0.032*	0.014**	-0.003
	(0.030)	(0.024)	(0.019)	(0.018)	(0.007)	(0.006)
HG-1% × Female					0.021***	-0.007
					(0.007)	(0.006)
HG-1% × Male					-0.001	0.006
					(0.005)	(0.007)
Control Mean (Female)	0.278	0.132	0.009	0.013	0.007	0.013
Control Mean (Male)	0.294	0.104	0.019	0.007	0.008	0.012
p(HG20: F = M)	[0.273]	[0.074]	[0.972]	[0.972]	[0.388]	[0.320]
p(HG01: F = M)					[0.011]	[0.151]
p(HG20 = HG01, Female)					[0.080]	[0.088]
p(HG20 = HG01, Male)					[0.052]	[0.219]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2935	2935	2935	2935	5421	5421

Notes: This table reports OLS estimates of treatment effects on workers' reporting in Replications 1 and 2, as a robustness check in which the reports are recoded to zero for workers who answer less than two questions correctly. Each column in the table reports the estimated coefficient from a separate regression. Panel A reports main effects; Panel B reports heterogeneous effects by gender (the female main effect is included but omitted from display). Each regression includes fixed effects, and PDS lasso-selected controls (as described in notes to Table 6). Robust standard errors in Replication 1; clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents) in Replication 2; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.10: Replications 1 and 2: Heterogeneity by education

	Replication 1				Replication 2	
	(1)	(2)	(3)	(4)	(5)	(6)
	Shouting	Threats	Physical	Sexual	Physical	Sexual
Panel A: High educ. has ≥ 8 years of schooling						
HG-20% × High Educ.	0.002	0.044*	0.013	0.029	0.016	0.007
	(0.029)	(0.024)	(0.018)	(0.018)	(0.011)	(0.010)
HG-20% × Low Educ.	0.022	0.019	0.039**	0.031*	0.008	-0.005
	(0.028)	(0.024)	(0.017)	(0.017)	(0.014)	(0.015)
HG-1% × High Educ.					0.008	0.002
					(0.006)	(0.006)
HG-1% × Low Educ.					0.017**	0.001
					(0.008)	(0.007)
Control Mean (High Educ.)	0.266	0.098	0.016	0.016	0.007	0.014
Control Mean (Low Educ.)	0.301	0.138	0.012	0.006	0.007	0.011
p(HG20, High vs Low)	[0.625]	[0.463]	[0.296]	[0.937]	[0.747]	[0.634]
p(HG01, High vs Low)					[0.322]	[0.883]
p(HG20 vs HG01, High)					[0.501]	[0.684]
p(HG20 vs HG01, Low)					[0.546]	[0.727]
Panel B: High educ. has ≥ 9 years of schooling						
HG-20% × High Educ.	0.000	0.043	0.005	0.026	0.027**	-0.002
	(0.033)	(0.026)	(0.020)	(0.020)	(0.013)	(0.013)
HG-20% × Low Educ.	0.021	0.027	0.039**	0.034**	0.000	0.006
	(0.025)	(0.022)	(0.016)	(0.016)	(0.010)	(0.011)
HG-1% × High Educ.					0.004	0.002
					(0.006)	(0.007)
HG-1% × Low Educ.					0.018***	0.002
					(0.007)	(0.007)
Control Mean (High Educ.)	0.251	0.084	0.015	0.006	0.007	0.012
Control Mean (Low Educ.)	0.303	0.138	0.013	0.013	0.007	0.013
p(HG20, High vs Low)	[0.626]	[0.632]	[0.187]	[0.743]	[0.215]	[0.698]
p(HG01, High vs Low)					[0.109]	[0.946]
p(HG20 vs HG01, High)					[0.095]	[0.738]
p(HG20 vs HG01, Low)					[0.114]	[0.684]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2935	2935	2935	2935	5421	5421

Notes: This table reports OLS estimates of treatment effects by education heterogeneity on workers' reporting in Replications 1 and 2. Each column in the table reports the estimated coefficient from a separate regression. Panel A defines highly educated respondents as those with at least 8 years of schooling (the median in replication 2), Panel B those with at least 9 years of schooling (secondary or more; just above the mean in replication 2). Each regression includes the education dummy main effect (omitted from displayed coefficients), fixed effects, and PDS lasso-selected controls (as described in notes to Table 6). Standard errors are clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.11: Heterogeneity by education: $\geq$ 8 years (Panel A) and $\geq$ 9 years (Panel B) of schooling, controlling for gender interactions

	Replication 1				Replication 2	
	(1)	(2)	(3)	(4)	(5)	(6)
	Shouting	Threats	Physical	Sexual	Physical	Sexual
Panel A: High educ. has $\geq$ 8 years of schooling						
HG-20% × High Educ.	-0.014	0.066**	0.015	0.031	0.016	0.002
	(0.034)	(0.027)	(0.022)	(0.021)	(0.010)	(0.009)
HG-1% × High Educ.					-0.003	0.006
					(0.006)	(0.008)
HG-20% × Low Educ.	-0.000	0.053	0.043*	0.033	0.009	-0.012
	(0.039)	(0.034)	(0.025)	(0.023)	(0.019)	(0.019)
HG-1% × Low Educ.					0.001	0.007
					(0.009)	(0.009)
Control Mean (High Educ.)	0.266	0.098	0.016	0.016	0.007	0.014
Control Mean (Low Educ.)	0.301	0.138	0.012	0.006	0.007	0.011
p(HG20, High vs Low)	[0.739]	[0.707]	[0.278]	[0.920]	[0.778]	[0.578]
p(HG01, High vs Low)					[0.713]	[0.943]
p(HG20 vs HG01, High)					[0.076]	[0.738]
p(HG20 vs HG01, Low)					[0.682]	[0.345]
Panel B: High educ. has $\geq$ 9 years of schooling						
HG-20% × High Educ.	-0.014	0.064**	0.008	0.029	0.026**	-0.004
	(0.036)	(0.029)	(0.023)	(0.022)	(0.012)	(0.011)
HG-1% × High Educ.					-0.005	0.006
					(0.006)	(0.008)
HG-20% × Low Educ.	-0.001	0.063**	0.045*	0.038*	-0.002	0.004
	(0.038)	(0.032)	(0.023)	(0.023)	(0.015)	(0.014)
HG-1% × Low Educ.					0.003	0.008
					(0.008)	(0.008)
Control Mean (High Educ.)	0.251	0.084	0.015	0.006	0.007	0.012
Control Mean (Low Educ.)	0.303	0.138	0.013	0.013	0.007	0.013
p(HG20, High vs Low)	[0.765]	[0.970]	[0.164]	[0.711]	[0.231]	[0.753]
p(HG01, High vs Low)					[0.392]	[0.834]
p(HG20 vs HG01, High)					[0.017]	[0.453]
p(HG20 vs HG01, Low)					[0.737]	[0.785]
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Controlling for interactions with gender	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2935	2935	2935	2935	5421	5421

Notes: This table reports OLS estimates of treatment effects by education heterogeneity on workers' reporting in Replications 1 and 2, adding gender and its interaction with the HG arms as controls. Each column in the table reports the estimated coefficient from a separate regression. Panel A defines highly educated respondents as those with at least 8 years of schooling (the median in replication 2), Panel B those with at least 9 years of schooling (secondary or more; just above the mean in replication 2). Each regression includes the education dummy main effect, gender and its interaction with the HG arms (omitted from displayed coefficients), fixed effects, and PDS lasso-selected controls (as described in notes to Table 6). Standard errors are clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.12: Primary experiment & Replication 2 (pooled): Impacts on Mental Health

	Reduced Form		OLS		2SLS	
	(1)	(2)	(3)	(4)	(5)	(6)
HG-20%	0.012					
	(0.010)					
HG-1%	0.015					
	(0.014)					
HG-20% x Female		0.000				
		(0.012)				
HG-20% x Male		0.030*				
		(0.016)				
HG-1% x Female		0.028				
		(0.019)				
HG-1% x Male		0.008				
		(0.019)				
Reported physical harassment			0.017		0.546	
			(0.013)		(0.443)	
Reported sexual harassment				-0.001		0.524
				(0.013)		(0.475)
Rapport-building (RB) arm	0.007	0.008	0.009	0.009	0.012	-0.000
	(0.013)	(0.013)	(0.013)	(0.012)	(0.017)	(0.018)
Low-PII arm	-0.008	-0.009	-0.010	-0.009	-0.019	-0.010
	(0.014)	(0.014)	(0.014)	(0.014)	(0.018)	(0.018)
DE-HG explained	0.020	0.022	0.010	0.010	0.012	0.010
	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)
Observations	7104	7104	7104	7104	7104	7104
Data-source FE	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Kleibergen-Paap Wald F	.	.	.	.	9.6	6.5
p(HG-20% = HG-1%)	[0.776]					
p(HG-20% x F = HG-1% x F)		[0.122]				
p(HG-20% x M = HG-1% x M)		[0.227]				
p(HG-20% x F = HG-20% x M)		[0.122]				
p(HG-1% x F = HG-1% x M)		[0.459]				
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Baseline control Mean	-0.009	-0.009	-0.009	-0.009	-0.009	-0.009

Notes: This table reports reduced form, OLS and 2SLS results for respondents' mental-health, measured in the follow-up surveys, in the Primary experiment and Replication 2. Columns 1-2 report reduced form effects of assignment to HG20 or HG01 (column 1), or of either treatment condition interacted with either sex (column 2). Columns 3-4 report OLS results and columns 5-6 report 2SLS results using the randomized assignment to the HG treatments as the instrumental variables for for the effect of reporting physical or sexual harassment, respectively. Reporting of physical and sexual harassment are based on the baseline reports, and are bindary indicators for DE, or adjusted by the garbling rate for HG arms. All regressions include data source/experiment fixed effects and strata fixed effects (factory, production team or section and sex), treatment indicators for the other treatment arms (rapport-building and reduced personally identifying information for the Primary experiment, or having HG explained to a subset of DE respondents for Replication 2), and controls selected post-double lasso. age, schooling and tenure (standardized and squared), ever-married, having children, as well as department indicators (sewing, finishing, cutting, washing) and position (helper, iron/folding, operator, quality, packing). Lasso-selected controls (union across PDS columns of the combined table): has children, washing deptartment. Standard errors clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table E.13: Primary experiment & Replic. 2 (pooled): Impacts on Job Satisfaction

	Reduced Form		OLS		2SLS	
	(1)	(2)	(3)	(4)	(5)	(6)
HG-20%	0.023					
	(0.020)					
HG-1%	0.042					
	(0.027)					
HG-20% x Female		0.038				
		(0.023)				
HG-20% x Male		-0.002				
		(0.035)				
HG-1% x Female		0.006				
		(0.036)				
HG-1% x Male		0.069*				
		(0.039)				
Reported physical harassment			0.016		1.181	
			(0.026)		(0.920)	
Reported sexual harassment				0.003		1.025
				(0.025)		(0.968)
Rapport-building (RB) arm	0.020	0.019	0.023	0.023	0.029	0.006
	(0.026)	(0.026)	(0.026)	(0.026)	(0.035)	(0.036)
Low-PII arm	-0.046	-0.045	-0.049*	-0.049*	-0.070*	-0.050
	(0.029)	(0.029)	(0.029)	(0.029)	(0.040)	(0.037)
DE-HG explained	-0.037	-0.040	-0.061*	-0.061*	-0.057	-0.061
	(0.040)	(0.041)	(0.037)	(0.037)	(0.038)	(0.038)
Observations	7104	7104	7104	7104	7104	7104
Data-source FE	Yes	Yes	Yes	Yes	Yes	Yes
Strata FE	Yes	Yes	Yes	Yes	Yes	Yes
Kleibergen-Paap Wald F	.	.	.	.	9.6	6.4
p(HG-20% = HG-1%)	[0.443]					
p(HG-20% x F = HG-1% x F)		[0.342]				
p(HG-20% x M = HG-1% x M)		[0.050]				
p(HG-20% x F = HG-20% x M)		[0.316]				
p(HG-1% x F = HG-1% x M)		[0.225]				
PDS lasso controls	Yes	Yes	Yes	Yes	Yes	Yes
Baseline control Mean	0.059	0.059	0.059	0.059	0.059	0.059

Notes: This table reports reduced form, OLS and 2SLS results for respondents' job-satisfaction, measured in the follow-up surveys, in the Primary experiment and Replication 2. Columns 1-2 report reduced form effects of assignment to HG20 or HG01 (column 1), or of either treatment condition interacted with either sex (column 2). Columns 3-4 report OLS results and columns 5-6 report 2SLS results using the randomized assignment to the HG treatments as the instrumental variables for for the effect of reporting physical or sexual harassment, respectively. Reporting of physical and sexual harassment are based on the baseline reports, and are bindary indicators for DE, or adjusted by the garbling rate for HG arms. All regressions include data source/experiment fixed effects and strata fixed effects (factory, production team or section and sex), treatment indicators for the other treatment arms (rapport-building and reduced personally identifying information for the Primary experiment, or framing HG explained to a subset of DE respondents for Replication 2), and controls selected post-double lasso. age, schooling and tenure (standardized and squared), ever-married, having children, as well as department indicators (sewing, finishing, cutting, washing) and position (helper, iron/folding, operator, quality, packing). Lasso-selected controls (union across PDS columns of the combined table): washing deptartment. Standard errors clustered by HG batch within a given factory, strata, treatment arm (HG20 respondents) or by respondent (DE or HG01 respondents); $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.