Discussion Paper Series

IZA DP No. 18699

May 2026

De-biasing or *Backlash*? Confronting Prejudice Among Police Officers in India

Sofia Amaral
World Bank, IZA@LISER and CESifo

Kim Chaney
University at Buffalo

Victoria Kaiser
Bavarian Ministry of Economic Affairs

Nishith Prakash
Northeastern University, BREAD, CESifo, IZA@LISER, GLO, HiCN, and CReAM

Abhilasha Sahay
World Bank

IZA DP No. 18699 May 2026

De-biasing or *Backlash*? Confronting Prejudice Among Police Officers in India*

Abstract

Police officers' discretionary handling of gender-based violence (GBV) complaints is a critical and largely neglected barrier to justice in developing countries. We collaborate with the Madhya Pradesh Police in India, the second largest state, to conduct a lab-in-the-field randomized experiment in which 323 male and female officers participate, and study the effect of randomly confronting officers with evidence of their biased handling of a fictitious GBV case on officer behavior and attitudes towards GBV. We find no statistically significant average effect, but sharply divergent and robust responses by officer gender. Confronted female officers prioritize the victim's statement by 23 percentage points more than controls, a 27 percent increase relative to the control mean. Male officers exhibit a backlash: they deprioritize the victim's statement, elevate the offender's statement, and on a computerized stereotyping task assign significantly more negative stereotypes to GBV victims one week after confrontation. We find no effects on deeper attitudinal outcomes such as beliefs in the truthfulness of rape complaints. A likely explanation for the heterogeneous response is the stark difference in baseline bias: 72 percent of female officers display only mild bias, while 51 percent of male officers are strongly biased. Because policing is male-dominated, the average female officer perceives a work environment more biased than her own, and women are thus willing to de-bias their case handling while men are not. Interventions targeting officer bias must account for these gender-differentiated responses to avoid unintended consequences.

JEL classification

J16, J45, K42, K14, C93, D91, O12, O15

Keywords

prejudice confrontation, gender heterogeneity, gender-based violence, police bias, backlash, stereotype reduction, lab-in-the-field experiment, India

Corresponding author

Nishith Prakash
n.prakash@northeastern.edu

* We are grateful for the collaboration of the Madhya Pradesh Police, the Central Academy for Police Training (CAPT), Bhopal, Pawan Kumar Srivastava (Indian Police Service), and Praveen Vashistha (Indian Police Service). We are grateful to Prashant Bharadwaj, Gabriela Deschamps, Michael Kaiser, Paul Niehaus, C. Schmidt-Padilla, and Helmut Rainer, and participants at the Annual Meeting of the Society for Experimental Social Psychology (2022), North East Universities Development Consortium (2024), World Bank Gender Research in South Asia Seminar (2025), and Pacific Conference for Development Economics (2025) for valuable comments and suggestions. We thank Vishakha Wadhwani and Asmi Khushi for their outstanding fieldwork and research assistance, and Diego Ramirez Ramirez for excellent research assistance. We thank DAI Research and Advisory Services for data collection. The views expressed in this paper do not necessarily reflect those of the Madhya Pradesh Police or CAPT. This study received bioethics approval from the University of Connecticut under protocol X21-0091 and is registered in the AEA RCT Registry under AEARCTR-0008611. All errors and omissions are our own. Sofia Amaral and Victoria Kaiser gratefully acknowledge financial support from the Leibniz Association. Victoria Kaiser also acknowledges support from the Joachim Herz Foundation. Findings, interpretations, and conclusions expressed in this paper do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

1 Introduction

Two explanations dominate existing accounts of inadequate police responses to gender-based violence (GBV), a problem most acute in low- and middle-income countries.[1] The first points to officer attitudes: misogynistic beliefs, victim-blaming cultures, and norms that classify GBV as a private family matter lead officers to doubt victims' accounts, deny formal remedies, and reclassify criminal offenses as domestic disputes (Amaral et al., 2026; Belknap, 2010; García-Moreno et al., 2015). The second emphasizes institutional failures: absent protocols, weak supervisory oversight, and no consequences for misconduct predictably produce adverse responses. Both explanations are important, but they leave a gap. Even officers without strong gender-based hostility may apply habitual scripts that systematically disadvantage complainants (dismissing complaints, reframing offenses as private disputes, or dissuading women from filing) without awareness that their responses reflect bias. Correcting such behavior requires confrontation: an external mechanism that reveals to an officer the gap between their response and a norm they recognize as legitimate, generates the guilt needed to motivate self-regulation, and produces lasting behavioral change—though whether this guilt-to-rumination pathway operates as theorized depends on the characteristics of those confronted, and, as we show, the pathway is not confirmed in our setting. Since formal reporting is the primary pathway to deterrence, complaints that are dismissed or discredited not only dilute police responsiveness but undermine victims' trust and willingness to report (Belknap, 2010; García-Moreno et al., 2015), making the correction of officer bias a necessary condition for equal access to justice.

Confronting people with their prejudices is an effective tool for reducing bias and changing behavior, particularly in contexts of racism and sexism (Alesina et al., 2024; Mallett and Wagner, 2011).[2] The mechanism operates through a well-documented chain: confrontation generates negative self-directed affect, particularly guilt, which motivates prolonged rumination; this rumination establishes "cues for control" that trigger behavioral inhibition when similar occasions for bias arise in the future (Chaney and Sanchez, 2018; Czopp et al., 2006; Monteith et al., 2002). Chaney and Sanchez (2018) show that a single confrontation reduces stereotype application seven days later,

[1] Gender-based violence can include sexual, physical, mental, and economic harm, threats of violence, coercion, and manipulation. While GBV can affect people of all genders, rates of GBV against women and girls are particularly high (`https://ncadv.org/STATISTICS`). GBV takes many forms, including intimate partner violence, sexual violence, child marriage, female genital mutilation, and so-called "honor crimes" (Degener and Koster-Dreese, 1995). Reliable cross-country estimates are unavailable for most forms; yet Sardinha et al. (2022) estimate that 27 percent of ever-partnered women aged 15–49 have experienced intimate partner violence in their lifetime.

[2] Prejudice confrontations are defined as verbal challenges directed at those who commit a blatant, subtle, or unspoken act of discrimination (Chaney et al., 2015; Czopp et al., 2006).

with effects mediated by guilt-induced rumination rather than by conscious egalitarian motivation or prejudice suppression. Three features of this evidence, however, limit its relevance to our setting. First, it derives almost entirely from high-income countries where egalitarian norms are strong (Jayachandran, 2015). Second, it focuses predominantly on racial rather than gender-based bias, where the guilt-generating force of confrontation is weaker (Czopp and Monteith, 2003). Third, feedback that threatens self-image can trigger defensive responses [denial, downplaying, or dismissal (Howell et al., 2017)], potentially amplifying rather than reducing bias. Whether confrontation-based interventions produce durable behavioral change among police in low-income, non-egalitarian-norm settings remains an open empirical question.

We address this gap by examining prejudice confrontation in the context of GBV and Indian police officers, a setting where egalitarian norms are weak and the bias being targeted is gender-based rather than racial. To our knowledge, this is the first study to examine the impact of prejudice confrontation on state actors' biases around GBV in a workplace setting. We collaborated with the Madhya Pradesh Police, the second largest state in India by area, to design and implement a lab-in-the-field randomized controlled trial. Participating officers reviewed two cases, one GBV and one non-GBV, and completed a computer-based survey on how they would handle each. Officers were then randomly assigned to a treatment condition in which a senior officer confronted them, in a private in-person conversation, about their responses to the GBV case; control officers received neutral feedback. One week later, all officers reviewed a new GBV case and completed a computerized reaction-time task in which images of potential victims were presented and officers rapidly categorized descriptions that might or might not apply.

We examine whether confronted officers respond in a less biased manner to GBV cases one week later, and whether effects vary by officer gender. Female officers may be more responsive to confrontation given their shared gender identity with the predominantly female victims of GBV. At the same time, women constitute only approximately 8 percent of police officers in India, and may face workplace pressure to conform to prevailing norms that could attenuate their response.[3]

Behavioral change is measured one week after the confrontation in two ways: officer responses to a new GBV case, which capture deliberate case-handling decisions, and a reaction-time task, which captures automatic stereotype activation (Chaney and Sanchez, 2018; Czopp et al., 2006). Male targets in non-GBV trials serve as a baseline to isolate GBV-specific stereotyping from general

[3]National Crime Records Bureau data show that women constituted approximately 8 percent of police personnel in India in 2018, well below the government's own target of 33 percent representation (National Crime Records Bureau, 2019).

response tendencies. Reduced stereotype activation on this task would indicate that confrontation established "cues for control" (Monteith et al., 2002), reflecting deeper cognitive change than behavioral adjustment in structured scenarios alone.

We find no statistically significant average effect of confrontation, but the responses by officer gender are sharply divergent. Female officers in the treatment group prioritize the victim's statement by 23 percentage points more than control-group female officers, a 27% increase relative to the control mean. Male officers move in the opposite direction: treated males place significantly less weight on the victim's statement and significantly more weight on the offender's statement. These results are robust to multiple hypothesis testing corrections and alternative specifications.

A likely explanation for the heterogeneous response is the difference in baseline bias across groups (Blattman et al., 2023). Among female officers, 72% exhibit mild pre-treatment bias in GBV case handling; among male officers, 51% exhibit strong bias.[4] Confrontation leads female officers to de-bias their case handling, consistent with a setting in which they perceive prevailing workplace norms as more biased than their own responses, making them receptive to the feedback. Male officers exhibit backlash (Rudman and Fairchild, 2004), consistent with evidence that gender bias confrontations generate greater anger toward the confronting party and defensive responding than racial bias confrontations (Czopp and Monteith, 2003), and the effect is driven primarily by strongly biased officers. The results suggest that the share of strongly biased officers in a target group is likely to determine whether prejudice confrontation induces de-biasing or backlash, an important design consideration for scaling such interventions.

We next turn to the reaction-time task, which measures automatic stereotype activation when officers assess a woman reporting GBV. The task yields results consistent with the case-handling findings. Confronted male officers assign significantly more negative stereotypes to women reporting GBV than control-group male officers, confirming that the backlash documented in case handling extends to automatic stereotype activation. Female officers show no statistically significant change in stereotype assignment after confrontation. Assignment of negative stereotypes to targets in non-GBV cases does not differ by treatment condition for either male or female officers, establishing that the backlash effect among male officers is specific to GBV contexts rather than a general response to the confrontation (Rudman and Fairchild, 2004). Taken together, the case-handling and reaction-time results point to a consistent pattern: confrontation de-biases female officers in deliberate responses

[4]Strong bias is defined as, for example, filing a report against the GBV victim rather than the accused. Mild bias is defined as, for example, ranking an investigation of the victim among the top three of five investigative steps.

and produces no significant backlash in automatic responses, while male officers, particularly those with strong baseline bias, exhibit backlash across both measures (Czopp and Monteith, 2003).

We extend this literature in three directions. First, we test confrontation in a low-income, non-egalitarian-norm setting. Second, we target gender-based rather than racial bias, a domain where confrontation is predicted to generate weaker guilt and greater defensiveness (Czopp and Monteith, 2003). Third, we examine confrontation in a workplace context among state actors, where institutional hierarchy shapes both the delivery and reception of the intervention. Our results show that these features matter: confrontation de-biases female officers but triggers backlash among strongly biased male officers, a heterogeneity not documented in the existing laboratory literature.

Our paper also relates to evidence that cognitive behavioral therapy-informed interventions produce durable non-cognitive change in developing-country settings (Blattman et al., 2023). Unlike cognitive behavioral therapy, which builds internal skills through structured exercises, confrontation operates through guilt and rumination triggered by an external actor; our results speak to whether this externally driven mechanism can produce lasting behavioral change in a workplace setting.

The second body of literature concerns the drivers and prevention of GBV. Scholars have linked GBV to social norms (Bandiera et al., 2020; Green et al., 2020), cultural factors (Guarnieri and Rainer, 2021; Tur-Prats, 2019), labor market conditions (Aizer, 2010; Anderberg et al., 2016), liquidity constraints (Hidrobo et al., 2016), divorce legislation (Stevenson and Wolfers, 2006), and emotional cues (Card and Dahl, 2011). Evidence on prevention is considerably thinner. Police are the primary institutional gateway for GBV victims: their responses determine victims' willingness to engage with formal support services (Amaral et al., 2026; Palermo et al., 2014) and shape deterrence (Amaral et al., 2023). Dube et al. (2025) show that cognitive retraining reduces use of force and discretionary arrests among Chicago police officers; we differ in targeting GBV-specific bias through interpersonal confrontation in a developing-country context, and in documenting the gender heterogeneity that such confrontation produces.

We make three contributions to this literature. First, a low-cost confrontation intervention can shift GBV case-handling behavior among female officers in a developing-country setting, increasing victim statement prioritization. Second, the same intervention triggers backlash among strongly biased male officers, both in deliberate case handling — deprioritizing the victim and elevating the offender — and in automatic stereotype activation, a pattern that policymakers must account for when designing and scaling such programs. Third, the distribution of baseline bias within a target group appears to be the key design parameter determining whether confrontation de-biases or

backfires, with implications for how interventions should be tailored.

2 Conceptual Framework

Prejudice confrontation triggers behavioral change through a well-documented affective pathway (Chaney and Sanchez, 2018; Czopp et al., 2006; Monteith et al., 2002). When an individual is confronted with evidence that their behavior reflects bias, the gap between their action and their self-image as a fair person generates guilt, which motivates sustained rumination. This rumination establishes "cues for control": inhibitory signals that intercept habitual biased responses in future encounters (Monteith et al., 2002). Crucially, this mechanism is self-regulatory rather than attitudinal — it installs a monitoring process without necessarily revising underlying beliefs, so behavioral change can occur in the absence of attitudinal change, though, as we show, the guilt pathway itself is not confirmed in our affect data. The guilt pathway, however, is not unconditional. Feedback that threatens self-image can instead trigger defensive responding — denial, downplaying, or counterattack — particularly among individuals with strong prior attitudes or when the social cost of conceding bias is high (Czopp and Monteith, 2003; Howell et al., 2017). Whether guilt or defensiveness prevails depends on the characteristics of those confronted.

We follow Blattman et al. (2023) in treating the distribution of pre-existing attitudes as the key parameter governing receptivity. An officer with mild bias occupies a position close to the egalitarian norm the confronter invokes, making the feedback credible and guilt the more likely response. An officer with strong bias faces a larger and more threatening gap; conceding bias requires a more fundamental revision of self-image, raising the psychological cost of compliance and making defensiveness more likely (Howell et al., 2017). The share of strongly biased individuals in a target group thus determines whether the aggregate effect of confrontation is de-biasing or backfire-inducing.

This threshold interacts with the institutional environment in ways that predict the gender heterogeneity we document. In a male-dominated institution where approximately 8 percent of officers are women (National Crime Records Bureau, 2019), perceived workplace norms are set largely by male officers' behavior and attitudes. The average female officer, whose own bias is predominantly mild, perceives her work environment as more biased than her own responses; confrontation therefore reinforces a perception she already holds, making her receptive (Czopp and Monteith, 2003). Strongly biased male officers, by contrast, operate in an environment where their

attitudes are closer to the perceived norm; the confrontation message is both self-threatening and socially incongruent, raising the likelihood of backlash (Rudman and Fairchild, 2004).

These considerations jointly generate the predictions our results confirm. Confrontation produces no significant average effect because de-biasing among mildly biased officers and backlash among strongly biased officers offset one another in a mixed sample. Effects are sharply heterogeneous by gender because female officers are disproportionately mildly biased and operate in an environment where the confrontation message is socially congruent, while strongly biased male officers face a message that is both self-threatening and norm-inconsistent. And behavioral change in either direction occurs without attitudinal change, consistent with the mechanism operating through automatic inhibition or automatic defensiveness rather than conscious belief revision, though the reaction-time evidence for female officers is directionally consistent with automatic inhibition rather than statistically conclusive.

3 Experimental Design

3.1 Context

The experiment was conducted in Madhya Pradesh, India. The state recorded the highest number of rape cases among all Indian states in 2018, accounting for 16 percent of all cases nationally, and ranks third in dowry deaths (National Crime Records Bureau, 2019). The police are the primary institutional point of contact for GBV victims, yet officer quality in Madhya Pradesh is low relative to other states. In the Indian Police Foundation Citizen Satisfaction Survey on Smart Policing, the state scores below the national average on both fair and unbiased policing and integrity and corruption-free service, with an overall index score of 6.15 out of 10 (Indian Police Foundation, 2021). The civil police in Madhya Pradesh comprises roughly 9,000 officers at the ranks of Sub-Inspector (SI) and Assistant Sub-Inspector (ASI), who handle crime reports at the early stages of investigations; approximately 13% are female.[5] The police force is highly hierarchical, and senior officers command considerable deference from junior colleagues, making a confrontation delivered by a senior officer a potentially effective intervention in this context.

[5]Bureau of Police Research and Development, *Data on Police Organizations 2021* (Bureau of Police Research and Development, 2021). This figure refers to female officers at the SI and ASI ranks specifically; the overall share of women across all police ranks in Madhya Pradesh is lower at approximately 7 percent.

3.2 Participants

A total of 323 officers participated in the experiment, drawn from the districts of Bhopal, Harda, Hoshangabad, Raisen, Sehore, and Vidisha. All participants held the rank of Sub-Inspector (SI) or Assistant Sub-Inspector (ASI), the ranks that typically handle GBV reports at the early stages of investigations. Recruitment proceeded in two stages: station heads were first notified of the study and asked to provide lists of eligible officers, who were then contacted directly by phone. Of the 360 officers approached, 323 agreed to participate, a response rate of 90%; non-participation was attributable almost entirely to duty scheduling conflicts or local operational priorities. To limit spillovers between treated and control officers, we invited on average two officers per station.

The final analysis sample comprised 239 male (80%) and 58 female (20%) officers, aged 24 to 60 (mean age: 42), after excluding the approximately 8% of participants who did not return for the endline session one week later. Absences were work-related and attritors did not differ on baseline characteristics from those who remained. Female officers constitute 20% of the analysis sample, somewhat above their 13% share among SI and ASI officers in Madhya Pradesh, reflecting variation in availability and scheduling across sessions rather than deliberate oversampling. Officers' participation was counted as on-duty time and lunch was provided on site.

3.3 Randomization

Data collection took place in November and December 2021. Each experimental day ran two parallel sessions, one per available computer room. Officers present on a given day were divided into the two sessions by alternating assignment as they entered, a procedure chosen for logistical reasons rather than identification purposes. Within each session, individual assignment to treatment or control was executed automatically by the survey software on the computer at which each officer sat (see Figure 1). Table A1 confirms that randomization produced balanced treatment and control groups across officer characteristics.

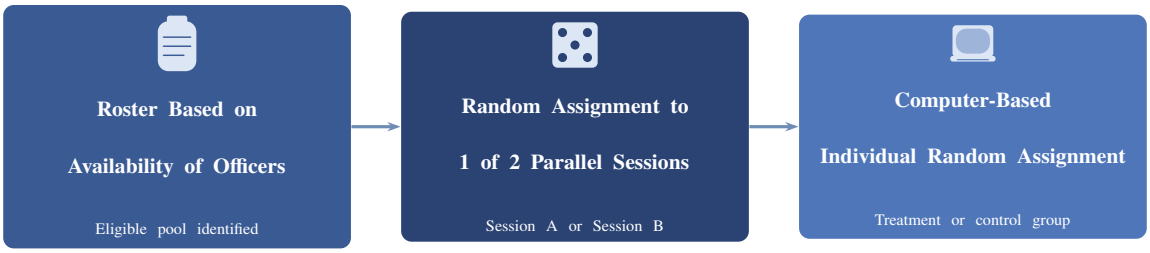

Figure 1: Three-stage randomization protocol. Officers are rostered based on availability, assigned to one of two parallel sessions, then individually randomized to treatment or control by the survey software.

3.4 Experimental Protocol

Baseline. The experiment relies on a between-subjects design. After giving consent, participants reviewed one neutral, non-GBV case followed by one GBV case in a computer-based survey. For each case, officers answered a structured set of questions: whether a crime had been committed, how they would proceed with the investigation step by step, and how they would rank different investigative actions in order of priority. From these responses we construct our key outcome variables: whether the crime is detected, whether the officer registers the complaint formally, whether the complaint is pursued or dropped, and how the officer prioritizes statements from the victim relative to the accused. We included two GBV cases, randomly assigned to subjects, to ensure the generalizability of findings following a confrontation: one involving domestic violence and one involving a dowry dispute. Assignment was executed automatically by the survey software. All cases were developed by the research team in consultation with senior officers from Madhya Pradesh Police: they were closely aligned with real cases reported to the police in the region, involved multiple parties and competing accounts to mirror the complexity of actual cases, and were extensively pilot-tested in separate officer samples before the main experiment (see Appendix C for details).

After completing the two cases, officers notified field staff and waited to enter a private one-on-one session with a senior officer (see Figure 2).[6] Officers in the control condition, and treatment officers who demonstrated no bias in their GBV case handling (10% of the treatment group), received neutral feedback unrelated to GBV.[7] The control condition reflects the realistic counterfactual of an officer who receives no GBV-specific feedback from a supervisor; while the neutral feedback

[6]We recruited four male senior officers to deliver the treatment, motivated by two considerations. First, the experiment was designed to mimic officers' interactions while on duty. Second, given the hierarchical nature of the police force, having external parties conduct the confrontation would have been challenging and likely ineffective. Senior officers were recruited in the same fashion as the SI and ASI officers and were not the direct managers or supervisors of participating officers, which mitigates potential concerns about retaliatory responses since officers operate in different jurisdictions.

[7]In a neutral feedback session, the senior officer informed participants that their case handling was generally satisfactory but that they had not paid adequate attention and care while responding. This feedback was not entirely content-free: the mild criticism it contains may itself have affected subsequent case handling in the control group, biasing our treatment effect estimates toward zero and making them conservative.

contains mild criticism unrelated to GBV, it does not draw supervisory attention to officers' GBV case handling, and the design therefore identifies the joint effect of confrontation and GBV salience relative to this baseline. A design with GBV-specific neutral feedback in the control would isolate confrontation from salience, but such a condition is difficult to construct credibly: any feedback that draws a supervisor's attention to an officer's GBV case handling without confronting bias would itself likely trigger reflection. To the extent that GBV salience alone drives de-biasing among female officers, it would have to operate selectively—producing no corresponding shift among control-group male officers exposed to the same neutral feedback—a pattern inconsistent with a pure salience explanation and more naturally attributed to the confrontation itself. Throughout the paper, we interpret estimated effects as the joint impact of confrontation and GBV salience; we return to this point in the conclusion.

Treatment officers who demonstrated bias received a confrontation adapted to the specific GBV case they had completed and to the extent of bias shown: officers with strong bias received a high-intensity confrontation and officers with mild bias received a low-intensity confrontation.[8] Officers were given the opportunity to respond; the interaction was capped at two minutes by field staff and was audio recorded, with a member of the research team present throughout.

After the feedback session, officers returned to the computer to complete a brief survey on self-directed affect (guilt and shame), affect directed toward the senior officer (anger), and measures of respect for the senior officer, self-assessed performance, and demographic characteristics (Chaney and Sanchez, 2018; Czopp et al., 2006). Officers were then dismissed and informed they would return in one week.

[8]In the high-intensity confrontation, the senior officer stated: "[. . .] it looks like you sided with the defendant's mother even though somebody else is the victim who came up with the complaint of domestic abuse. It seems like your actions were influenced by biased beliefs against women in society. Police officers cannot do their job based on their assumptions [. . .]". In the low-intensity confrontation: "[. . .] during the investigation you prioritized the statements of the defendant and his friends [. . .]. It seems that in this case your action was influenced by prevailing beliefs against women [. . .]. Police officers cannot do their job based on their perceptions." Senior officers were instructed to deliver all feedback in a neutral, non-aggressive, and educational manner.

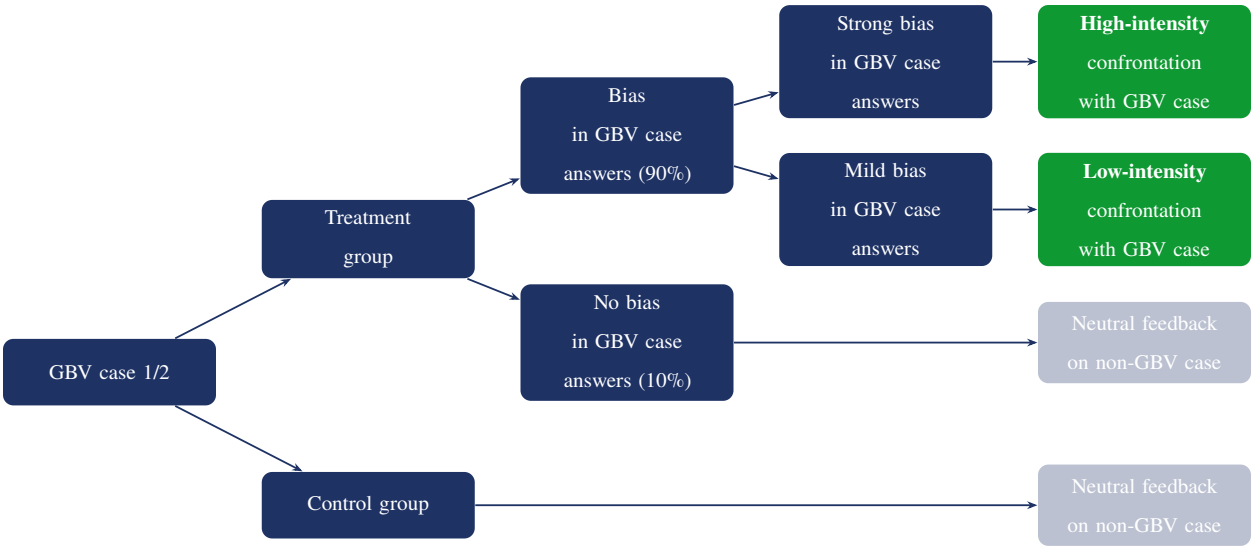

Figure 2: Experimental protocol at baseline

Endline. One week later, participants returned to complete a computer-based survey and a measure of automatic stereotyping. The survey included a novel GBV case developed by the research team, with the same structured questions on case handling used at baseline. Officers also completed measures of reflection on their handling of GBV cases, perceived credibility of GBV complaints, interest in future training, and secondary outcomes including perceived norms about GBV case handling, attitudes toward women, and empathy.

After the survey, officers completed a computerized stereotype reaction-time task adapted from Chaney and Sanchez (2018) and administered on Inquisit (Millisecond).[9] On each trial, officers viewed a photograph of a victim and a one-line complaint for 8 seconds, after which nine words appeared sequentially on screen.[10] Officers indicated whether each word "applies to this person" or "does not apply to this person" by pressing the corresponding key; each word remained on screen for 2.5 seconds or until a response was recorded. The nine words on each trial were drawn randomly from a pool of six victim-blaming words (e.g., liar, at fault, manipulative) and six innocent words (e.g., innocent, good, moral). Officers completed two practice trials followed by 40 test trials: 8 with male targets reporting non-GBV complaints, 16 with female targets reporting non-GBV complaints, and 16 with female targets reporting GBV complaints. Male-target trials serve as an additional within-person baseline for isolating GBV-specific stereotyping from general response tendencies; they are not analyzed as a separate outcome. The two outcomes from the task are the counts of "applies" responses to innocent words for victim images in the GBV and non-GBV

[9]Images were drawn from the Chicago Face Database (India).

[10]An example one-line complaint is: "This woman reported molestation by a group of boys at the bus stop near her college."

contexts respectively.

3.5 Outcomes of Interest

We measure the impact of the confrontation intervention on three sets of outcomes, each capturing a distinct dimension of officer behavior and attitudes toward GBV.

Case handling. Our primary outcomes measure how officers handle a new GBV case presented one week after the confrontation. The first is whether the officer *recognizes that a crime has been committed*, coded as a binary indicator. This outcome is reported in Table 1. The second is whether the officer *prioritizes the victim's statement*, defined as a dummy equal to one if the victim's statement is ranked as the most important input to the investigation. The third is whether the officer *prioritizes the offender's statement*, defined as a dummy equal to one if the offender's statement is ranked as the most important input to the investigation. The fourth is officers' *beliefs about the truthfulness of rape complaints*, measured as the number of rape complaints out of ten that the officer considers to be false, elicited from a case-independent survey completed at endline. The fifth is the *likelihood of registering a formal complaint* against the accused, coded as a dummy equal to one if a formal complaint is registered. Notably, 93 percent of control-group officers register the complaint, likely reflecting the administrative sanctions officers face for non-compliance; this outcome therefore has limited scope for treatment variation. These five outcomes are reported in Tables 2 and 3. As a secondary outcome, we examine whether the officer *drops the complaint* rather than pursuing it after registration; this outcome is reported in Appendix Table A10, which uses it to test peer-spillover specifications.

Automatic stereotype activation. Our second set of outcomes comes from the computerized reaction-time task administered at endline and described in Section 3. Words were drawn randomly from a pool of six victim-blaming descriptors (e.g., liar, at fault, manipulative) and six pro-victim descriptors (e.g., innocent, good, moral). We construct two outcomes: the count of "applies" responses to *innocent words* in trials featuring female targets reporting GBV complaints, and the analogous count for trials featuring female targets reporting non-GBV complaints (placebo). The GBV outcome captures automatic stereotype activation directed specifically at GBV victims; the non-GBV outcome serves as a within-person placebo, isolating GBV-specific stereotyping from general response tendencies. Male targets in non-GBV trials provide an additional within-person

baseline; they are not analyzed as a separate outcome. A lower count of "applies" responses to innocent words indicates greater automatic activation of victim stereotypes. These outcomes are reported in Table 4.

Attitudinal outcomes. Our third set of outcomes measures deeper attitudinal change, which prior work predicts to be more resistant to short-run confrontation (Chaney and Sanchez, 2018; Czopp et al., 2006). The first group captures *victim-blaming attitudes* across three scenarios: when a husband beats his wife, when a woman is harassed, and when a woman is raped. For each scenario, officers indicate in how many out of ten cases they consider the violence to be the woman's fault. The second group captures *perceived social norms*: the extent to which officers believe their friends, family, and partner consider GBV to be the woman's fault, measured on a five-point scale (1 = 0%, 2 = 25%, 3 = 50%, 4 = 75%, 5 = 100%). These outcomes are reported in Table A2 in the appendix.

Finally, we examine *self-reported affect* immediately following the feedback session, comprising six items: guilty, angry, disappointed, regretful, ashamed, and annoyed. Guilty is coded as an index following Chaney et al. (2021); the remaining outcomes use a 1–7 scale (1 = not at all; 7 = very much). These outcomes are reported in Table 5.

3.6 Deviations from Pre-Analysis Plan

Our empirical analysis closely follows our registered pre-analysis plan (RCT ID: AEARCTR-0008611), with a few exceptions. The PAP focused on the overall confrontation treatment and did not separately document the intensity gradient (high- versus low-intensity confrontation based on the degree of baseline bias), or the neutral feedback assigned to the small share of treatment officers (10%) who showed no bias at baseline, nor did it pre-specify the gender heterogeneity analysis; both the intensity gradient and the heterogeneity by officer gender are described in full in Section 3. Although the gender heterogeneity analysis was not pre-specified, it is directly predicted by the conceptual framework in Section 2, which generates the female–male asymmetry from the distribution of baseline bias and the institutional environment prior to seeing the data. Because the gender heterogeneity is the central finding of the paper and was not pre-specified, we encourage readers to interpret these results alongside the pre-registered average effects, which show no statistically significant impact on any primary outcome.[11]

[11] A recent paper by Banerjee et al. (2020) discusses the costs and benefits of adhering to PAP and recommends that the final research paper be written and judged as a distinct object from the "results of the PAP".

3.7 Validity of the Intervention

A natural concern with any confrontation-based experiment is whether the treatment was actually received as intended — that is, whether officers experienced the confrontation as a meaningful challenge rather than a routine interaction. We provide two sets of checks: randomization balance and delivery fidelity.

Randomization balance. Randomization produced well-balanced treatment and control groups. We assess balance across 11 baseline characteristics for the full sample and separately by gender (Table A1). Among male officers, two of the 11 variables show imbalances: treated males prioritize the victim's statement less often in GBV case 1, and take longer to respond in the non-GBV case (Column 5). Among female officers, the only imbalance is in GBV case 2, where treated females prioritize the victim's statement slightly less often than controls (Column 6). Within the treatment group, female officers are significantly younger than their male counterparts, and treated males exhibit longer response times in the non-GBV case (Column 7). Our results are robust to controlling for these imbalanced characteristics (Table A6), and our preferred specification includes officer age and an indicator for posting in the capital city of Bhopal.

Delivery fidelity. Three pieces of evidence support the conclusion that the confrontation was delivered and perceived as intended. First, all feedback sessions were audio recorded and monitored in real time by a member of the research team, who coded the tone of each senior officer's delivery on a four-point scale (rebuking, neutral, explanatory, or reading from script). Tone scores are balanced across treatment and control sessions (Table A1), confirming that the confrontation was delivered consistently and non-aggressively across the sample. Second, the in-person, hierarchical structure of the delivery lends the confrontation institutional credibility: senior officers were drawn from the same force as participants but were not their direct supervisors, striking a balance between authority and impartiality, and the interaction was capped at two minutes to ensure consistency across sessions. Third, self-reported affect immediately following the feedback session is informative. Although we do not find a statistically significant average effect on guilt or shame — the primary affective pathway through which confrontation is theorized to operate (Chaney and Sanchez, 2018; Czopp et al., 2006) — the affective results are consistent with the pattern noted in Section 3: treated male officers report significantly lower regret relative to controls, indicating a defensive response to the confrontation. The pattern of affective responses mirrors the behavioral results: female officers,

who de-bias their case handling, show no evidence of the defensive affect observed among males (Table 5).

4 Empirical Strategy

The effect of interest is the causal impact of prejudice confrontation on police bias. We estimate the following linear model:

$$Y_{i,s,o} = \beta_1 + \beta_2 \, Treatment_i + \mathbf{X}_i + \gamma_s + \alpha_o + \epsilon_{i,s,o} \tag{4.1}$$

where $Y_{i,s,o}$ is an outcome of interest for officer i — such as the priority given to the victim's statement or the likelihood of registering a complaint. $\mathbf{X}_i$ is a vector of individual controls including officer gender, age, and posting. γ_s and α_o denote baseline session and senior officer fixed effects, which absorb any systematic differences in how sessions and confrontations were conducted. The main explanatory variable $Treatment_i$ is an indicator equal to one for officers randomly assigned to the confrontation condition. All estimates are intention-to-treat (ITT); the 10% of treatment-assigned officers who showed no bias at baseline and received neutral feedback are retained in the treatment group. Standard errors are clustered at the officer group (session) level; the sample spans 40 sessions. With 40 clusters, asymptotic cluster-robust inference may be unreliable, so we supplement clustered standard errors with bootstrapped p-values throughout.

Randomization inference. As a further check, we conduct a randomization inference test following Heß (2017). Under the sharp null hypothesis of no treatment effect, the probability of observing a male treatment effect on victim statement prioritization as large as the one we estimate is below 0.2% (Figure A1). This test does not rely on asymptotic approximations and provides an additional check that is robust to the small number of clusters.

5 Results

We organize the results in four parts: average effects on case handling, heterogeneous effects by officer gender, automatic stereotype activation, and attitudinal outcomes. Robustness checks follow in Section 5.5.

5.1 Average Effects on Case Handling

We begin by asking whether the confrontation shifts case-handling behavior on average. Over 95% of officers recognize that a crime has occurred after reviewing the GBV case; the confrontation has no impact on this outcome in either specification (Table 1). For the remaining case-handling outcomes, Table 2 reports the average treatment effect on prioritization of the victim's statement, beliefs about the truthfulness of rape complaints, formal complaint registration, and prioritization of the offender's statement. The confrontation produces no statistically significant effect on any of these four outcomes. The null on complaint registration is unsurprising: 93% of control-group officers register the complaint, likely reflecting administrative sanctions for non-compliance rather than genuine officer discretion. Point estimates and standard errors are nearly identical across specifications with and without controls, and the bootstrapped p-values confirm that the null findings are not an artifact of clustering. Point estimates on the truthfulness of rape complaints are directionally consistent with the treatment intent — treated officers consider slightly fewer complaints to be false — though the effect falls short of conventional significance levels.

Table 1: The effects of confrontation on the recognition of crime

Variable	Crime recognized (dummy)	
	Without heterogeneity	With gender heterogeneity
	(1)	(2)
GBV treatment	0.014	−0.002
	(0.028)	(0.036)
Treatment × Female officer		0.081
		(0.054)
Female officer		−0.011
		(0.046)
Controls	Yes	Yes
Control group mean	0.949	0.949
R^2	0.050	0.058
N	293	293

Notes: Outcome is a dummy equal to one if the officer recognizes that a crime has been committed. Column (1) reports the average intent-to-treat (ITT) effect; Column (2) adds a treatment × female officer interaction. Both specifications include senior officer and session fixed effects, officer age, and a Bhopal posting indicator. $N = 293$; four officers did not respond to this item. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table 2: The effects of confrontation on the handling of GBV cases

Variable	Prioritize victim's statement		Prioritize offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GBV treatment	−0.058	−0.055	0.047	0.043	−0.654	−0.686	0.034	0.038
	(0.050)	(0.049)	(0.041)	(0.041)	(0.454)	(0.448)	(0.032)	(0.029)
	[0.377]	[0.652]	[0.304]	[0.313]	[0.132]	[0.132]	[0.305]	[0.305]
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control group mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
R^2	0.038	0.053	0.064	0.072	0.058	0.080	0.066	0.092
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): dummy equal to one if the officer prioritizes the offender's statement. Columns (5)–(6): number of rape complaints (out of 10) considered false. Columns (7)–(8): dummy equal to one if a formal complaint is registered. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. Bootstrapped p-values are reported in square brackets; multiple-hypothesis-testing adjusted p-values are reported in Table A3. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

5.2 Heterogeneous Effects by Officer Gender

The null average effect masks sharply divergent responses by officer gender. Table 3 introduces a treatment–female interaction and reveals a striking asymmetry. Among female officers, confrontation increases the probability of prioritizing the victim's statement by 23 percentage points relative to control-group females — a 27% change relative to the control mean — and simultaneously reduces the probability of prioritizing the offender's statement by 15 percentage points. Male officers move in the opposite direction: treated males place significantly less weight on the victim's statement and significantly more weight on the offender's statement, consistent with a backlash response (Rudman and Fairchild, 2004). Appendix Table A10 reports complaint dropping in peer-spillover specifications and shows no statistically significant effects. The bootstrapped p-value on the interaction term for both victim statement prioritization and offender statement prioritization is 0.001, confirming robustness to multiple-hypothesis-testing corrections.

A likely mechanism is the stark difference in baseline bias across groups (Figure A3 in the appendix; Figures A2 and A4 show the full distribution of first-course-of-action responses pooled and by treatment status). Among female officers, 72% exhibit only mild pre-treatment bias; among male officers assigned to treatment, 51% exhibit strong bias.[12] Because policing is highly male-dominated,

[12]Strong bias is defined as, for example, filing a report against the GBV victim rather than the accused, and is classified from baseline case responses for all officers. The 51% figure refers to the share of treated male officers exhibiting strong bias at baseline (58 of 112 treated male officers), as strong bias determines treatment intensity for the treated group. Among ASI and SI officers in Madhya Pradesh, 13% are female (Bureau of Police Research and Development, 2021).

the average female officer perceives her work environment as more biased than her own responses, making her receptive to confrontation feedback in ways that strongly biased male officers are not — consistent with evidence that intervention receptivity depends on baseline characteristics (Blattman et al., 2023). Male officers instead exhibit backlash, driven primarily by the strongly biased subgroup: removing strongly biased males from the sample renders the negative effect on victim statement prioritization substantially smaller and statistically insignificant (Table A5). This suggests that the share of strongly biased officers in a target group is likely to determine whether prejudice confrontation induces de-biasing or backlash — an important design consideration for scaling such interventions.

An alternative interpretation is that the gender heterogeneity reflects treatment intensity rather than gender *per se*: because male officers are more strongly biased at baseline, they mechanically receive more intense confrontations, so the differential response could be an intensity effect in disguise. Two pieces of evidence argue against this interpretation. First, the reaction-time task captures automatic stereotype activation one week after a two-minute conversation — a response unlikely to be driven by the verbal intensity of the feedback itself, yet the backlash pattern among male officers persists on this measure. Second, removing strongly biased male officers from the sample eliminates the backlash entirely (Table A5), but this group is defined by baseline bias level, not by confrontation intensity — consistent with bias level rather than feedback intensity as the operative mechanism. Because intensity and bias level are correlated by design, the reaction-time result is the cleaner of the two pieces of evidence on this point.

Table 3: Heterogeneous effects of confrontation on the handling of GBV cases

Variable	Prioritize victim's statement		Prioritize offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GBV treatment	−0.129**	−0.127**	0.096*	0.092*	−0.677	−0.706	0.044	0.048
	(0.055)	(0.055)	(0.045)	(0.046)	(0.403)	(0.411)	(0.039)	(0.036)
Treatment × Female officer	0.361***	0.360***	−0.246***	−0.246***	0.131	0.112	−0.050	−0.050
	(0.072)	(0.075)	(0.050)	(0.052)	(0.813)	(0.870)	(0.040)	(0.038)
	[0.001]	[0.001]	[0.001]	[0.001]	[0.700]	[0.700]	[0.285]	[0.285]
Female officer	−0.092	−0.141**	0.113**	0.126***	0.877**	0.517	0.067**	0.036
	(0.066)	(0.063)	(0.041)	(0.040)	(0.364)	(0.501)	(0.023)	(0.026)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control group mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
R^2	0.073	0.084	0.091	0.098	0.071	0.084	0.074	0.094
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): dummy equal to one if the officer prioritizes the offender's statement. Columns (5)–(6): number of rape complaints (out of 10) considered false. Columns (7)–(8): dummy equal to one if a formal complaint is registered. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. The interaction term equals one for treated female officers. Bootstrapped p-values for the interaction term are reported in square brackets; multiple-hypothesis-testing adjusted p-values are reported in Table A3. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

5.3 Automatic Stereotype Activation

The reaction-time task allows us to examine whether the behavioral patterns above extend to automatic, rather than deliberate, responses. The outcome is the count of "applies" responses to *innocent* words when officers assess a GBV victim; a lower count indicates greater automatic stereotype activation — fewer pro-victim descriptors attributed at the automatic level. The results mirror the case-handling findings. Treated male officers apply significantly fewer innocent words to GBV victims relative to controls (Column 1, Table 4). The point estimate of −6.9 against a control mean of 44.8 corresponds to a 15% reduction in innocent-word attribution, indicating substantially greater automatic stereotyping one week after the confrontation. The interaction term for female officers is positive, indicating a relative shift in the opposite direction compared to males and directionally consistent with the de-biasing pattern documented in case handling, though the net effect for treated female officers is not statistically significant.

Crucially, neither group differs from controls on the non-GBV placebo trials (Column 2), confirming that the male backlash is specific to GBV contexts rather than a general reaction to being

confronted (Rudman and Fairchild, 2004), and ruling out demand effects or generalized changes in response tendencies as alternative explanations. The interaction term for female officers on the non-GBV placebo trials is large in magnitude but imprecisely estimated ($\hat{\beta} = -5.373$, $SE = 3.180$), reflecting the small number of female officers in the reaction-time subsample ($N = 47$) rather than a systematic pattern; the GBV-specificity argument rests on the null for both groups on the placebo, which holds. Taken together, the case-handling and reaction-time results point to a consistent pattern: confrontation de-biases female officers in deliberate responses and produces no significant backlash in automatic responses, while strongly biased male officers exhibit backlash across both margins (Czopp and Monteith, 2003).

Table 4: The effects of confrontation on stereotypes based on victim's appearance

Variable	GBV trials	Non-GBV trials (placebo)
	(1)	(2)
GBV treatment	−6.914**	0.676
	(2.512)	(1.242)
Treatment × Female officer	3.157	−5.373
	(4.051)	(3.180)
Female officer	1.725	1.987
	(3.303)	(2.300)
Controls	Yes	Yes
Control group mean	44.828	45.867
R^2	0.235	0.253
N	233	233

Notes: Outcome is the count of "applies" responses to innocent words per trial type; a lower count indicates greater automatic victim stereotyping. Column (1): female targets reporting GBV complaints. Column (2): female targets reporting non-GBV complaints (placebo). Specification follows Table 3 and adds a control for total key presses. $N = 233$ due to insufficient time on a small number of intervention days, primarily one day, due to laptop availability constraints (47 female, 186 male officers). Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

5.4 Attitudinal Outcomes

We find no statistically significant treatment effects on any deeper attitudinal outcome, with the exception of regret among treated male officers. Self-reported affect following the feedback session

is otherwise largely unaffected by the confrontation (Table 5). Regret decreases significantly among treated male officers — consistent with the defensive response documented in case handling and the reaction-time task, and in contrast to the guilt-to-rumination pathway through which confrontation is theorized to produce lasting de-biasing (Chaney and Sanchez, 2018). Victim-blaming attitudes and perceived social norms about GBV are likewise unaffected (Table A2).

The absence of attitudinal change alongside the behavioral shifts documented above is itself informative. Notably, the guilt-to-rumination pathway theorized by Chaney and Sanchez (2018) is not confirmed in the affect data: treated officers report no significant increase in guilt or shame, and treated male officers report *less* regret than controls, consistent with defensive dismissal rather than self-directed affect. De-biasing among female officers thus appears to operate through automatic inhibition of habitual scripts — establishing "cues for control" (Monteith et al., 2002) — rather than through conscious guilt-motivated reflection, a pattern directionally consistent with, though not confirmed by, the reaction-time evidence for female officers, while the backlash among male officers reflects an automatic defensive reaction rather than a deliberate rejection of the intervention's message (Czopp and Monteith, 2003).

Table 5: The effects of confrontation on guilt and self-reflection

Variable	Guilty	Angry	Disappointed	Regretful	Ashamed	Annoyed
	(1)	(2)	(3)	(4)	(5)	(6)
GBV treatment	−0.353	0.197	−0.133	−0.569**	0.164	−0.155
	(0.259)	(0.213)	(0.159)	(0.232)	(0.186)	(0.217)
Treatment × Female officer	−0.133	−0.604	0.081	0.057	−0.155	−0.695
	(0.610)	(0.726)	(0.355)	(0.512)	(0.329)	(0.401)
Female officer	−0.073	0.611	−0.241	−0.044	0.012	0.200
	(0.398)	(0.369)	(0.365)	(0.443)	(0.196)	(0.403)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Control group mean	0.214	2.225	1.975	2.819	1.525	2.206
R^2	0.077	0.131	0.068	0.106	0.061	0.107
N	297	297	297	297	297	297

Notes: Outcomes are self-reported affect measured immediately following the feedback session. Guilty is coded as an index following Chaney et al. (2021); remaining outcomes use a 1–7 scale (1 = not at all; 7 = very much). Specification follows Table 3. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

5.5 Robustness Checks

Alternative specifications. Our main results are robust to an array of alternative specifications. We present three checks. First, we address multiple hypothesis testing by computing bootstrapped p-values for all main outcomes in Table A3, following a procedure that asymptotically controls the family-wise error rate (List et al., 2019). The key findings remain statistically significant at conventional levels. Second, Table A4 shows that the results hold when we drop all fixed effects and officer controls and replace clustered standard errors with heteroskedasticity-robust standard errors. Point estimates are very similar in size and precision to the main specification. Third, we control for all baseline characteristics that show statistically significant imbalances in Table A6. The main results are qualitatively unchanged, though somewhat less precisely estimated.

Experimenter demand. We assess the potential threat that officers may have adjusted their endline responses to confirm what they perceived the senior officer's expectations to be. Following De Quidt et al. (2018), recent evidence finds that strong experimenter demand effects are small in magnitude and insufficient to reverse directional comparative statics (Winichakul et al., 2025). Several features of our design further limit this concern. The endline GBV case was new and unseen at baseline, so officers could not have anticipated the specific scenario. The reaction-time task captures automatic stereotype activation and is difficult to consciously manipulate. Crucially, treated and control officers do not differ on the non-GBV placebo trials of the reaction-time task (Table 4), ruling out general demand-driven shifts in response tendencies as an alternative explanation. Finally, the backlash documented among male officers runs counter to any simple hypothesis-confirmation story.

Endline attrition. Approximately 8% of participants did not return for the endline session. Table A7 reports balance checks across baseline characteristics for completers and attritors. Attritors do not differ systematically from completers; the only marginally significant difference is in Bhopal posting ($p < .10$), which is included as a control in all specifications. Absences were work-related, and results are robust to their exclusion.

Reaction-time task attrition. The reaction-time task has a higher missing-data rate than the case-handling outcomes ($N = 233$ vs. $N = 297$), attributable to insufficient time on a small number of intervention days, primarily one day, due to laptop availability constraints. On affected days, officers were selected randomly to complete the task. Because treatment was assigned within

session rather than by day, participation on an affected day is orthogonal to treatment status by design, allaying concerns about treatment-selective attrition. Table A8 reports balance checks across baseline characteristics, confirming that treatment assignment is balanced between included and missing officers. One variable shows a marginally significant difference, GBV Case 2 victim prioritization ($p < .10$), but this imbalance is uncorrelated with treatment assignment.

Statistical power. The study is well powered for its central finding: the treatment–female interaction on victim statement prioritization has implied power of essentially 100% and far exceeds its minimum detectable effect. The null average effects and null attitudinal outcomes should not be interpreted as evidence of no effect — the design was underpowered to detect effects of modest size on these outcomes, with a minimum detectable effect of 0.118 (14% of the control mean) for the full sample. A full ex-post power analysis is reported in Appendix B.

Additional heterogeneity. Analyses reported in the appendix examine effects by confrontation intensity (Table A9), potential peer spillovers (Tables A10 and A11), and heterogeneity by officer age (Table A12); none of these analyses undermines the main findings.

6 Conclusion

Police officers are the primary institutional gateway for victims of gender-based violence, yet their discretionary handling of complaints remains a critical and largely neglected barrier to justice. This paper provides the first experimental evidence on whether confronting officers with their own bias can shift case-handling behavior in a developing-country setting where egalitarian norms are weak and gender-based bias is the target.

Confrontation produces no statistically significant average effect, but generates sharply divergent responses by officer gender. Female officers de-bias their case handling, prioritizing the victim's statement and deprioritizing the offender's; male officers exhibit backlash across both deliberate responses (deprioritizing the victim and elevating the offender) and in automatic stereotype activation. Neither group shifts on deeper attitudinal outcomes, indicating that confrontation changes behavioral responses without revising underlying beliefs — and without activating the guilt-to-rumination pathway theorized in the confrontation literature; de-biasing among female officers instead appears to operate through automatic inhibition of habitual scripts rather than conscious guilt-motivated

reflection, a pattern directionally consistent with, though not confirmed by, the reaction-time evidence for female officers. The distribution of baseline bias is the likely mechanism, and the share of strongly biased officers in a target group thus appears to be the key design parameter determining whether a confrontation intervention de-biases or backfires.

The policy implication is clear. A uniform confrontation risks entrenching bias among the most strongly biased — the very group most in need of change. Effective programs must account for the distribution of baseline bias and tailor delivery accordingly, pairing confrontation with complementary approaches for high-bias groups. The response to confrontation is contingent on institutional hierarchy, social norms, and cultural context (Jayachandran, 2015), and our findings derive from one Indian state with particularly high rates of GBV and below-average policing quality scores; generalizability to settings with stronger baseline norms should be assessed carefully. Our design identifies the joint effect of confrontation and GBV salience rather than confrontation alone, and all outcomes are measured in a lab-in-the-field setting rather than from administrative case records; whether the behavioral shifts documented here translate to real case outcomes remains an open question. Nevertheless, the distribution of baseline bias as a design parameter is likely to travel across hierarchical institutions where confrontation-based feedback is used, including other developing-country police forces, other judicial and quasi-judicial settings, and male-dominated workplaces more broadly. Two questions remain open: whether the behavioral shifts documented here persist beyond one week, and whether peer accountability, incentives, or structured training can achieve lasting de-biasing among strongly biased male officers.

References

Aizer, A. (2010). The gender wage gap and domestic violence. *American Economic Review 100*(4), 1847–59.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review 114*(7), 1916–1948.

Amaral, S., G. Borker, N. Prakash, and M. M. Sviatschi (2026). Debiasing law enforcement officers: Evidence from an expressive arts intervention in india. Working paper.

Amaral, S., G. B. Dahl, T. Hener, V. Kaiser, and H. Rainer (2023). Deterrence or backlash? arrests and the dynamics of domestic violence. Technical report, National Bureau of Economic Research.

Anderberg, D., H. Rainer, J. Wadsworth, and T. Wilson (2016). Unemployment and domestic violence: Theory and evidence. *Economic Journal 126*(597), 1947–1979.

Bandiera, O., N. Buehren, R. Burgess, M. Goldstein, S. Gulesci, I. Rasul, and M. Sulaiman (2020). Women's empowerment in action: evidence from a randomized control trial in africa. *American Economic Journal: Applied Economics 12*(1), 210–59.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. *NBER Working Paper 26993*.

Belknap, J. (2010). Rape: Too hard to report and too easy to discredit victims. *Violence Against Women 16*(12), 1335–1344.

Blattman, C., S. Chaskel, J. C. Jamison, and M. Sheridan (2023). Cognitive behavioral therapy reduces crime and violence over ten years: Experimental evidence. *American Economic Review: Insights 5*(4), 527–545.

Bureau of Police Research and Development (2021). Data on police organizations 2021. Technical report, Ministry of Home Affairs, Government of India. Available at http://www.bprd.nic.in.

Card, D. and G. B. Dahl (2011). Family violence and football: The effect of unexpected emotional cues on violent behavior. *The quarterly journal of economics 126*(1), 103–143.

Chaney, K. E. and D. T. Sanchez (2018). The endurance of interpersonal confrontations as a prejudice reduction strategy. *Personality and Social Psychology Bulletin 44*(3), 418–429.

Chaney, K. E., D. T. Sanchez, N. P. Alt, and M. J. Shih (2021). The breadth of confrontations as a prejudice reduction strategy. *Social Psychological and Personality Science 12*(3), 314–322.

Chaney, K. E., D. M. Young, and D. T. Sanchez (2015). Confrontation's health outcomes and promotion of egalitarianism (c-hope) framework. *Translational Issues in Psychological Science 1*(4), 363.

Czopp, A. M. and M. J. Monteith (2003). Confronting prejudice (literally): Reactions to confrontations of racial and gender bias. *Personality and social psychology bulletin 29*(4), 532–544.

Czopp, A. M., M. J. Monteith, and A. Y. Mark (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of personality and social psychology 90*(5), 784.

De Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and bounding experimenter demand. *American Economic Review 108*(11), 3266–3302.

Degener, T. and Y. Koster-Dreese (1995). Declaration on the elimination of violence against women: by general assembly resolution 48/104 of 20 december 1993. In *Human Rights and Disabled Persons*, pp. 416–422. Brill Nijhoff.

Dube, O., S. J. MacArthur, and A. K. Shah (2025). A cognitive view of policing. *The Quarterly Journal of Economics 140*(1), 745–791.

García-Moreno, C., C. Zimmerman, A. Morris-Gehring, L. Heise, A. Amin, N. Abrahams, O. Montoya, P. Bhate-Deosthali, N. Kilonzo, and C. Watts (2015). Addressing violence against women: a call to action. *The Lancet 385*(9978), 1685–1695.

Green, D. P., A. M. Wilke, and J. Cooper (2020). Countering violence against women by encouraging disclosure: A mass media experiment in rural uganda. *Comparative Political Studies 53*(14), 2283–2320.

Guarnieri, E. and H. Rainer (2021). Colonialism and female empowerment: A two-sided legacy. *Journal of Development Economics 151*, 102666.

Heß, S. (2017). Randomization inference with Stata: A guide and software. *The Stata Journal 17*(3), 630–651.

Hidrobo, M., A. Peterman, and L. Heise (2016). The effect of cash, vouchers, and food transfers on intimate partner violence: evidence from a randomized experiment in northern ecuador. *American Economic Journal: Applied Economics 8*(3), 284–303.

Howell, J. L., L. Redford, G. Pogge, and K. A. Ratliff (2017). Defensive responding to iat feedback. *Social Cognition 35*(5), 520–562.

Indian Police Foundation (2021). IPF citizen satisfaction survey on SMART policing 2021. Technical report, Indian Police Foundation.

Jayachandran, S. (2015). The roots of gender inequality in developing countries. *Annual Review of Economics 7*(1), 63–88.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*, 773–793.

Mallett, R. K. and D. E. Wagner (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology 47*(1), 215–220.

Monteith, M. J., L. Ashburn-Nardo, C. I. Voils, and A. M. Czopp (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology 83*(5), 1029–1050.

National Crime Records Bureau (2019). Crime in india 2018, volume I. Technical report, Ministry of Home Affairs, Government of India.

Palermo, T., J. Bleck, and A. Peterman (2014). Tip of the iceberg: reporting and gender-based violence in developing countries. *American journal of epidemiology 179*(5), 602–612.

Rudman, L. A. and K. Fairchild (2004). Reactions to counterstereotypic behavior: the role of backlash in cultural stereotype maintenance. *Journal of personality and social psychology 87*(2), 157.

Sardinha, L., M. Maheu-Giroux, H. Stöckl, S. R. Meyer, and C. García-Moreno (2022). Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *The Lancet 399*(10327), 803–813.

Stevenson, B. and J. Wolfers (2006). Bargaining in the shadow of the law: Divorce laws and family distress. *The Quarterly Journal of Economics 121*(1), 267–288.

Tur-Prats, A. (2019). Family types and intimate partner violence: A historical perspective. *Review of Economics and Statistics 101*(5), 878–891.

Winichakul, K. P., G. Lezama, P. Mustafi, M. Lepper, A. Wilson, D. Danz, and L. Vesterlund (2025). Effect size, experimenter demand and inference. Working paper, University of Pittsburgh.

Appendix A: Additional Tables and Figures

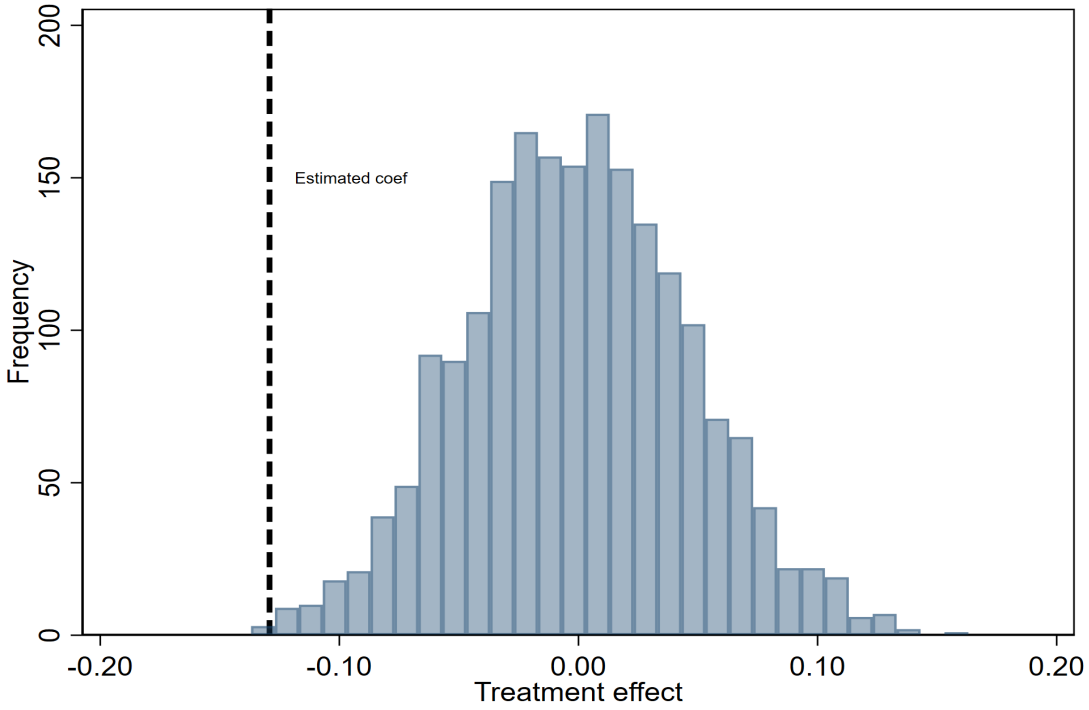

Figure A1: Randomization inference test

Notes: Distribution of treatment effects under the sharp null hypothesis of no effect, obtained via randomization inference following Heß (2017). The dashed vertical line marks the estimated coefficient on the male treatment effect (priority given to the victim's statement) from the heterogeneous specification in Table 3. Under the null, the observed estimate would occur with probability 0.002.

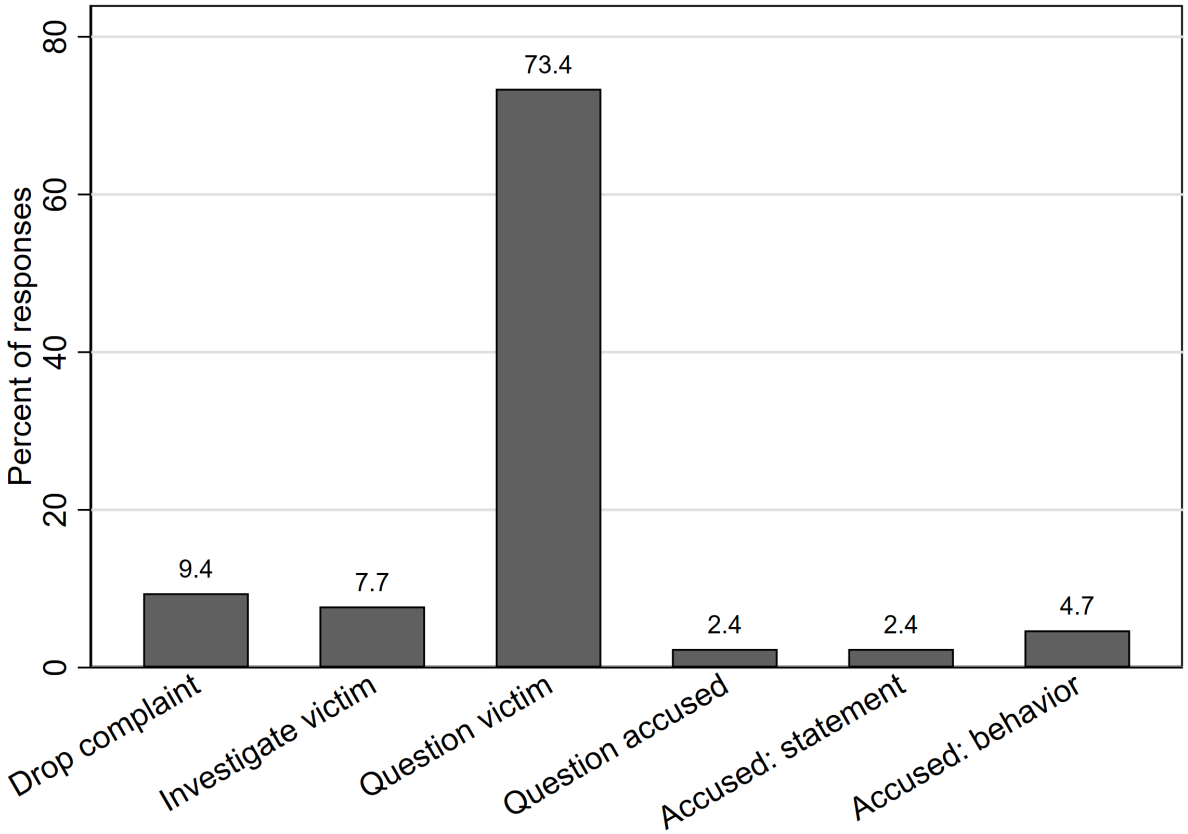

Figure A2: Distribution of prioritization items

Notes: Distribution of officers' stated first course of action at the initial stage of GBV case handling. Bars report the percentage of responses in each category, computed over the full sample.

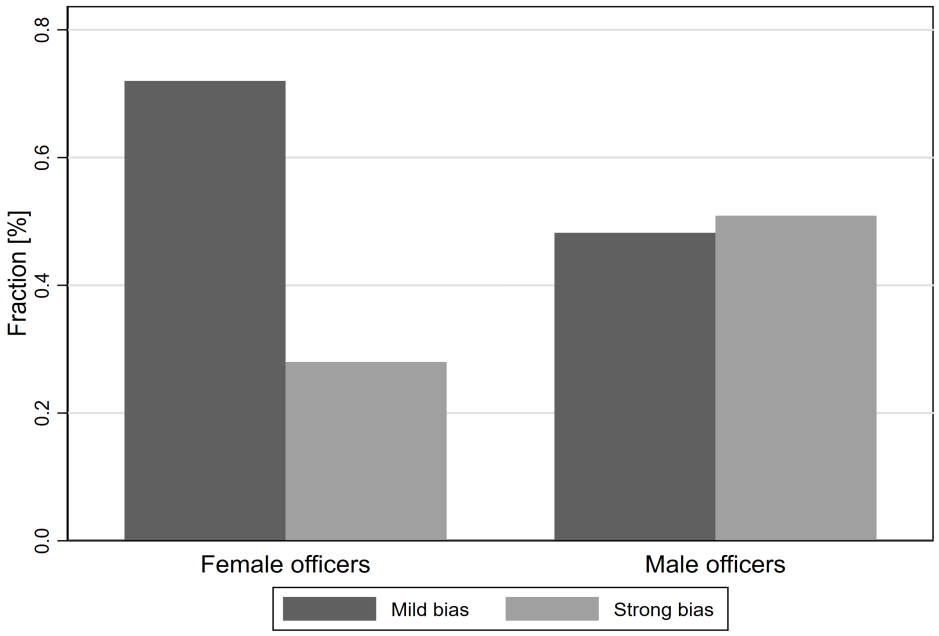

Figure A3: Baseline bias in GBV case handling by officer gender

Notes: Fraction of female and male officers exhibiting mild or strong bias in GBV case handling at baseline. Strong bias is defined as, for example, filing a report against the victim rather than the accused. Mild bias is defined as ranking an investigation of the victim among the top three of five investigative steps. The 51% figure cited in the text refers to the share of *treated* male officers exhibiting strong bias at baseline (58 of 112 treated male officers).

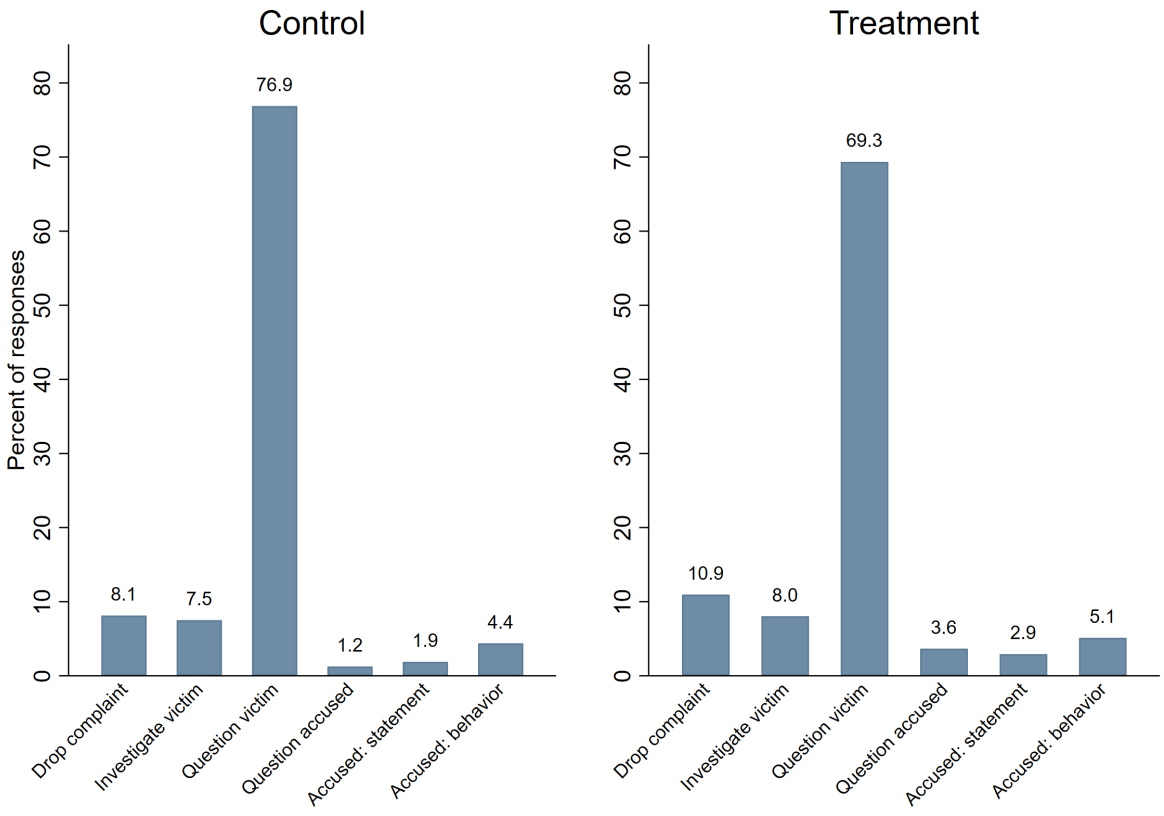

Figure A4: Distribution of prioritization items by treatment status

Notes: Distribution of officers' stated first course of action at the initial stage of GBV case handling, reported separately for treatment and control groups. Bars report the percentage of responses in each category.

Table A1: Balance tests

Variable	Mean treated men	Mean control men	Mean treated women	Mean control women	Diff: treated vs control men	Diff: treated vs control women	Diff: treated men vs treated women
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Officer age (years)	45.152	44.543	31.440	31.909	−0.540	−0.554	−9.383***
	(9.687)	(9.414)	(7.030)	(8.431)	(1.321)	(2.161)	(2.626)
Posted in Bhopal	0.420	0.394	0.320	0.273	0.119	0.032	−0.119
	(0.496)	(0.491)	(0.476)	(0.452)	(0.085)	(0.161)	(0.254)
GBV Case 1: Victim prioritized	0.109	0.324	0.222	0.368	−0.214***	0.144	0.105
	(0.315)	(0.471)	(0.441)	(0.496)	(0.072)	(0.294)	(0.127)
GBV Case 1: Response time (sec)	222.184	215.766	201.661	184.330	−7.182	29.721	1.759
	(90.360)	(95.273)	(98.161)	(128.191)	(23.536)	(27.710)	(45.451)
GBV Case 2: Victim prioritized	0.351	0.464	0.438	0.500	−0.053	−0.302*	−0.022
	(0.481)	(0.503)	(0.512)	(0.519)	(0.118)	(0.145)	(0.208)
GBV Case 2: Response time (sec)	231.783	201.703	171.268	172.088	32.871	−0.590	−79.033
	(117.034)	(105.139)	(60.482)	(59.732)	(35.745)	(22.505)	(45.244)
Non-GBV Case 1: Offender investigated	0.643	0.661	0.640	0.788	0.036	−0.125	−0.072
	(0.481)	(0.475)	(0.490)	(0.415)	(0.070)	(0.195)	(0.102)
Non-GBV Case 1: Response time (sec)	335.614	283.763	259.262	330.310	35.977*	−48.319	−91.789*
	(182.296)	(136.544)	(123.890)	(237.295)	(20.416)	(56.253)	(50.696)
Non-GBV Case 2: Victim prioritized	0.938	0.921	0.920	0.970	0.034	−0.057	−0.077
	(0.243)	(0.270)	(0.277)	(0.174)	(0.049)	(0.066)	(0.092)
Non-GBV Case 2: Response time (sec)	71.293	65.774	53.542	56.433	5.369	−7.202	−19.587**
	(44.921)	(40.579)	(23.140)	(40.938)	(4.113)	(6.626)	(7.780)
Senior officer tone	2.080	2.142	2.240	2.242	0.001	0.000	−0.022
	(0.304)	(0.431)	(0.523)	(0.561)	(0.004)	(0.000)	(0.024)
N	112	127	25	33	239	58	137

Notes: Columns (1)–(4) report simple means. Columns (5)–(7) report regression-adjusted differences controlling for baseline session and senior officer fixed effects, clustered at the session level. *Victim prioritized*: dummy equal to one if the officer ranked the victim's statement as most important. *Response time*: seconds to complete the prioritization question. *Offender investigated*: dummy equal to one if the officer ranked investigation of the offender first. *Senior officer tone*: coded by the research team (1 = rebuking; 2 = neutral; 3 = explanatory; 4 = reading from script). Standard errors in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A2: Effects of confrontation on victim-blaming attitudes and perceived social norms

Variable	Victim blaming: husband beats	Victim blaming: harassment	Victim blaming: rape	Perceived bias: friends	Perceived bias: family	Perceived bias: partner
	(1)	(2)	(3)	(4)	(5)	(6)
GBV treatment	−0.407	−0.384	0.104	−0.168	−0.087	−0.015
	(0.265)	(0.339)	(0.273)	(0.167)	(0.138)	(0.128)
Treatment × Female officer	0.508	0.073	−0.489	0.307	0.294	0.225
	(0.884)	(0.733)	(0.729)	(0.372)	(0.272)	(0.290)
Female officer	0.532	0.650	0.441	−0.282	−0.226	0.045
	(0.707)	(0.576)	(0.633)	(0.194)	(0.190)	(0.178)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Control group mean	3.675	2.587	5.544	2.094	2.094	1.744
R^2	0.107	0.102	0.105	0.124	0.093	0.090
N	297	297	297	297	297	297

Notes: Columns (1)–(3): number of cases (out of 10) in which the officer considers the violence to be the woman's fault, for three scenarios (husband beats wife; harassment; rape). Columns (4)–(6): officer's perception of the share of friends, family, and partner who consider GBV to be the woman's fault, on a five-point scale (1 = 0%; 2 = 25%; 3 = 50%; 4 = 75%; 5 = 100%). All columns include senior officer, session fixed effects, officer age and a Bhopal posting indicator. Specification follows Table 3. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A3: Multiple-hypothesis-testing adjusted p-values for main outcomes

Outcome	Coefficient	Raw bootstrap p-value	MHT-adjusted p-value
	(1)	(2)	(3)
Panel A: Average treatment effect			
Prioritize on victim's statement	-0.055	0.260	0.375
Prioritize on offender's statement	0.043	0.313	0.313
Truth of rape complaints	-0.686	0.162	0.470
Register complaint	0.038	0.231	0.475
Panel B: Treatment × female officer			
Prioritize on victim's statement	0.360	0.001	0.001
Prioritize on offender's statement	-0.246	0.004	0.010
Truth of rape complaints	0.112	0.884	0.884
Register complaint	-0.050	0.220	0.398

Notes: The table reports multiple-hypothesis-testing adjusted p-values for the preferred specifications with senior officer and session fixed effects, officer age, and a Bhopal posting indicator. Panel A tests the average treatment effect. Panel B tests the treatment × female officer interaction. Raw bootstrapped p-values and List et al. (2019) stepdown adjusted p-values are computed using `mhtreg` with 999 bootstrap repetitions clustered at the session level.

Table A4: Robustness: effects of confrontation on GBV case handling without clustering

Variable	Prioritize on victim's statement	Prioritize on offender's statement	Truth of rape complaints	Register complaint
	(1)	(2)	(3)	(4)
GBV treatment	-0.135**	0.088**	-0.722*	0.042
	(0.053)	(0.039)	(0.390)	(0.032)
	[0.012]	[0.027]	[0.067]	[0.167]
Treatment × Female officer	0.347***	-0.239***	0.292	-0.042
	(0.089)	(0.074)	(0.886)	(0.032)
	[0.000]	[0.001]	[0.749]	[0.167]
Female officer	-0.070	0.096	0.889	0.087***
	(0.078)	(0.066)	(0.548)	(0.025)
	[0.363]	[0.132]	[0.109]	[0.000]
Controls	No	No	No	No
Control group mean	0.844	0.075	5.325	0.931
R^2	0.045	0.031	0.031	0.021
N	297	297	297	297

Notes: Column (1): dummy equal to one if the officer prioritizes the victim's statement. Column (2): dummy equal to one if the officer prioritizes the offender's statement. Column (3): number of rape complaints (out of 10) considered false. Column (4): dummy equal to one if a formal complaint is registered. Models are estimated without fixed effects or additional controls. Heteroskedasticity-robust standard errors in parentheses. Bootstrapped p-values in square brackets. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A5: Robustness: effects of confrontation excluding strongly biased male officers

Variable	Prioritize on victim's statement	Prioritize on offender's statement	Truth of rape complaints	Register complaint
	(1)	(2)	(3)	(4)
GBV treatment	-0.085	0.093*	-0.906*	0.033
	(0.063)	(0.048)	(0.454)	(0.050)
Treatment x Female	0.287***	-0.234***	0.451	-0.040
	(0.082)	(0.060)	(0.983)	(0.051)
Female officer	-0.154**	0.137**	0.376	0.038
	(0.063)	(0.047)	(0.546)	(0.031)
Controls	Yes	Yes	Yes	Yes
Control Group Mean	0.844	0.075	5.325	0.931
R^2	0.093	0.146	0.086	0.105
N	239	239	239	239

Notes: Column (1): dummy equal to one if the officer prioritizes the victim's statement. Column (2): dummy equal to one if the officer prioritizes the offender's statement. Column (3): number of rape complaints (out of 10) considered false. Column (4): dummy equal to one if a formal complaint is registered. Male officers with strongly biased baseline beliefs are excluded from the sample. All columns include senior officer, session fixed effects, officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A6: Robustness: main results including baseline controls

Variable	Prioritize on victim's statement	Prioritize on offender's statement	Truth of rape complaints	Register complaint
	(1)	(2)	(3)	(4)
Panel A: Baseline non-GBV case controls				
GBV treatment	-0.133**	0.097*	-0.641	0.052
	(0.055)	(0.048)	(0.408)	(0.035)
Treatment x Female	0.378***	-0.258***	-0.057	-0.060
	(0.082)	(0.062)	(0.888)	(0.035)
Female officer	-0.151**	0.131**	0.609	0.042
	(0.070)	(0.048)	(0.520)	(0.026)
Control Group Mean	0.844	0.075	5.325	0.931
R^2	0.089	0.104	0.091	0.099
N	297	297	297	297
Panel B: Baseline non-GBV case + GBV case 1 controls				
GBV treatment	-0.105*	0.089*	-1.005*	0.016
	(0.057)	(0.047)	(0.573)	(0.049)
Treatment x Female	0.295***	-0.183**	-1.163	-0.048
	(0.079)	(0.075)	(2.029)	(0.035)
Female officer	-0.073	0.072	0.959	0.042
	(0.072)	(0.066)	(0.702)	(0.034)
Control Group Mean	0.844	0.075	5.325	0.931
R^2	0.121	0.177	0.139	0.155
N	154	154	154	154
Panel C: Baseline non-GBV case + GBV case 2 controls				
GBV treatment	-0.178	0.087	-0.699	0.085*
	(0.105)	(0.081)	(0.679)	(0.044)
Treatment x female	0.436**	-0.336**	0.766	-0.079
	(0.151)	(0.116)	(1.175)	(0.047)
Female officer	-0.230*	0.242**	0.070	0.010
	(0.118)	(0.100)	(1.409)	(0.043)
Control Group Mean	0.844	0.075	5.325	0.931
R^2	0.180	0.184	0.193	0.179
N	142	142	142	142

Notes: Replicates Table 3 adding baseline controls that showed statistically significant imbalances in Table A1. Panel A adds non-GBV case response time, officer age, and Bhopal posting indicator. Panel B additionally controls for priority given to the victim's statement in GBV case 1; N falls to 154 as only half of participants were assigned this case at baseline. Panel C replaces the GBV case 1 control with the analogous variable for GBV case 2 ($N = 142$). Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A7: Balance test: Attritors vs. Completers

Variable	Completers	Attritors	Difference
	(1)	(2)	(3)
Officer age (years)	42.266	41.636	-0.630
	(10.576)	(10.684)	(2.316)
Posted in Bhopal	0.384	0.200	-0.184*
	(0.487)	(0.414)	(0.107)
Female officer	0.195	0.267	0.071
	(0.397)	(0.458)	(0.117)
GBV Case 1: Victim prioritized	0.247	0.182	-0.065
	(0.433)	(0.405)	(0.122)
GBV Case 2: Victim prioritized	0.420	0.467	0.047
	(0.495)	(0.516)	(0.136)
Non-GBV Case 1: Offender investigated	0.667	0.654	-0.013
	(0.472)	(0.485)	(0.098)
Non-GBV Case 2: Victim prioritized	0.933	1.000	0.067***
	(0.251)	(0.000)	(0.015)
N	297	26	323

Notes: Column (1) reports baseline means for endline completers. Column (2) reports baseline means for attritors. Column (3) reports the difference between attritors and completers. Standard deviations are in parentheses for means; robust standard errors are in parentheses for differences. Eleven attritors are missing district/posting metadata. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A8: Reaction-time task: balance between included and missing officers

Variable	In reaction-time sample	Missing from reaction-time sample	Difference (missing − included)
	(1)	(2)	(3)
Treatment assignment	0.451	0.500	0.049
	(0.499)	(0.504)	(0.071)
Officer age (years)	42.056	43.031	0.975
	(10.487)	(10.944)	(1.526)
Female officer	0.202	0.172	−0.030
	(0.402)	(0.380)	(0.054)
Posted in Bhopal	0.382	0.391	0.009
	(0.487)	(0.492)	(0.069)
GBV Case 1: Victim prioritized	0.246	0.250	0.004
	(0.432)	(0.440)	(0.086)
GBV Case 2: Victim prioritized	0.459	0.281	−0.178*
	(0.501)	(0.457)	(0.093)
Non-GBV Case 1: Offender investigated	0.661	0.688	0.027
	(0.474)	(0.467)	(0.066)
Non-GBV Case 2: Victim prioritized	0.936	0.922	−0.014
	(0.246)	(0.270)	(0.037)
N	233	64	297

Notes: Column (1) reports baseline means for officers included in the reaction-time analysis sample ($N = 233$). Column (2) reports baseline means for officers missing from the RT sample ($N = 64$). Column (3) reports the regression-adjusted difference (missing minus included). One variable shows a marginally significant difference: GBV Case 2 victim prioritization ($p < .10$), which is uncorrelated with treatment assignment (Column 1 vs. Column 2 on treatment row). Standard deviations in parentheses for columns (1) and (2); robust standard errors in parentheses for column (3). $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A9: Heterogeneity by confrontation intensity: effects on general GBV disposition

Variable	Prioritize on victim's statement		Prioritize on offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mild confrontation	-0.034	-0.037	0.060	0.057	-0.908**	-0.994**	0.028	0.028
	(0.079)	(0.076)	(0.053)	(0.049)	(0.405)	(0.427)	(0.052)	(0.048)
Strong confrontation	-0.197***	-0.190**	0.100*	0.096*	-0.462	-0.437	0.056	0.062*
	(0.065)	(0.072)	(0.052)	(0.054)	(0.504)	(0.485)	(0.035)	(0.032)
Female officer	-0.087	-0.131*	0.107**	0.119**	0.886**	0.498	0.066***	0.034
	(0.068)	(0.065)	(0.042)	(0.041)	(0.364)	(0.498)	(0.022)	(0.026)
Mild × female officer	0.273***	0.276***	-0.212***	-0.215***	0.020	0.006	-0.035	-0.032
	(0.088)	(0.087)	(0.068)	(0.065)	(0.951)	(1.038)	(0.048)	(0.046)
Strong × female officer	0.410***	0.410***	-0.246***	-0.243***	0.751	0.796	-0.057	-0.059
	(0.090)	(0.096)	(0.050)	(0.054)	(1.110)	(1.191)	(0.043)	(0.039)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control Group Mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
R^2	0.084	0.094	0.086	0.094	0.076	0.090	0.074	0.095
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): dummy equal to one if the officer prioritizes the offender's statement. Columns (5)–(6): number of rape complaints (out of 10) considered false. Columns (7)–(8): dummy equal to one if a formal complaint is registered. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A10: Peer spillovers and GBV case handling

Variable	Prioritize on victim's statement		Prioritize on offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GBV treatment	-0.035	-0.032	0.059	0.053	-0.778	-0.850	0.077	0.081
	(0.078)	(0.081)	(0.064)	(0.064)	(0.626)	(0.608)	(0.064)	(0.060)
Peer in other treatment arm	0.085	0.091	-0.050	-0.064	-0.657	-0.822	0.083*	0.094**
	(0.062)	(0.063)	(0.035)	(0.038)	(0.550)	(0.518)	(0.040)	(0.041)
GBV treatment × Peer in other arm	-0.049	-0.048	-0.014	-0.011	0.294	0.364	-0.081	-0.083
	(0.089)	(0.090)	(0.067)	(0.065)	(0.681)	(0.635)	(0.067)	(0.064)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control Group Mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
R^2	0.045	0.061	0.073	0.084	0.066	0.091	0.083	0.113
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): dummy equal to one if the officer prioritizes the offender's statement. Columns (5)–(6): number of rape complaints (out of 10) considered false. Columns (7)–(8): dummy equal to one if a formal complaint is registered. *Peer in opposite treatment arm* is a dummy equal to one if the officer has at least one colleague in the same posting assigned to the opposite treatment arm. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A11: Heterogeneous effects of confrontation by officer age and gender

Variable	Prioritize on victim's statement		Prioritize on offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GBV treatment	-0.136*	-0.131*	0.075	0.071	-0.765	-0.779	0.081*	0.086**
	(0.070)	(0.073)	(0.066)	(0.069)	(0.631)	(0.631)	(0.044)	(0.039)
GBV treatment x Female	0.194**	0.166*	-0.127	-0.103	0.388	0.478	-0.094**	-0.118**
	(0.081)	(0.092)	(0.075)	(0.089)	(1.479)	(1.644)	(0.043)	(0.043)
GBV treatment x Young	0.011	0.012	0.049	0.047	0.129	0.119	-0.093	-0.092*
	(0.085)	(0.086)	(0.081)	(0.083)	(1.205)	(1.191)	(0.054)	(0.049)
Female x Young	-0.298***	-0.327***	0.236***	0.245***	-0.989	-1.032	-0.173***	-0.183***
	(0.089)	(0.099)	(0.047)	(0.058)	(0.869)	(0.925)	(0.034)	(0.053)
GBV treatment x Female x Young	0.186	0.216	-0.165	-0.191	-0.326	-0.422	0.105*	0.131**
	(0.138)	(0.142)	(0.121)	(0.130)	(2.384)	(2.532)	(0.057)	(0.055)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control Group Mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
TE: Young Female	0.255	0.264	-0.168	-0.176	-0.574	-0.605	-0.001	0.007
SE (Young Female)	0.061	0.060	0.052	0.049	1.058	1.072	0.009	0.013
p-value (Young Female)	0.001	0.001	0.005	0.003	0.596	0.581	0.919	0.599
R^2	0.085	0.092	0.098	0.106	0.089	0.091	0.094	0.107
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): number of rape complaints (out of 10) considered false. Columns (5)–(6): dummy equal to one if a formal complaint is registered. *Young* is an indicator equal to one for officers aged 45 or below, the sample median officer age. The interaction term tests whether treatment effects differ for officers with cross-arm peers. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Table A12: Heterogeneous effects of confrontation by officer age

Variable	Prioritize on victim's statement		Prioritize on offender's statement		Truth of rape complaints		Register complaint	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GBV treatment	-0.132*	-0.128*	0.073	0.070	-0.760	-0.765	0.076*	0.079*
	(0.069)	(0.070)	(0.062)	(0.065)	(0.588)	(0.583)	(0.042)	(0.038)
GBV treatment x Young	0.138*	0.142*	-0.051	-0.054	0.127	0.118	-0.085*	-0.081*
	(0.075)	(0.076)	(0.067)	(0.070)	(0.988)	(0.974)	(0.044)	(0.039)
Young	0.020	-0.067	0.019	0.036	0.989	0.673	0.089***	0.051
	(0.062)	(0.112)	(0.034)	(0.067)	(0.645)	(0.613)	(0.029)	(0.064)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Control Group Mean	0.844	0.844	0.075	0.075	5.325	5.325	0.931	0.931
Treatment effect (Young)	0.006	0.014	0.022	0.016	-0.633	-0.647	-0.009	-0.002
SE (Young effect)	0.056	0.053	0.044	0.041	0.732	0.723	0.032	0.029
p-value (Young effect)	0.913	0.796	0.624	0.706	0.401	0.385	0.781	0.940
R^2	0.054	0.061	0.066	0.074	0.081	0.083	0.085	0.100
N	297	297	297	297	297	297	297	297

Notes: Columns (1)–(2): dummy equal to one if the officer prioritizes the victim's statement. Columns (3)–(4): dummy equal to one if the officer prioritizes the offender's statement. Columns (5)–(6): number of rape complaints (out of 10) considered false. Columns (7)–(8): dummy equal to one if a formal complaint is registered. *Young* is an indicator equal to one for officers aged 45 or below, the sample median officer age. Odd columns include senior officer and session fixed effects; even columns add officer age and a Bhopal posting indicator. Standard errors clustered at the session level in parentheses. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Appendix B: Ex-Post Power Analysis

Table A13 reports ex-post power for the main estimates. Ex-post power is the probability, under the estimated effect size, of obtaining $p < 0.05$ in a two-sided test, computed as $\Phi(|\hat{\beta}|/SE - 1.96)$ using coefficients and standard errors from the main tables.[13] Because ex-post power is mechanically high for statistically significant results and low for null results, we place greater weight on the minimum detectable effect (MDE), defined as the smallest true effect detectable at 80% power, as a benchmark less sensitive to the realized effect size. MDEs are computed using actual treatment and control cell sizes from Table A1; the primary binary outcome has a control-group mean of 0.844 and standard deviation of 0.363.

The study is well powered for its central finding. The treatment–female interaction on victim statement prioritization ($\hat{\beta} = 0.360$, $SE = 0.075$) has implied power of essentially 100%. The male treatment effect on the reaction-time stereotyping task ($\hat{\beta} = -6.914$, $SE = 2.512$) reaches 79% power, just below the conventional threshold.

Average effects and attitudinal outcomes are underpowered. The average treatment effect on victim prioritization has implied power of 20%, and the MDE for the full sample is 0.118 (14% of the control mean), slightly below the estimated male backlash effect of 0.127. Null average results should therefore not be interpreted as evidence of no effect; the design lacked power to detect plausible effects of modest magnitude. The male-specific backlash estimate ($\hat{\beta} = -0.127$, $SE = 0.055$) has implied power of 64%; the result is statistically significant because the estimated effect is large, though effect size estimates from underpowered designs are subject to inflation and the magnitude should be interpreted accordingly. Beliefs about the truthfulness of rape complaints and the complaint-dropping interaction are underpowered at 33% and 45% respectively.

The small female subsample ($N = 58$, $N_T = 25$) limits standalone inference for female officers: the female subsample MDE is 0.270 (32% of the control mean), and the net female effect of 0.233 is estimated with a conservative standard error of 0.093, implying a lower bound of 71% power. The interaction specification recovers power by pooling information across groups, which accounts for the gender heterogeneity result being well powered despite the limited female subsample.

[13] We use the standard normal critical value $z = 1.96$; the $t(39)$ critical value for 40 clusters is 2.023, which reduces each power figure by at most 2.5 percentage points but does not change any conclusion.

Table A13: Ex-post power by outcome

	$\hat{\beta}$	SE	$\lvert t \rvert$	Power	MDE (80%)
Primary outcome: victim statement prioritization					
Average TE	−0.055	0.049	1.12	20%	0.118
Male TE	−0.127	0.055	2.31	64%	0.132
Interaction $T\times F$	0.360	0.075	4.80	∼100%	—
Net female TE†	0.233	0.093	2.51	≥71%	0.270
Secondary outcomes					
Male TE: offender priority	0.088	0.039	2.26	62%	—
$T\times F$: drop complaint	−0.108	0.059	1.83	45%	—
Avg TE: rape beliefs	−0.686	0.448	1.53	33%	—
Reaction-time task					
Male TE: GBV trials	−6.914	2.512	2.75	79%	—

Notes: Power computed as $\Phi(|\hat{\beta}|/SE - 1.96)$, two-sided $\alpha = 0.05$. MDEs use actual cell sizes ($N_T = 137$, $N_C = 160$ for the full sample; $N_T = 112$, $N_C = 127$ for males; $N_T = 25$, $N_C = 33$ for females) and $SD = 0.363$ for the primary outcome; dashes indicate outcomes where MDE is not the relevant diagnostic. †SE is a conservative upper bound ($\sqrt{SE^2_{\text{treat}} + SE^2_{\text{int}}}$); power of ≥71% is therefore a lower bound.

Appendix C: Case Vignettes

GBV Case I: Reshma

A woman named Reshma comes to the police station to report domestic violence by her in-laws and other family members. You are the SHO of the police station on duty along with ASI XX, who is the day officer. There is no one else at the police station. Reshma has been married to Rajan for three years. She reports that many members of her marital family have been calling her abusive names, commenting on her character, and harassing her mentally and emotionally. Reshma tells you that her father-in-law, mother-in-law, and husband's older brother have all abused her. She reports a specific incident in which her mother-in-law dragged her by the hair in the kitchen while her brother-in-law hit her. The father-in-law and brother-in-law have also complained, in front of visitors, that the dowry they received was insufficient for the bride they got. Whenever they speak to Reshma's parents, they repeat this. Reshma also claims that Rajan has been physically abusive, but she is very quiet when the ASI asks her to show evidence of injuries; she refuses to speak. When the ASI asks about an alleged extramarital affair, she is again silent. Reshma wishes to file complaints of domestic violence and under the Dowry Act against her marital family members.

Q1. Do you think a serious criminal offense has been committed in the above case?

- Yes
- No

Q2. If yes, please describe the offense.

Q3. What would be your first step?

1. Drop the complaint as not important to police

2. Investigate by asking Reshma more questions
3. Investigate by asking the husband's family questions
4. Call a female police officer to speak with Reshma and extract the details of her case
5. Register Reshma's complaint

Upon investigation by the ASI, Rajan's family reveals that their behavior toward Reshma began only after they discovered that she had been in regular correspondence with a man named Ali in her natal village. Reshma's mother-in-law claims to have found many letters exchanged between the two. The ASI speaks to Rajan, who says he has not physically abused Reshma.

Q4. According to you, what are the most relevant parts of the story that require further investigation? Please rank the options in order of importance.

1. Investigate Ali to confirm the husband's account
2. Investigate Reshma's parents to ask whether they received dowry
3. Ask Reshma whether she wants to file a DIR and initiate separation from her husband
4. Ask neighbors about Reshma's household
5. Ask Reshma to first end her correspondence with Ali and then return to the police if she faces further violence

Q5. Against whom are you most likely to file an FIR, and under what section? (Select multiple)

A. Rajan

1. Section 498A (domestic violence)

B. Rajan's father-in-law, brother-in-law, and mother-in-law

2. Section 498A
3. Dowry Act

C. Ali

4. Section 498 (Enticing a married woman)
5. Section 497 (Adultery)
6. Will not file an FIR, as doing so will cause more problems for Reshma and disrupt the household
7. Will not file an FIR, as this is not a criminal case

GBV Case II: Tina

Tina Kumari, a 28-year-old school teacher, has come to the police station to file a complaint. She tells you the following: she is employed at a government primary school and was serving the mid-day meal to students when two parents, Guddu Singh and Pratibha Singh, arrived at the school. They drew Tina aside to a classroom, saying they wanted to discuss their son's academic performance. They then asked why Tina was serving the children eggs. When Tina explained that this was a Government mandate, both shouted indecent and casteist slurs at her and threatened to have her dismissed from her job. Pratibha snatched Tina's gold chain, slapped her, and left to call another parent. While Pratibha was away, Guddu Singh continued to abuse Tina, pulled off her dupatta, and stared at her. He made comments about her body and background. Tina reports that Guddu then raised his hand as if to outrage her modesty; in self-defense, she screamed and hit him in the nose. The commotion drew children into the classroom and Tina fled. Tina tells you that Guddu Singh has threatened retaliation and is well-connected to influential people in the area. She wishes to file an FIR and seek protection.

Q1. Do you think a serious criminal offense has been committed in the above case?

- Yes
- No

Q2. If yes, please describe the offense.

Q3. What would be your first step?

1. Drop the complaint as not important to police
2. Do further investigation by asking Guddu more questions
3. Do further investigation by asking Tina more questions
4. Register Tina's complaint

Upon speaking with Guddu Singh, he reveals that he and his wife had gone to the school to ask Tina about their son Somu's performance. During that conversation, they also informed Tina that they were aware of her irregular attendance and intended to report a complaint to the headmaster. Tina reportedly lost her temper, threatened to invoke the SC/ST Act against them, and abused them verbally. When Pratibha left to call more parents, Tina struck Guddu in the face and fled; his nose started bleeding as a result (he shows you a stained handkerchief). Guddu says he is willing to forgo a criminal complaint against Tina if she apologizes publicly and agrees to attend school regularly. Tina denies this account entirely, adheres to her original version, and requests your help.

Q4. According to you, what are the most relevant parts of the story that require further investigation? Please rank the options in order of importance.

1. Whether Tina has been attending school regularly
2. Whether Tina was serving eggs at school
3. Whether Tina's chain was stolen

4. Whether Guddu Singh was threatened by Tina with the SC/ST Act
5. Whether Tina hit Guddu in the nose
6. Whether Guddu sexually harassed Tina
7. Whether Guddu used casteist remarks against Tina

Q5. Whose testimony are you likely to collect for the investigation? Please rank in the order of importance you will assign each.

1. Headmaster, teachers, and children at the school
2. Guddu and Pratibha's friends
3. People from Tina's community

Q6. Against whom are you most likely to file an FIR, and under what section? (Select multiple)

A. Guddu / Pratibha

1. SC/ST Act
2. Section 509 (Indecent remarks or gestures to outrage the modesty of a woman)
3. Section 354B (Attempt to outrage modesty by disrobing)
4. Section 379 (Theft)

B. Tina

1. Hurt
2. Intentionally causing hurt
3. Causing hurt on provocation

C. Do not file an FIR, as this is an internal matter of the village that could be resolved by the village head

GBV Case III: Vidisha

A 20-year-old student, Vidisha, comes to the police station to file a complaint of sexual assault against her professor (aged 35 years) at Barkatullah University. You are the SHO on duty along with a male ASI. There is no one else at the police station. Vidisha is a second-year B.A. student and is taught history by Prof. Shrivastava.

Vidisha reports that the first incident occurred six months ago during a cultural function at the college. Prof. Shrivastava asked her to meet him in the college library. When she arrived, she found no one else there, as everyone was attending the function in the auditorium. Vidisha recounts that Prof. Shrivastava pulled her toward him and forcibly tried to kiss her. She resisted, and after scuffling for several minutes, managed to free herself and run out of the library. Too scared and traumatized to tell anyone, she continued her normal routine. Over the following weeks, however,

Prof. Shrivastava began regularly calling her to his office after class hours and molesting her. When the ASI asked for additional details, Vidisha remained silent. When asked why she continued to visit the professor and did not come to the police sooner, she said she feared it would damage her reputation and result in her expulsion from the university.

One day, Vidisha saw Prof. Shrivastava groping another student, Mary, in an empty classroom. She then realized that he may have been harassing multiple students and decided to file a complaint against him for sexual assault with intent to outrage her modesty (IPC 354). She also told the ASI that she wishes to file a complaint on behalf of Mary.

Q1. Do you think a serious criminal offense has been committed in the above case?

- Yes
- No

Q2. If yes, please describe the offense.

Q3. What would be your first step?

1. Drop the complaint and ask Vidisha to approach college authorities instead
2. Investigate by asking Vidisha more questions
3. Ask Vidisha questions about Prof. Srivastava's behavior
4. Ask Prof. Srivastava questions about Vidisha's behavior
5. Ask Prof. Srivastava questions about Vidisha's grades
6. Ask Prof. Srivastava questions about his behavior

Q4–Q11. For each of the following courses of action, please indicate how likely you are to pursue it. (Scale: 1 = Very likely; 2 = Likely; 3 = Somewhat likely; 4 = Neither likely nor unlikely; 5 = Somewhat unlikely; 6 = Unlikely; 7 = Very unlikely)

Q4. Drop the complaint and ask Vidisha to approach college authorities instead

Q5. Investigate by asking Vidisha more questions

Q6. Ask Vidisha questions about Prof. Srivastava's behavior

Q7. Ask Prof. Srivastava questions about Vidisha's behavior

Q8. Ask Prof. Srivastava questions about Vidisha's grades

Q9. Ask Prof. Srivastava questions about his behavior

Q10. Call a female police officer to speak with Vidisha and extract additional details of the case

Q11. Register Vidisha's complaint

On further investigation by the ASI, Prof. Shrivastava states that his relationship with Vidisha is consensual. He says "she is a very modern girl with a loose character; she wears tight jeans and short skirts to college and has several male friends." He also claims that when he first called her to the library, it was for academic purposes, but that she signaled interest in him and continued to visit his office on academic pretexts. When asked about Mary, he dismisses having interacted with her outside the classroom. He adds that both Vidisha and Mary are low-performing students who are trying to falsely implicate him in order to pass their upcoming final exams.

Q12. According to you, what are the most relevant parts of the story that require further investigation? Please rank the options in order of importance.

1. Ask students and faculty about Vidisha's character
2. Ask Vidisha to share her academic records to confirm Prof. Shrivastava's account
3. Ask Mary whether she faced harassment from Prof. Shrivastava
4. Ask other students about Prof. Shrivastava's character
5. Ask Vidisha to share any additional evidence she may have, such as mobile phone messages, calls, or letters
6. Investigate further with Vidisha about the nature of her relationship with Prof. Shrivastava

Q13. Against whom are you most likely to file an FIR, and under what section? (Select multiple)

1. Drop the complaint as unfounded; this is a case of rivalry
2. File FIR against Prof. for sexual harassment of Vidisha (Section 354)
3. Approach Mary to ask whether she wishes to file an FIR against Prof. for sexual harassment
4. Give Prof. Shrivastava a strong warning

Non-GBV Case: Dhairya

A young man with a bleeding head approaches you. You are in the police station with an ASI and two constables. He identifies himself as Dhairya Gupta, the owner of an electronics store near the city center. He has been renting the shop from Rahul Kumar for the past five years. The lease is set to expire next year, but two months ago Rahul found a tenant willing to pay a higher rent and has since been pressuring Dhairya to vacate. Three days ago the two had an argument in which Dhairya refused to leave before his lease expired. This morning, Rahul Kumar's wife Sunita devi and his two sons Shubham (age 18) and Sunil (age 15) stormed into the shop, began throwing items around, and shouted at Dhairya. They broke several tube-lights and pieces of kitchen equipment. When Dhairya tried to intervene, they threw items at him and a brawl ensued. The two sons were physically stronger and Dhairya was unable to defend himself. They left after threatening him to vacate within twenty-four hours or face a larger group.

Q1. Do you think a criminal offense has been committed in the above case?

- Yes

- No

Q2. If yes, please describe the offense.

Q3. What would be your first step?

1. Drop the complaint as not important to police
2. Do further investigation by asking Rahul and his family questions
3. Do further investigation by asking Dhairya more questions
4. Register Dhairya's complaint
5. Tell Dhairya to go to the hospital

Upon speaking with Rahul, he says the altercation occurred because he had discovered that Dhairya was using illegal drugs (charas) in the storeroom of the shop, and had asked him to vacate for this reason. That morning, Sunita heard disturbing sounds from the shop below and went to check. Dhairya began shouting at her, and Shubham and Sunil came to their mother's defense. A fight broke out, during which Sunil and Shubham admit to hitting Dhairya on the head with a tube-light because he was saying offensive things about their family.

Q4. According to you, what are the most relevant parts of the story that require further investigation? Please rank the options from 1 to 5 in order of importance, where 1 is most important and 5 is least important.

1. Whether Dhairya has been taking illegal drugs
2. What the lease agreement between Dhairya and Rahul says
3. Whether valuable items were broken in Dhairya's shop
4. Who was hurt as a result of the fight between Shubham, Sunil, and Dhairya
5. Whether Rahul has found a higher-paying tenant for the shop

Q5. Whose testimony are you likely to collect for the investigation? Please rank from 1 to 3 in order of importance, where 1 is most important and 3 is least important.

1. Nearby shopkeepers
2. Sunita
3. Friends of Dhairya

Q6. What arrests will you make and on what charges? (Select multiple)

1. Arrest Dhairya under the Narcotic Drugs and Psychotropic Substances Act
2. Arrest Rahul under the Madhya Pradesh Accommodation Control Act
3. Arrest Sunil and Shubham under Section 334 (voluntarily causing hurt on provocation)
4. Arrest Sunil and Shubham under Section 335 (voluntarily causing grievous hurt on provocation)
5. Arrest Rahul under Sections 503 and 506 (criminal intimidation)
6. Arrest Sunil and Shubham for mischief by damaging expensive property