

The Under-Reporting of Sexual Assault: An Experiment on Victims, Perpetrators, and Social Judgment*

Kennedy Johnston

August 2, 2026

Most Recent Version: [Click Here](#)

Abstract

Sexual assault imposes substantial economic costs, yet existing data likely understate its prevalence. I investigate how social judgment suppresses disclosure by victims and perpetrators. In an online experiment with 1,600 college students and graduates, respondents answer questions about rape and unwanted sexual contact using direct elicitation or “hard-garbling,” which provides plausible deniability. Among women, deniability increases reports of non-rape victimization by 23 percent. Among men, it doubles reports of rape perpetration, with 13.1 percent admitting perpetration under hard-garbling. I also find widespread beliefs that victims may be lying and misperceptions about victim-blaming prevalence, suggesting victims anticipate negative judgment for disclosing. *JEL Codes:* J16, K42

*Department of Economics, Texas A&M University. Email: kennedyjohnston@tamu.edu. I am grateful for the support and guidance of my advisor Danila Serra and my committee members Laura Dague, Catherine Eckel, and Jonathan Meer. I thank Abigail Adams, Laura Boudreau, Alex Brown, Marco Castillo, Jennifer Doleac, Rachel Heath, Joanna Lahey, Emily Nix, Ragan Petrie, Kaitlyn Sims, Daniel Tannenbaum, Clementine Van Effenterre, and Alessandra Voena for useful comments. I acknowledge funding from Arnold Ventures and from the John Van Huyck Fellowship through Texas A&M University. The study received IRB approval from the Texas A&M University IRB under study number STUDY2024-1379. The study is preregistered in the AEA Registry with ID AEARCTR-0015001.

1 Introduction

Survey evidence suggests that over 25 percent of undergraduate women have been sexually assaulted (Cantor et al., 2020), with victims facing long-term economic setbacks (Nix et al., 2026). Given these consequences, accurate measurement of sexual assault is critical for both policy and research. However, nearly two-thirds of sexual assaults are never reported to law enforcement (Morgan and Thompson, 2021). In the absence of reliable administrative data, institutions rely on survey data to guide prevention and support efforts. For example, ten leading U.S. institutions participated in the 2024 Higher Education Sexual Misconduct and Awareness (HESMA) survey, using an online survey to elicit anonymous self-reports of victimization and inform policy.¹ Yet even survey-based prevalence estimates among college women vary widely, ranging from 2 to 34 percent (Fedina et al., 2018), indicating substantial under-reporting and mismeasurement even in anonymous surveys meant to encourage disclosure. As a result, existing data provide a limited picture of both victimization and perpetration, impeding our understanding of who commits sexual assault, the broader costs to society, and the design of effective policies to protect victims and deter offenders.

Despite this limitations, little is known about decisions to report or disclose sexual assault experiences. For victims and perpetrators alike, reporting or disclosing experiences of sexual assault may carry large expected costs, like expectations of social disapproval and reputational harm. On college campuses where strong social ties are common, expected social costs and perceived social norms may play an especially important role in reporting and disclosure decisions.²

In this paper, I investigate the decision to report sexual assault victimization and perpetration and the role of social judgment through a conceptual framework and an online experiment involving over 1,600 college students and graduates. The experiment measures perceptions of relevant social norms and tests whether a survey technique designed to provide plausible deniability can increase self-reports of sexual assault. Unlike previous studies,

¹Similarly, employers and professional associations, including the American Economics Association, often administer workplace climate surveys to monitor workplace sexual harassment.

²These social forces are not unique to sexual assault: similar dynamics may arise in other settings where disclosure is socially costly, including corruption, workplace misconduct, intimate partner violence, and illegal drug use.

I examine not only reports of victimization (among women) but also reports of perpetration (among men).³ I use anonymous self-reports of victimization and perpetration to mirror the anonymous survey setting that institutions frequently use to measure sexual assault prevalence.⁴ These informal reports do not request information regarding the identity of the perpetrator (for victims) or victim (for perpetrators). This feature allows me to abstract away from fears of retaliation and other institutional barriers and better isolate expected social costs as a barrier to reporting.⁵

I first develop a simple framework for the informal reporting decision that considers both high-severity assaults (such as rape) and lower-severity assaults (such as unwanted kissing). The framework focuses on expected social costs for victims and perpetrators, shaped by two distinct types of norms: norms about the broader acceptability of assault, and norms about reporting. I distinguish between these to capture the possibility that society may disapprove of sexual assault yet still hold negative views of victims who report. The framework assumes that expected social costs are highest for low-severity victimization and high-severity perpetration: less severe incidents are more easily minimized by others, generating skepticism toward victims who report, while more severe assaults are harder to dismiss or justify, exposing perpetrators to greater social condemnation. This is consistent with evidence that more serious assaults are more likely to be reported (Spencer et al., 2020; FBI, 2026). The framework predicts that a plausible-deniability survey technique will be most effective where expected social costs are highest: for less severe victimizations, more severe perpetrations, and among individuals who are more sensitive to social judgment.

In the experiment, study participants are randomized to answer two binary questions on “more severe” assaults (i.e., rape) and less severe assaults (i.e., all other types of unwanted sexual contact) using either direct elicitation (DE) or “hard-garbling” (HG) (Boudreau et al., 2025; Warner, 1965). In the DE treatment, responses are recorded as the subjects intend. In the HG treatment, a subset of responses to the binary questions are randomly recorded as “Yes,” regardless of the true response. This randomization provides respondents with

³This choice was driven by statistical power: detecting comparable treatment effects among male victims would require substantially larger samples than feasible for this study. See Section 3 for more details.

⁴For an overview of sexual assault surveys, see Appendix Table A1.

⁵Even in this anonymous setting, respondents may be concerned about report leaks or may anticipate social judgment from researchers.

plausible deniability, as researchers cannot know for sure whether a “Yes” response was truly a “Yes” or an intended “No” randomly switched by the system. The experiment also contains an incentivized social norm elicitation exercise to measure two victim-focused norms. I first elicit individual agreement with two statements on whether victims are responsible for their assaults (“victim-blaming”) and the possibility that victims are lying (“lying”).⁶ I then elicit their perceptions about the extent to which other respondents agree with the statements.

The study generates two main findings. First, half of men and over a quarter of women in the study agree that women who say they were sexually assaulted may be lying, compared to only 16 percent of men and 8 percent of women who believe that women who were sexually assaulted may be to blame. Although men and women accurately guess others’ agreement with the lying statement, both genders drastically overestimate the rates of agreement with the victim-blaming statement. While the social norm exercise is descriptive in nature, both the widespread agreement with the lying statement and the misperception of the victim-blaming norm could amplify the anticipated social costs of reporting or disclosing for sexual assault victims.

Second, I find that hard-garbling is effective in encouraging reports of both experienced and perpetrated sexual assault, but along different margins for victims and perpetrators. Under direct elicitation, 45.6 percent of women report an unwanted sexual contact other than rape, and 34.3 percent report rape. Hard-garbling increases women’s reports of non-rape sexual assault victimization by 23 percent with no impacts on reports of rape. Conversely, hard-garbling has large effects on self-reported rape perpetration with no impacts on reports of less severe perpetration. While 6.2 percent of men in the direct elicitation treatment admit to having committed rape, hard-garbling more than doubles reports: 13.1 percent of men in the hard-garbling treatment admit to having raped someone. Thus, among victims, plausible deniability primarily raises reporting of less severe assaults, whereas among perpetrators it primarily raises reporting of the most severe behavior.

These findings align with the conceptual framework, which predicts that hard-garbling should most strongly increase reporting for incidents with the largest expected social costs

⁶These norms speak most directly to how victims might expect to be judged if they disclose an assault, and do not directly capture attitudes toward perpetrators or norms about admitting perpetration.

of disclosure. The framework further implies that expected social costs are highest for individuals who are more sensitive to social judgment. Using a social desirability scale as a proxy for this sensitivity, I find that the treatment effects of hard-garbling are concentrated among those more sensitive to social judgment. The results demonstrate that sexual assault experiences are under-reported even in anonymous, online survey contexts and provide evidence consistent with hard-garbling increasing reports by attenuating expectations of social judgment.⁷ More broadly, these findings suggest that anonymous surveys on socially costly behaviors may systematically understate true prevalence, and that survey design choices can meaningfully affect measured estimates of stigmatized behaviors.

To assess the validity of the experimental design, I conduct a replication of the experiment using a placebo outcome: theft victimization. In this setting, where individuals have little incentive to conceal the truth, hard-garbling should have no impact on reporting. I find that hard-garbling has no statistically significant effect on reporting for theft outcomes, consistent with hard-garbling increasing reports for only experiences with expectations of social judgment.

I contribute to three primary strands of literature: 1) the reporting and prevalence of sexual assault; 2) the hard-garbling survey technique and the measurement of sensitive information; and 3) the misperception of social norms. First, the literature on sexual assault focuses largely on incidence rather than reporting decisions. Studies document how alcohol consumption affects sexual assault rates, attributing changes to incidence rather than reporting behavior (Lindo et al., 2018; Topper, 2026), and show that rape victims experience significant long-run economic setbacks (Nix et al., 2026). A smaller set of papers studies reporting of sexual assault directly. The #MeToo movement increased formal reporting of sex crimes (Levy and Mattsson, 2024; Gauthier, 2024), and local support services in the UK raised formal reports (Denti and Iammarino, 2022). Similarly, the existing theoretical work (Lee and Suen, 2020; Chassang and Miquel, 2019) largely concerns formal reporting

⁷One alternative interpretation is that hard-garbling works through a generic privacy preference rather than social-judgment concerns specifically. This is inconsistent with the pattern of results. A pure privacy preference would predict the largest hard-garbling effects for the most sensitive disclosures, but I find no effect on reports of rape victimization, which is likely the most private experience in the survey. Effects are also concentrated among individuals more sensitive to social judgment, a pattern that a generic privacy preference would not predict.

to authorities.⁸ Relatively little is known about perpetration, the drivers of under-reporting in anonymous survey contexts, or methods to increase honest disclosure. I advance this literature first by directly eliciting self-reports of perpetration, contributing new evidence on the prevalence of sexual assault perpetration and the effectiveness of survey design in disclosing it. Second, by working in an anonymous, non-identifying survey setting and using a plausible-deniability technique, I can focus on social judgment as a barrier to reporting and am the first to experimentally test a specific survey design to increase reporting of sexual assault victimization and perpetration.

Second, I contribute to the literature on survey techniques for sensitive topics, including hard-garbling, randomized response, and list experiments. Previous work shows that hard-garbling can dramatically increase reports of workplace harassment in a Bangladeshi garment factory (Boudreau et al., 2025) and sustain informative reporting under high retaliation costs in laboratory settings (Chassang and Zehnder, 2024). Other approaches, such as list experiments and randomized response techniques, also introduce plausible deniability but can be undermined by non-compliance (Chuang et al., 2021). I extend this literature by evaluating hard-garbling in a novel setting: anonymous surveys on sexual assault victimization and perpetration. This setting limits fears of retaliation and other institutional barriers, allowing the estimated effects of hard-garbling to more directly reflect the influence of expected social judgment on disclosure behavior. Finally, existing studies focus on reports of victimization rather than perpetration. I contribute to the literature by testing whether hard-garbling can also increase self-reports of perpetration, providing new evidence on the role of plausible deniability in disclosing a range of socially undesirable behaviors.

Lastly, this study speaks broadly to the economics literature on the measurement and misperception of social norms (Bursztyn et al., 2023; Bursztyn et al., 2020; Bursztyn and Yang, 2022; Fields and Schuman, 1976; D. T. Miller and Prentice, 1994; Bordalo et al., 2016) and the literature on social norms related to violence against women and consent (Bermek et al., 2023; Hoover and Krupka, 2025; Seabright and Marti, 2022). However, the literature specifically addressing how social norms may impact reporting remains limited.

⁸Related work on other forms of violence against women, including sexual harassment, also largely focus on formal reporting or require that victims report against a specific perpetrator (A. Miller and Segal, 2019; Cheng and Hsiaw, 2022; Dahl and Knepper, 2026).

Related insights come from the work in psychology on “rape myth acceptance” (RMA), a framework encompassing beliefs that justify or downplay sexual violence (Lonsway and Fitzgerald, 1994).⁹ I add to this literature by quantifying perceptions of these “rape myths” as social norms and by investigating how they impact the reporting decision.

The remainder of this paper is as follows. Section 2 discusses the conceptual framework and characterizes the hard-garbling survey technique mathematically. Section 3 describes the experimental design, and Section 4 provides details on the estimation strategy. Section 5 discusses the primary results and exploratory analysis, and Section 6 describes robustness checks. Finally, Section 7 discusses and concludes.

2 Conceptual Framework

I model the decision to report sexual assault experiences using a simple costs and benefits framework. Prior work (Boudreau et al., 2025; Chassang and Miquel, 2019; Chassang and Zehnder, 2024) employs a principal-agent-monitor framework to study the decision to report bad agents, requiring victims to report a specific bad agent. Thus, fear of retaliation plays an important role in the reporting decision. I modify this framework by focusing on reporting that does not require identification of a perpetrator, like reporting in surveys. I refer to this type of disclosure as “informal” reporting. Because survey respondents are not asked about the identity of their perpetrator, I am able to abstract away from retaliation fears and focus on other barriers to reporting, like social disapproval.¹⁰

I model the victim’s informal reporting decision, then extend the framework to perpetrators in Section 2.1.1 and characterize the hard-garbling technique in Section 2.2. Online Appendix B discusses how the framework extends to formal reporting and the possibility of false reports.

⁹Rape myth acceptance, or RMA, is broadly defined as beliefs in ideas that justify violence against women. Many rape myth scales that are designed to measure rape myths consider six dimensions: 1) she was asking for it; 2) it was not really rape; 3) she is lying; 4) rape is a deviant event; 5) rape is a trivial event; and 6) he did not mean to (Lonsway and Fitzgerald, 1994).

¹⁰An act of retaliation by the perpetrator following an informal report could reveal their identity where it was otherwise hidden, likely eliminating incentives to retaliate.

2.1 Framework

Reporting decisions for sexual assault likely depend on incident characteristics, like severity. Less severe assaults are plausibly more prevalent yet less frequently reported than more severe assaults, consistent with FBI data and prior work in psychology (Fisher et al., 2003; Spencer et al., 2020; FBI, 2026). Thus, I distinguish between sexual assaults of high and low severity, s , with $s \in \{h, l\}$.¹¹ Denote $v_{is} = 1$ if student i has experienced a victimization of severity s and $v_{is} = 0$ if not, and let $r_{is} \in \{0, 1\}$ denote a binary reporting decision.

I assume that the utility function for a student i from an informal reporting decision r_{is} given a victimization status v_{is} is a function of costs and benefits:

$$U_{is}(r_{is}|v_{is}) = IB_{is}(r_{is}|v_{is}) + SB_{is}(r_{is}|v_{is}) - IC_{is}(r_{is}|v_{is}) - SC_{is}(r_{is}), \quad (1)$$

where:

- $IB_{is}(r_{is}|v_{is})$ denotes intrinsic or emotional benefits. I assume there are positive emotional benefits from a positive report of severity s given that a victimization of severity s occurred and that otherwise emotional benefits are zero. Thus, I assume $IB_{is}(1|1) > 0$, and $IB_{is}(1|0) = IB_{is}(0|0) = IB_{is}(0|1) = 0$.¹² Intrinsic benefits are likely larger for $s = h$: victims of more severe assaults plausibly gain more utility from feelings of empowerment and from protecting potential future victims
- $SB_{is}(r_{is}|v_{is})$ denotes social benefits. Following prior literature (Boudreau et al., 2025), these benefits include utility gained from informing policy.¹³ I assume that $SB_{is}(1|1) > 0$, $SB_{is}(1|0) < 0$, and $SB_{is}(0|0) = SB_{is}(0|1) = 0$.¹⁴ Similar to intrinsic benefits, we

¹¹This framework focuses on a binary of assault severity, but other dimensions of victims and incidents also matter for reporting. I abstract away from these for simplicity.

¹²This paper focuses on the role of social judgment in the decision to report. The assumption that $IB_{is}(1|0) = IB_{is}(0|0) = IB_{is}(0|1) = 0$ is made largely for simplicity. Relaxing this assumption would complicate notation but not fundamentally alter the resulting conclusions regarding the role of social costs and perceived social norms in the reporting decision.

¹³These benefits could also include social support after a positive report. However, an individual's expectations of social support are likely the inverse of their expectations of social costs (as an individual expects higher social costs and/or more damage to their reputation, they likely also expect less social support). Because this dynamic is captured by the social costs term $SC_{is}(r_{is})$, I make a simplifying assumption $SB_{is}(r_{is}|v_{is})$ is only utility from informing policy.

¹⁴I assume people have negative social utility from reporting a victimization when one has not occurred.

may expect social benefits to be larger for assaults of severity h .

- $IC_{is}(r_{is}|v_{is})$ denote intrinsic or emotional costs. IC capture emotional costs from acknowledging an experience as assault, identifying as a victim, or potential guilt from not reporting. I assume that $IC_{is}(1|1) \geq 0$, $IC_{is}(0|1) \geq 0$, and $IC_{is}(0|0) = IC_{is}(1|0) = 0$.
- $SC_{is}(r_{is})$ denotes social or reputation costs, including reputational damage and utility lost from negative social reactions. I assume $SC_{is}(1) > 0$ and $SC_{is}(0) = 0$ and discuss these costs in more detail below.¹⁵

I consider perceptions of two types of social norms shaping $SC_{is}(r_{is})$: 1) *social norms on sexual assault* (θ_{is}), capturing the broader social acceptability of assault, including what “counts” as assault and whether it is viewed as important; and 2) *social norms on reporting and perceptions of victims* (herein “social norms on reporting,” δ_{is}), capturing expectations of social judgment toward victims who report, including whether victims are perceived as dishonest or at fault. I distinguish between these because society may view assault as wrong yet still judge victims negatively for reporting—for instance, due to beliefs that perpetrators should not face serious consequences.

I assume that higher values of θ_{is} imply sexual assault is (perceived to be) viewed as more socially acceptable. Higher values of δ_{is} imply less victim-friendly perceived norms on reporting assaults of severity s .¹⁶ I assume both $0 \leq \delta_{is} \leq 1$ and $0 \leq \theta_{is} \leq 1$.

We should also expect social costs to be larger for individuals who care more about social judgment, i.e., for those who lose relatively more utility from having a damaged reputation or from negative social reactions. Thus, I denote a social judgment sensitivity parameter for individual i $\phi_i > 0$ such that higher values of ϕ_i denote more sensitivity to social judgment.

I assume that greater sensitivities to social judgment and greater expectations of negative

For example, consider the case in which individuals receive utility from informing policy. Then, if $v_{is} = 0$ but $r_{is} = 1$, policy-makers are now developing policy based on incorrect information, which I assume causes negative utility.

¹⁵I interpret the relevant audience for social judgment broadly. Expected social costs may arise from anticipated reactions of others or from judgment by the researcher administering the survey.

¹⁶Consider a social norm statement measuring victim-blaming. If a survey respondent perceives that 50% of people agree with the statement, then she expects higher levels of victim-blaming than someone who reports that 10% of people would agree. Thus, higher values of δ_{is} and θ_{is} denote “worse” perceived norms.

social judgments result in greater expected social costs.¹⁷ Thus, I let:

$$E[SC_{is}(r_{is})] = \phi_i K(\delta_{is}, \theta_{is}, r_{is}), \quad (2)$$

where δ_{is} is student i 's perceptions of social norms on reporting of sexual assaults of severity s , θ_{is} is student i 's perception of social norms regarding the acceptability of sexual assaults of severity s , ϕ_i is student i 's sensitivity to social judgment, and K is a positive and strictly increasing function of δ_{is} and θ_{is} .

Both δ_{is} and θ_{is} depend on the assault severity s . We may expect social norms on assault to be more favorable to victims of more severe assaults: rape is more likely to be socially condemned than sexual assaults like unwanted kissing. Therefore, we may expect $\theta_{il} \geq \theta_{ih}$. Similarly, reporting rape is less likely to be socially condemned than reporting less severe assaults. Thus, we may also expect $\delta_{il} \geq \delta_{ih}$. This implies that $SC_{is}(r_{is})$ may be larger for assaults of severity l .

Then, following Equation 1, an individual with $v_{is} = 1$ will find it optimal to report $r_{is} = 1$ when the expected utility from a positive report is larger than the expected utility from not reporting, i.e., when $U_{is}(1|1) \geq U_{is}(0|1)$. A victimized student has the following payoffs from $r_{is} = 1$ and $r_{is} = 0$:

$$\begin{aligned} U_{is}(1|1) &= IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - SC_{is}(1) \\ U_{is}(0|1) &= -IC_{is}(0|1) \end{aligned}$$

Thus, a victimized student will only submit $r_{is} = 1$ when¹⁸:

$$IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - \phi_i K(\delta_{is}, \theta_{is}, 1) + IC_{is}(0|1) \geq 0 \quad (3)$$

Equation (3) demonstrates that as δ_{is} , θ_{is} , or ϕ_i increase (i.e., as individuals perceive “worse” social norms or are more sensitive to judgment), expected social costs rise, making reporting less likely, all else equal.

¹⁷For example, victims who believe assault is “not a big deal” (norms on assault) or that accusations of lying are widespread (norms on reporting) will anticipate greater social disapproval.

¹⁸Expectations have been omitted from equation for simplicity of notation

2.1.1 Informal Reporting Decision for Perpetration

Let $a_{is} = 1$ if a student has perpetrated a sexual assault of severity $s \in \{h, l\}$ and $a_{is} = 0$ if not. A student i has a binary reporting decision for perpetration $z_{is} \in \{0, 1\}$. The utility function for reporting closely follows the victimization decision:

$$U_{is}(z_{is}|a_{is}) = IB_{is}(z_{is}|a_{is}) + SB_{is}(z_{is}|a_{is}) - IC_{is}(z_{is}|a_{is}) - SC_{is}(z_{is}) \quad (4)$$

Assumptions for costs and benefits are nearly identical to the victimization decision; I omit them here and discuss only deviations in expected social costs. I present the full framework in Online Appendix B.

We may expect only norms on assault to be relevant to a perpetrator's decision to report. Social norms on reporting, like those regarding whether women should report or whether victims are perceived as lying, will not be relevant when self-reporting an act of perpetration and thus δ_{is} is excluded from the expected social cost term.¹⁹ Norms on the broader acceptability of assault will likely still be relevant to a perpetrator's decision to self-report their behavior. If a perpetrator thinks that sexual assault is more acceptable and thus that he is less likely to be socially condemned for the behavior, his expected social costs likely decrease. Recall that higher values of θ_{is} imply that social norms such that sexual assault is viewed as more socially acceptable. For perpetrators, this now implies that higher values of θ_{is} *decrease* expected social costs. To maintain that higher values of the social norm parameter imply that social costs are increasing, I let $SC_{is}(z_{is})$ be a function of $\psi_{is} = 1 - \theta_{is}$. Therefore, I let:

$$E[SC_{is}(z_{is})] = \phi_i K(\psi_{is}, z_{is}), \quad (5)$$

where ψ_{is} are student i 's perceptions of social norms on assaults of severity s , ϕ_i is student i 's sensitivity to social judgment, and K is a positive and strictly increasing function of ψ_{is} .

¹⁹For example, if a perpetrator is admitting that he sexually assaulted someone, the victim will not somehow be accused of lying, and thus these norms should be irrelevant to the perpetrator's decision to report.

A perpetrator ($a_{is} = 1$ reports when (analogous to Equation 3):

$$IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - \phi_i K(\psi_{is}, 1) + IC_{is}(0|1) \geq 0 \quad (6)$$

Equation (6) then demonstrates that as individuals perceive that assault is less socially accepted and as a perpetrator is more sensitive to social judgment (i.e., as ψ_{is} and ϕ_i increase in value), expected social costs also increase. Thus, holding all else equal, individuals with higher values of ψ_{is} and ϕ_i are less likely to report.

2.2 A Survey Technique to Reduce Under-Reporting: Hard-Garbling

A survey technique that provides plausible deniability may attenuate expected social costs and thus increase rates of reporting. The “hard-garbling” survey technique, for binary reports r_{is} and z_{is} , automatically records a random subset of responses as “Yes” responses. No one, including researchers, can know whether an observed “Yes” response in the data was truly a “Yes” or an intended “No” randomly switched by the system. Thus, garbling may change the expected value of social costs by offering deniability.

This section characterizes the hard-garbling survey technique mathematically and closely follows the framework developed by Boudreau et al. (2025). I distinguish between intended reports r_{is} and z_{is} and recorded reports $\tilde{r}_{is}$ and $\tilde{z}_{is}$. An intended report r_{is} or z_{is} is the report the respondent intended to submit. A recorded report $\tilde{r}_{is}$ and $\tilde{z}_{is}$ is the report actually observed in the data. Under direct elicitation, $r_{is} = \tilde{r}_{is}$ and $z_{is} = \tilde{z}_{is}$. Under hard-garbling, if $r_{is} = 1$ or $z_{is} = 1$, then $\tilde{r}_{is} = 1$ and $\tilde{z}_{is} = 1$ with certainty, but if $r_{is} = 0$ or $z_{is} = 0$, then:

$$\tilde{r}_{is}, \tilde{z}_{is} = \begin{cases} 0 & \text{with probability } 1 - \pi \\ 1 & \text{with probability } \pi \end{cases}$$

Therefore, π is the level of garbling, i.e., the rate at which responses are randomly flipped to “Yes” responses. I assume reports are intended to be confidential, like in many large-scale surveys on sexual assault. A student now assumes their recorded report $\tilde{r}_{is}$ or $\tilde{z}_{is}$ is leaked with some perceived probability $p \in [0, 1]$. As in Boudreau et al. (2025), this does not require actual positive leaking probability. Even in anonymous or confidential surveys,

respondents may worry their report could be linked to identifying information (e.g., via IP address tracking).

Following Equations (1) and (4), a student i 's utility associated with an intended report r_{is} or z_{is} depends on their actual victimization or perpetration status v_{is} or a_{is} , their intended report r_{is} or z_{is} , and the recorded report $\tilde{r}_{is}$ or $\tilde{z}_{is}$:

$$\begin{aligned} U_{is}(r_{is}|v_{is}) &= IB_{is}(r_{is}|v_{is}) + SB_{is}(\tilde{r}_{is}|v_{is}) - IC_{is}(r_{is}|v_{is}) - p \times SC_{is}(\tilde{r}_{is}) \\ U_{is}(z_{is}|a_{is}) &= IB_{is}(z_{is}|a_{is}) + SB_{is}(\tilde{z}_{is}|a_{is}) - IC_{is}(z_{is}|a_{is}) - p \times SC_{is}(\tilde{z}_{is}) \end{aligned}$$

I assume social benefits and social costs are now functions of *realized* rather than intended reports.²⁰ Thus, $SC_{is}(\tilde{r}_{is})$ and $SC_{is}(\tilde{z}_{is})$ are reputation and social reaction costs in the event that the recorded report $\tilde{r}_{is}$ or $\tilde{z}_{is}$ is leaked. I assume that the social costs are a function of perceived social norms surrounding college sexual assault and the posterior belief that a student i meant to submit a positive report given a positive recorded report. Thus, I let $SC_{is}(\tilde{r}_{is}) = \phi_i K(\delta_{is}, \theta_{is}, \text{prob}(r_{is} = 1|\tilde{r}_{is}))$, where K is now a positive and strictly increasing function of δ_{is} , θ_{is} , and $\text{prob}(r_{is} = 1|\tilde{r}_{is})$. Analogously, I let $SC_{is}(\tilde{z}_{is}) = \phi_i K(\psi_{is}, \text{prob}(z_{is} = 1|\tilde{z}_{is}))$, where K is a positive and strictly increasing function of both ψ_{is} and $\text{prob}(z_{is} = 1|\tilde{z}_{is})$.

A victimized student with $v_{is} = 1$ then has the following expected utilities (notation has omitted the expectations for simplicity):

$$\begin{aligned} U_{is}(1|1) &= IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - p \times SC_{is}(1) \\ U_{is}(0|1) &= \pi \times [SB_{is}(1|1) - p \times SC_{is}(1)] - IC_{is}(0|1) \end{aligned} \tag{7}$$

Thus, a victimized student will report if and only if:

$$IB_{is}(1|1) + (1 - \pi) \times [SB_{is}(1|1) - p \times SC_{is}(1)] - IC_{is}(1|1) + IC_{is}(0|1) \geq 0 \tag{8}$$

²⁰Intrinsic costs are unchanged: hard-garbling still requires individuals to submit their intended response. Hard-garbling can therefore only attenuate utility costs tied to the recorded output — i.e., social costs. Mechanisms that operate at the answering stage, including emotional costs of recall and self-image concerns from acknowledging an experience to oneself, are unaffected by the hard-garbling technique.

where $SC_{is}(1)$ is:

$$SC_{is}(1) = \phi_i K(\delta_{is}, \theta_{is}, prob(r_{is} = 1 | \tilde{r}_{is} = 1)) , \text{ and}$$

$$prob(r_{is} = 1 | \tilde{r}_{is} = 1) = \frac{1}{1 + \pi \frac{prob(r_{is}=0)}{1-prob(r_{is}=0)}} \quad (9)$$

The perpetrator's reporting decision under garbling is analogous and omitted for brevity.

2.3 Predictions

The conceptual framework reveals expected social costs as a driver of under-reporting and provides the basis for four testable implications. Prediction 1 discusses the effectiveness of hard-garbling broadly, and Predictions 2 to 4 focus on when expected social costs may be highest and thus when hard-garbling may be more effective.

Prediction 1: (The effectiveness of hard-garbling)

The hard-garbling survey technique with $\pi = 0.2$ increases reporting relative to a standard direct elicitation survey technique with $\pi = 0$.

This prediction comes directly from the framework. Equation 9 demonstrates that increasing π reduces the size of the expected social costs associated with a positive recorded report for both victims and perpetrators by providing plausible deniability, thus increasing the likelihood that an individual sends $r_{is} = 1$ and $z_{is} = 1$ (Boudreau et al., 2025).²¹ In this study, I set $\pi = 0.2$, following Boudreau et al. (2025).

Prediction 2: (The role of social norms on assault θ_{is} and ψ_{is})

2a) *Hard-garbling is more effective at increasing reports of victimization when θ_{is} is larger, i.e., when $s = l$.*

2b) *Hard-garbling is more effective at increasing reports of perpetrations when ψ_{is} is larger, i.e., when $s = h$.*

²¹Increasing π also reduces the weight on social benefits in the reporting condition (Equation 8). Prediction 1 thus presumes the social-cost reduction from plausible deniability dominates this offsetting effect, which is consistent with the sexual assault context, where social benefits of disclosure are plausibly small.

The size of the expected social costs term depends on assault severity s . Rape may be broadly viewed as less socially acceptable than other types of sexual assault, implying $\theta_{il} > \theta_{ih}$. The plausible deniability offered by hard-garbling may thus be more effective at increasing reports for the severities with higher expected social costs: less severe victimizations (higher θ_{is}) and more severe perpetrations (higher ψ_{is}).

Prediction 3: (The role of sensitivity to social judgment ϕ_i)

Hard-garbling is more effective for individuals more sensitive to social judgment, i.e., those with higher values of ϕ_i .

Greater sensitivity to social judgment increases the size of expected social costs. Thus, we may expect the plausible deniability induced by hard-garbling to be most effective for individuals who are more sensitive to social judgment.

Prediction 4: (The role of social norms on reporting δ_{is} for victims)

Hard-garbling is more effective for victims who expect relatively more negative social judgment, i.e., those with higher values of δ_{is} .

While social norms on reporting δ_{is} do not enter the perpetrator’s reporting decision, victims with higher values of δ_{is} face higher expected social costs. Thus, hard-garbling may be more effective at increasing reports of victimization for these victims.

Predictions 1 and 3 follow directly from the conceptual framework without additional assumptions; the proof of Prediction 3 is in Online Appendix B.4.²² Predictions 2 and 4, that hard-garbling is more effective where norm-driven social costs are higher, require that K is multiplicatively separable in the norm parameters and the posterior belief:

$$E[SC_{is}(\tilde{r}_{is})] = \phi_i K(\delta_{is}, \theta_{is}) \times \text{prob}(r_{is} = 1 | \tilde{r}_{is}) \quad (10)$$

This implies that the reduction in social costs from plausible deniability scales proportionally with perceived norms: individuals facing “worse” perceived norms benefit more from

²²The proof of Prediction 3 is presented in Online Appendix B.4 under multiplicative separability for notational continuity with the proofs of Predictions 2 and 4, but does not rely on this assumption.

the same level of deniability. This is sufficient but not necessary for the qualitative predictions. A formal proof is in Online Appendix B.4.

3 Experimental Design

I employ an online experiment involving 1,621 individuals that are college students and recent college graduates, aged 18-27, that were recruited using Prolific (an online platform of survey takers) to: 1) quantify perceptions of social norms relevant to the sexual assault reporting decision, and 2) test whether the HG survey technique increases self-reports of sexual assault experiences. In this between-subjects experiment with two treatments, subjects complete two primary modules on their sexual assault experiences and social norms.

Motivated by my conceptual framework, I provide proxies for the parameters included in the expected social costs term (δ_{is} , θ_{is} , and ϕ_i). Specifically, I measure participants' perceptions of social norms around victim-blaming and false accusations as a proxy for perceptions of social norms on reporting δ_{is} . To indirectly proxy for perceptions of social norms on assault θ_{is} , my experimental design contains two binary experience questions. The first pertains to rape, which is widely recognized as the least socially accepted form of assault. The second pertains to all other unwanted sexual contact, which is likely perceived as more socially accepted. Thus, the experiment contains variation in the social acceptability of assaults, allowing me to explore how social norms on assault θ_{is} may impact the reporting decision. Finally, I use a social desirability scale (Crowne and Marlowe, 1960; Dhar et al., 2022), to capture individual sensitivity to social judgment ϕ_i . This scale asks participants to agree or disagree with a set of "too-good-to-be-true" statements; individuals who agree with more may be more sensitive to social judgment.

All subjects begin the experiment by answering a brief demographic section, which also contains the social desirability scale used to generate the proxy for sensitivity to social judgment ϕ_i . At the end of the survey, all subjects saw two questions regarding their perceptions of the anonymity of the survey platform Prolific.²³

²³These two questions are as follows: 1) "What is the probability that Prolific shares your identifying information, like name or birth date, with researchers?," and 2) "What is the probability that researchers using Prolific to administer surveys are able to link your survey responses to your identity?"

I vary which survey technique is used to elicit sexual assault experiences in the assault module. Thus, I randomly assign participants into one of two treatment conditions:²⁴

- **T1:** Direct Elicitation (DE)
- **T2:** Hard-Garbling (HG)

I focus on women as potential victims and men as potential perpetrators in the experiment.²⁵ While this assumption does not capture the full complexity of sexual assault dynamics, they are broadly consistent with national data: in rapes reported from January 2021 to January 2026, nearly 90 (85) percent of victims (offenders) were female (male) (FBI, 2026). The experiment is described in more detail below. The full set of survey questions can be found in Online Appendix [D](#).

3.1 Social Norm Module

In the social norm module, subjects complete an incentivized social norm elicitation exercise. First, participants state their agreement or disagreement with two statements on a 5-point Likert scale from strongly disagree to strongly agree. These statements, herein referred to the “victim-blaming” or “lying” statements, are as follows:²⁶

1. **Victim-Blaming:** “If a woman experiences unwanted sexual contact, she may be at least somewhat responsible for what happened.”
2. **Lying:** “Women who say they were sexually assaulted sometimes lie. For instance, they may just regret otherwise consensual encounters.”

²⁴The hard-garbling survey technique is used only when eliciting assault experiences and is not implemented in the social norm elicitation exercise.

²⁵This decision was made predominantly for statistical power, though it is important to recognize that men can also be victims, and women can also be perpetrators. Given lower victimization rates among men, the most optimistic power calculation reveals that one would need *at least* 1.5-times more men to detect the same 20 percent increase in reporting (using sample means and standard deviations among male and female respondents from Cantor et al. (2020)). A more realistic power calculation that accounts for the mechanical increase in noise as a result of the hard-garbling treatment manipulation estimates a total required sample size of at least 5,000 men.

²⁶There are other beliefs and norms that likely matter to sexual assault reporting. However, this study aims to understand the under-reporting of sexual assault in the most general sense. Thus, I focus on two norms that may be relevant to all cases of sexual assault victimization. In addition, agreement is binarized for analysis. A respondent is considered to have agreed with a statement if they reported agreeing or strongly agreeing with the statements.

Participants are then asked to guess the percentage of previous male and female respondents who agreed or strongly agreed with each statement. I randomize the order in which subjects are asked about men and women’s agreement and which statement they see first. As victims expect more agreement with both the victim-blaming and lying statements, they likely also expect more social judgment for reporting. Thus, I use these predictions as a proxy for social norms on reporting, δ_{is} .²⁷

3.2 Experience of Assault Module

All subjects complete the experience module first. In the experience module, female respondents answer questions on victimization; male respondents answer analogous questions on perpetration. Subjects are provided with a definition of non-consent, including descriptions of incapacitation, physical force, and sexual coercion. The conceptual framework assumes that the perceived social acceptability of assault θ_{is} may vary by assault severity such that $\theta_{il} > \theta_{ih}$. Thus, I distinguish between two severities of assault: the most severe form (rape, $s = h$), and all other forms of sexual assault (including unwanted kissing and groping, $s = l$). Subjects answer two binary questions each. The first pertains to non-consensual sexual penetration or oral sex (i.e., rape). The second binary question asks about any other non-consensual sexual contact, including unwanted kissing or sexual touching.²⁸

Subjects are randomly assigned to answer the two binary questions using either direct elicitation or “hard-garbling” (HG). In the HG treatment, a subset of responses to the binary questions are randomly recorded as “Yes,” regardless of the true response. This randomization provides respondents with plausible deniability, as researchers – or anyone else viewing the data – cannot know for sure whether a “Yes” response was truly a “Yes” or an intended “No” randomly switched by the system. In the experiment, the HG survey technique was explained to subjects in the HG treatment using visual aids, a written description, and two

²⁷Subjects also guess what percentage of women are sexually assaulted while in their undergraduate programs, providing their best guess at a victimization rate. They are then asked to guess the average answer of prior male and female respondents to this question.

²⁸The definitions of non-consent, rape, and non-consensual sexual contact used in this study were based heavily on existing large-scale surveys, such as the Association of American Universities Climate Survey on Sexual Assault and Sexual Misconduct (Cantor et al., 2020), which was conducted in 2019 and involved 33 universities and over 180,000 students.

incentivized comprehension questions.²⁹

The rate of garbling in this experiment is set at 20%, meaning that 1 of every 5 responses is automatically recorded as a “Yes” response for individuals in hard-garbling treatments. This rate is 0% for individuals in direct elicitation treatments.

3.3 Implementation

This study was pre-registered in the AEA Pre-Registry (AEARCTR-0015001).³⁰ I recruit all participants from Prolific, an online and anonymous platform of survey-takers. The experiment was programmed using Qualtrics, and subjects were paid \$2.75 for participation, with a possible bonus payment of up to \$1.50 for correct guesses on social norm questions. Individuals in the hard-garbling treatment also answered two incentivized comprehension questions. Subjects were paid an additional \$0.15 for each correct answer, for an additional bonus payment of up to \$0.30. The average completion time was 15.6 minutes, and the average bonus payment was just under \$0.25.³¹ Thus, the average total payment to participants was approximately \$3.00, putting the total hourly earnings at nearly \$12/hour. Preliminary data collection began in December 2024, though the majority of responses were collected in February and March 2025.

This study aimed to investigate college sexual assault. Therefore, I recruited only individuals who met the following criteria: 1) current Bachelor’s degree students or recent college graduates; 2) between 18 and 27 years old; and 3) currently reside in the United States. For the victimization outcomes, I recruited only individuals whose biological sex was female, and for the perpetration outcomes, I recruited only individuals whose biological sex was male. Further details on the sample are provided in Section 5.

²⁹Respondents assigned to hard-garbling are aware that their answers may be garbled, but respondents who answer “No” are not ultimately aware of how their response is recorded.

³⁰The study was initially pre-registered on December 16, 2024 as a 2x2 design in which I varied both the survey technique used and the order in which subjects completed the two tasks. The order in which subjects completed the social norm task was intended to be a proxy for the salience of social norms. Preliminary data collected in December 2024 indicated that the original design was ineffective in manipulating the salience of perceived social norms. I therefore changed the design so that the social norms task always comes after the experience task. As a result, the study is now a 2x1 online experiment, and the pre-registration was modified to reflect this change in February 2025.

³¹The average bonus payment in hard-garbling treatments was \$0.38, and the average bonus payment in direct elicitation treatments was \$0.10.

A small-scale pilot study of approximately 100 men and 100 women was conducted in November 2024 for the purpose of incentivizing the social norm task. In this pilot, subjects were asked to indicate their agreement with the victim-blaming and lying statements and were asked what percent of women they think are sexually assaulted while in their undergraduate programs. Subjects in the full study are paid \$0.25 for each correct (within 2 percentage point) guess regarding men and women’s answers from the pilot study, for an additional bonus payment of up to \$1.50.

A total of 1,639 current undergraduate students and recent graduates participated in the experiment. The data quality is high, and Online Appendix Table A2 displays the number of missed attention checks and failed comprehension questions. Survey takers were highly attentive; over 90% of respondents passed all three attention checks included in the survey. In addition, individuals in the hard-garbling treatments largely understood the survey technique. Over 90% of respondents in HG correctly answered both comprehension questions. In line with my pre-specified analysis plan, subjects who incorrectly answered both hard-garbling comprehension questions have been excluded from analysis, resulting in a final sample size of 825 women and 796 men.

3.4 External Validity

To address concerns about external validity, I organize this section according to the SANS framework on selection, attrition, naturalness, and scaling (List, 2020).

3.4.1 Selection

Prolific allows me to recruit a diverse set of students and graduates from institutions of various sizes and locations, providing greater certainty that results are not driven by a single outlier institution. However, participants may differ from the wider college student population through selection into Prolific or into this study. Descriptive statistics are reported in Online Appendix Table A3, and balance tests are discussed in Section 5.

To assess representativeness, I compare my sample characteristics with national data. My sample appears broadly representative: the share of white students and public institu-

tion enrollment is consistent with National Center for Education Statistics data on undergraduate enrollment for fall 2021 (National Center for Education Statistics, 2023), and the first-generation student rate (28%) aligns with national estimates from the Postsecondary National Policy Institute (Postsecondary National Policy Institute, 2025). The share of Pell Grant recipients in my sample (49%) is somewhat higher than the national rate of 40% (Cameron et al., 2023), suggesting a slightly more low-income sample. However, heterogeneity analysis (Section 5.3) indicates that primary treatment effects do not differ significantly by income.³² This difference is therefore unlikely to substantively limit generalizability.

To further address concerns about selection into Prolific, I also conduct a small-scale replication of the experiment with students from Texas A&M University, a large four-year public university in the South. This replication exercise is discussed in more detail in Section 6.2. While under-powered, the results of this replication provide broad evidence of the generalizability of my findings.

To mitigate concerns that individuals selecting into a study on sexual assault might differ from those unwilling to do so, the study was described on Prolific as a survey on college campus climate rather than sexual assault. The title and description referred to common college campus crimes broadly but did not mention sexual violence specifically, minimizing self-selection based on interest in or prior experience with sexual assault.

3.4.2 Attrition

There may be concerns that individuals differentially exited the survey before full completion based on their prior sexual assault experiences, which could bias the treatment effects. While I discuss attrition in more detail in Section 6, attrition was largely not present in the study, as only 1% of respondents who started the survey did not complete it.

3.4.3 Naturalness

Naturalness refers to the extent to which experimental decisions mirror those made in the target setting. The use of an anonymous and online survey experiment to study reports of

³²Treatment effects are more pronounced among higher-income students, particularly for perpetration. Thus, given the higher proportion of low-income students in my sample, overall estimates may understate effects in a more nationally representative population.

sexual assault could be viewed as a limitation of this study. However, the reports used in the experiment mirror those routinely used by universities to measure sexual assault incidence and guide prevention and support efforts.³³ For example, ten prominent U.S. institutions, including Harvard, Stanford, and Yale, recently conducted the 2024 Higher Education Sexual Misconduct and Awareness (HESMA) survey, which included binary and anonymous reports of sexual assault, similar to those collected in this study. The questions used to measure sexual assault in this study were designed to be consistent with the questions commonly used in “real-world” large scale campus surveys.

While anonymous survey reports do not directly generalize to higher-stakes contexts like law enforcement or Title IX offices, improving formal reporting first requires understanding barriers to disclosure in the lowest-risk environment. If victims choose not to disclose in an anonymous survey, where perceived costs are minimal, they are unlikely to come forward in higher-stakes settings. By identifying what prevents disclosure even in this conservative setting, this study offers a foundation for designing interventions that encourage reporting in both surveys and more consequential contexts.

There may also be concerns in this setting about experimenter demand effects. However, recent work finds little evidence of experimental demand effects in online survey experiments: knowing the purpose of an experiment has essentially no impact on a participant’s behavior (de Quidt et al., 2018, Mummolo and Peterson, 2019). Because hard-garbling should have no impacts on reporting for an outcome in which individuals have no incentive to conceal their true answers, I run additional treatments using theft victimizations as a placebo outcome to further address concerns of experimenter demand (discussed further in Section 6.1).

3.4.4 Scaling

The hard-garbling survey technique is a practical tool and can be easily implemented for organizations seeking more accurate measurement of sensitive experiences. For example, this study implemented the hard-garbling survey technique using the survey platform Qualtrics.

³³Anonymous and informal reports are also used to study other types of violence against women, including sexual harassment. For example, the American Economic Association (AEA) conducted an anonymous survey of members in fall 2023 that included various measures of sexual harassment, elicited using a direct elicitation survey technique.

Because it can be embedded seamlessly into existing survey instruments, institutions can adopt this method without major changes to their current data collection practices.

4 Estimation Strategy

For individuals in hard-garbling treatments, I observe garbled responses instead of intended responses. Following Boudreau et al. (2025) and Blair et al. (2015), I normalize the observed garbled responses. Consequently, OLS estimates are consistent, and robust standard errors are correct.

Thus, following prior literature, I define normalized recorded reports $\hat{r}_i$ as:

$$\hat{r}_i = \frac{\tilde{r}_i - \pi}{1 - \pi}, \quad (11)$$

where $\tilde{r}_i$ is the garbled report, and π is the rate of garbling, equal to 0.2 for HG and 0 for DE.

These normalized recorded reports are used as the primary outcomes. Note that all men are asked about sexual assault perpetration, and all women are asked about sexual assault victimization. Thus, the primary outcomes in this study are reports of:

1. Rape victimizations (*women only*)
2. Other non-consensual sexual contact victimizations (*women only*)
3. Rape perpetrations (*men only*)
4. Other non-consensual sexual contact perpetrations (*men only*)

Then, I estimate the following equation:

$$\hat{r}_i = \beta_0 + \beta_1 HG_i + \delta X_i + \epsilon_i, \quad (12)$$

where HG_i is an indicator for being in the hard-garbling treatment, and X_i is a vector of individual controls, selected using the Double Lasso method (Belloni et al., 2014).³⁴

³⁴Controls are selected from the list of characteristics in Table 1 and survey-generated measures of percep-

I also test for heterogeneity on subject characteristics, including their sensitivity to social judgment and perception of social norms. For example, to test Prediction 3 from the conceptual framework regarding the role of sensitivity to social judgment (proxied for using a social desirability scale), I estimate a specification with indicators for both treatment, their sensitivity to judgment, and an interaction term:

$$\hat{r}_i = \beta_0 + \beta_1 HG_i + \beta_2 Judge_i + \beta_3 HG_i \times Judge_i + \delta X_i + \epsilon_i, \quad (13)$$

where HG_i is an indicator for being in the HG treatment, $Judge_i$ is an indicator equal to one if subject i is less sensitive to social judgment, and $HG_i \times Judge_i$ is an interaction term.

5 Results

The experiment has a final sample size of 1,621 respondents, including 825 women and 796 men. Table 1 shows balance across treatments for demographic characteristics for women and men, respectively. The average respondent is 23 years old, though male respondents are slightly older on average. The sample is approximately 55% white, and half of the sample are current undergraduate students. Male respondents in the sample are less likely to have attended undergraduate institutions in the South and less likely to be associated with the Democratic political party, relative to female respondents in the sample.

One characteristic out of the 11 survey measures reported is not balanced across treatments for women, and two characteristics for men. For female respondents, those in HG were less likely to indicate being low-income than those in DE. For male respondents, those in HG were more likely to have attended undergraduate institutions located in the South and more likely to have attended large institutions than those in DE. The joint F-test p-value for equality of all characteristics reported in Table 1 is 0.027 for female respondents and 0.165 for male respondents. However, for female respondents, this is driven predominantly

tions of social norms and sexual assault prevalence, agreement with the social norm statements, sensitivity to judgment as measured by a social desirability scale, beliefs that their survey responses were anonymous, and an indicator for whether the response was collected as part of the preliminary December 2024 data collection. These additional characteristics are balanced between HG and DE treatments for both men and women, and tests of equality of means and Kolmogorov-Smirnov tests for distributional equality across HG and DE treatments are non-significant ($p > 0.05$ for all characteristics).

by low-income status. When dropping this variable from the F-test, the p-value increases to 0.217. While I select controls in all regressions using the Double Lasso method (Belloni et al., 2014), I always include this low-income indicator as a control in regressions including female respondents from the Prolific sample due to this imbalance, unless otherwise specified.

5.1 Analysis of Social Norms

Following Bursztyn et al. (2023), an *actual* social norm for a given behavior or action is defined as the share of respondents who agreed with a statement about that behavior of interest. An individual’s *perceived* social norm is the percent of respondents they think agreed with a statement about that behavior of interest. An individual’s misperception of the norm is calculated as the difference between their perceived norm and the actual norm.

To provide an estimate for δ_{is} (social norms on reporting and perceptions of victims), the study contained two social norm statements of interest: the victim-blaming and lying statements. Subjects rated their agreement with each statement using a 5-point Likert scale, though agreement has been binarized for analysis.³⁵ Subjects also guessed the percent of prior male and prior female participants that they believe agreed or strongly agreed with each statement.³⁶ Figure 1 visualizes the average answers from all 825 female and 796 male respondents, and this data is reported in Online Appendix Table A4.³⁷ Women and men differ in both their levels of agreement with the statements and their perceptions of the norms. Sixteen percent of men and 8 percent of women agree with the victim-blaming statement, whereas 50 percent of men and 28 percent of women agree with the lying statement.

While purely descriptive, the differences in agreement across statements reveal an interesting dynamic. The lying statement captures whether an individual thinks a survivor might be lying, while the victim-blaming statement assumes the assault occurred and captures be-

³⁵A respondent is considered to have agreed with a statement if they reported agreeing or strongly agreeing with the statements.

³⁶The survey also contained a question on the descriptive norm of sexual assault occurrence in which subjects were asked what percent of undergraduate women they believe are sexually assaulted. Online Appendix Figure A2 displays the average answers from all respondents.

³⁷While a pilot study was conducted to incentivize the social norm elicitation exercise, all results reported are based on responses from the primary sample of 825 female and 796 male respondents. “Actual Agreement” refers to the rate of agree or strongly agree responses with each statement from female or male respondents. Rates of agreement between the pilot sample and primary sample were qualitatively similar.

liefs about fault. The variation between these may reflect extensive and intensive margins of judgment: does an individual believe an assault occurred, and if so, do they find the victim at fault? These results suggest high barriers for survivors to be believed, but conditional on being believed, victims are unlikely to be blamed.³⁸

While actual agreement with the victim-blaming statement is fairly low, there are large misperceptions of this social norm. On average, men expect 40 percent of men and 20 percent of women to agree, and women expect nearly 50 percent of men and 24 percent of women to agree. This is a clear distinction from the lying norm, in which respondents are far more accurate in their perceptions of others' agreement. On average, both men and women expect just over 50 percent of men and just under a third of women to agree—implying there are few misperceptions of the lying social norm on aggregate.³⁹ Following Bursztyn et al. (2023), I explore three mechanisms behind social norm misperception: false consensus, minority overweighting, and gender stereotyping. False consensus is a positive correlation between an individual's opinion and their expected prevalence of this opinion among others (Bursztyn et al., 2023; Fields and Schuman, 1976), while minority overweighting is overestimating the proportion holding a minority opinion (Bursztyn et al., 2023; D. T. Miller and Prentice, 1994). Gender stereotyping involves overestimating gender differences in agreement (Bursztyn et al., 2023; Bordalo et al., 2016). This analysis is reported in Online Appendix C.

5.2 The Impact of Hard-Garbling on Reporting

Respondents were asked two binary questions regarding their sexual assault histories: one on non-consensual sexual penetration or oral sex ("rape"), and one on any other non-consensual sexual contact, such as unwanted kissing or touching ("other SA"). In the direct elicitation treatment, 34.3 percent of women report rape victimization, and 45.6 percent of women report victimization for other types of sexual assault. Additionally, 6.2 percent of men report rape perpetration, and 9.0 percent report perpetration of other types of sexual assault.

Table 2 reports the main treatment effects. There are two primary findings. First, women

³⁸This dynamic may be more complicated in other settings, as individuals may use victim-blaming sentiments as a way to accuse survivors of lying, and these beliefs are likely not entirely separable.

³⁹Online Appendix Table A5 displays the difference in guesses between male and female respondents: women expect higher agreement among others in all cases.

in the hard-garbling treatment are 10.5 percentage points more likely to report a non-rape sexual assault, which represents a 23 percent increase in reporting over the direct elicitation mean of 45.6 percent. Hard-garbling has no impacts on reports of rape victimization. Second, men in hard-garbling are 6.9 percentage points more likely to report a rape perpetration, which is an 111 percent increase relative to the direct elicitation mean of 6.2 percent. While noisy and marginally insignificant, hard-garbling increases reports of non-rape perpetrations by 5.0 percentage points, a 55 percent increase over the direct elicitation mean of 9 percent. These findings are in line with the conceptual framework, which assumes that victims expect more social backlash from reporting more socially-accepted incidents that others are more likely to dismiss or minimize, like unwanted kissing, while perpetrators expect more backlash for admitting more serious, socially condemned assaults. As a result, hard-garbling is most effective at attenuating expectations of social judgment when these costs are higher, i.e., for non-rape victimization and rape perpetration.

5.3 Exploratory Analysis

Informed by the conceptual framework, I explore heterogeneous treatment effects by subject characteristics. The framework predicts that hard-garbling is more effective when social costs are highest, i.e., for individuals more sensitive to social judgment and for victims who expect more social disapproval for reporting. I proxy for these using a social desirability scale and social norm perception measures. Results are presented here and discussed in Section 7.

5.3.1 Sensitivity to Judgment

To proxy for sensitivity to social judgment (ϕ_i), I use a social desirability scale (Crowne and Marlowe, 1960; Dhar et al., 2022), which includes statements regarding “too-good-to-be-true” traits (e.g., “No matter who I’m talking to, I’m always a good listener”).⁴⁰ Individuals reporting more of these traits may be more likely to give socially desirable answers. Thus, these individuals may also react more to social judgment, benefiting more from the plausible deniability offered by hard-garbling. Following Dhar et al., 2022, I count the number of

⁴⁰The full scale is reported in Online Appendix Table A7.

traits each individual reports. Then, I create a binary indicator equal to 1 if an individual reports fewer than the within-gender average number of traits.⁴¹

I estimate this heterogeneity using the specification in Equation 13, and the results are reported in Table 3. In the direct elicitation treatment, individuals less sensitive to social judgment tend to report victimization and perpetration more frequently, though these relationships are not statistically significant. For those more sensitive to social judgment, hard-garbling has a 16.5 percentage point increase in reports of less severe victimization, a 9.6 percentage point increase in reports of rape perpetration, and a 9.4 percentage point increase in reports of less severe perpetration. These effects represent increases in reporting of 39 percent, 192 percent, and 118 percent, respectively, over the direct elicitation means for individuals with high sensitivity to judgment. Hard-garbling has no statistically significant impacts on reporting for individuals with low sensitivity to judgment.

5.3.2 Expectations of Social Judgment

I measure expectations of social judgment (δ_{is}) using the four incentivized social norm predictions: women’s and men’s expected agreement with victim-blaming and lying statements. I create a binary indicator for “high” and “low” expectations of social judgment towards victims, where “high” corresponds with expecting relatively more agreement with the statements.⁴² The conceptual framework does not include norms on victim reporting in the perpetration decision—e.g., accusations of lying towards victims should be irrelevant for a man’s self-report of perpetration. Thus, we should not expect meaningful heterogeneity for perpetrators along this dimension; for completeness, I report these results in Online Appendix Tables A8 and A9.

While there is no meaningful heterogeneity on the basis of this expectation of judgment indicator alone (Online Appendix Table A10), the impacts of hard-garbling are most pronounced among individuals who are both sensitive to social judgment and expect more

⁴¹Figures showing the full distribution of social desirability scores by gender can be found in Online Appendix Figure A1.

⁴²I aggregate and standardize all four predictions around the control mean, create an index using the Anderson method, and set the indicator equal to 1 for individuals at or above the within-gender median. All four predictions are balanced across treatments ($p > 0.05$ for tests of equality of means and Kolmogorov-Smirnov tests). A specification using the continuous index is reported in Online Appendix Table A6 with broadly similar results.

judgment. Using Equation 13, I split the sample of female respondents into two groups: those with higher expectations of judgment and those with lower expectations of judgment. The results of this exercise are reported in Table 4. For respondents who expect less judgment, there is no meaningful heterogeneity on the basis of sensitivity to judgment. However, for respondents who expect more judgment and who are sensitive to it, hard-garbling increases reports of less severe victimization by 19.3 percentage points. Further, hard-garbling has no impact on reports for those who expect more judgment but are less sensitive to it.

5.3.3 Additional Heterogeneity Analysis

Hard-garbling may be more effective for individuals in social communities with stronger or more salient social norms. For instance, close-knit groups like undergraduates, athletes, or Greek life members may respond differently. Geographic region and political affiliation may also shape social norms around sexual assault, potentially leading to heterogeneous treatment effects. I conduct heterogeneity on these margins and report the results of this exercise in Online Appendix Figures A3 and A4. The effects of hard-garbling on non-rape victimizations and rape perpetrations are broadly persistent across these subgroups, with no evidence of heterogeneous treatment effects, though the estimates are noisier.⁴³

6 Robustness Checks

In this section, I conduct and discuss a number of robustness checks. First, to rule out the possibility that the hard-garbling findings are driven by experimenter demand, I replicate the experiment with a placebo outcome – theft – for which respondents should have no incentive to conceal their true answers. I discuss my findings in Section 6.1. Second, to provide additional evidence of external validity, I conducted a replication of the primary survey experiment using a sample of current undergraduate students at a large Southern university. This replication exercise is discussed in Section 6.2. Finally, in Section 6.3, I

⁴³As discussed in Section 3.4, Prolific respondents are slightly more likely to have received Pell grants (i.e., to be low-income) than the broader college student population. Heterogeneity on this margin reveals that the treatment effects are not statistically different for individuals who report being low-income relative to those who do not, though the treatment effects are more pronounced among *higher* income students, particularly for perpetration outcomes.

conduct additional robustness checks by varying the set of controls in the regression analyses, dropping respondents who displayed low levels of comprehension and attention, accounting for attrition, and validating the existence of the non-zero leak probability $p > 0$ required by the mathematical framework for hard-garbling.

6.1 Robustness Check: Placebo Theft Experiment

Hard-garbling should have no impacts on reporting for an outcome in which individuals have no incentive to conceal their true answers. As a robustness check, I repeat the experiment for a placebo outcome: theft victimization. Unlike reporting sexual assault victimization, there is little reason for a respondent to hide whether they have been victims of theft.

The placebo theft victimization treatments are designed to be analogous to the sexual assault victimization treatments, involving women aged 18–27 that are current college students or recent college graduates ($N=306$). Respondents answer two binary questions on theft victimizations of different severities: (1) theft or burglary of property worth over \$1,000 (“severe”), and (2) theft or burglary worth under \$1,000 (“less severe”). Additional design details, including the survey questionnaire, can be found in Online Appendix E. Balance on observable characteristics is shown in Online Appendix Table A11.

In Table 5, I estimate Equation 12 separately for each of the four outcomes: rape victimizations, non-rape victimizations, severe theft, and less-severe theft. Rates of occurrence are similar across analogous outcomes (e.g., 34.3% reported rape victimizations in DE vs. 35.9% for severe theft). I find no impacts of hard-garbling on either theft outcome, providing additional evidence that experimenter demand effects are not driving the observed treatment effects. Following my pre-specified analysis plan, I also pool outcomes by severity across sexual assault and theft, interacting treatment with an indicator for sexual assault (Online Appendix Table A12), and again find no effects on theft outcomes.

6.2 Robustness Check: TAMU Replication

To assess generalizability, I conduct a replication with undergraduate students at Texas A&M University (TAMU), a large public university in the Southern U.S. The sample includes 495

women and 349 men recruited from the economics research lab and pool of economics majors and minors, with a survey nearly identical to the primary study.⁴⁴ Online Appendix Table A13 reports summary statistics comparing Prolific and TAMU samples; TAMU respondents are younger, more religious, and less likely to be white, low-income, or affiliated with the Democratic party.

The social norm results closely mirror the Prolific findings (Online Appendix Figure A5 and Tables A14-A15). Rates of agreement with victim-blaming were low while agreement with lying was relatively high, and these rates are not statistically different from the Prolific sample (with the exception that TAMU men are less likely to agree with victim-blaming). The social norm misperceptions observed in the Prolific sample also persist. While the guesses regarding others' agreement with the social norm statements in the TAMU sample are statistically different from the Prolific sample, they are qualitatively similar (within 5 percentage points in most cases). These findings, while descriptive in nature, provide additional evidence that the Prolific sample is informative of the broader population of interest, specifically regarding their beliefs about the social costs associated with reporting.

The replication targeted a sample size of 900 women and 900 men, though the final sample size obtained is just under 500 women and 350 men.⁴⁵ Thus, statistical power is limited, especially for estimating treatment effects of hard-garbling. However, estimates from Equation 12, and heterogeneity on the basis of sensitivity to judgment following Equation 13, are reported in the online appendix (Tables A17 and A18).⁴⁶ While hard-garbling does not appear to impact reports of sexual assault victimization or perpetration in the whole sample, I again find heterogeneity on the basis of sensitivity to social judgment. For individuals more sensitive to social judgment, hard-garbling increases reports of less severe victimizations by 11.3 percentage points ($p < 0.1$), which is a 54 percent increase over the direct elicitation

⁴⁴Students were recruited via email and paid a \$5 participation incentive. Due to funding constraints, the replication did not include incentives for comprehension or social norm questions. The payment system was set up to be entirely anonymous.

⁴⁵The replication initially recruited students signed up to participate in the university's economics research lab. I later attempted to increase the sample size by recruiting from the pool of all economics majors or minors at the university. Respondents were sent multiple follow-up emails.

⁴⁶A balance table, testing for differences in observable characteristics by treatment status in the TAMU sample, is reported in Online Appendix Table A16. There are no observable characteristics unbalanced by treatment. Respondents were intentionally over-sampled in the HG treatment to aid in statistical power.

mean of 21 percent. While the estimates are less precise, hard-garbling also increases reports of rape victimization by 8 percentage points and rape perpetration by 5.9 percentage points for those more sensitive to judgment. However, given the small sample size and subsequent lack of statistical power, these estimates may reflect noise rather than meaningful impacts.

The consistent social norm patterns and replicated heterogeneity by judgment sensitivity provide suggestive evidence of generalizability, though the limited statistical power of the TAMU replication precludes stronger claims.

6.3 Additional Checks: Controls, Comprehension, Attrition, and Leak Probability

To test whether the primary results are sensitive to the controls used, I conduct the primary analysis with no controls, additional controls, and controls selected using the Double Lasso method (which is the preferred specification). The results are robust across these specifications (Online Appendix Table [A19](#)).

To verify that the results are not driven by individuals who misunderstood the hard-garbling survey technique or by inattentive survey takers, I run robustness checks that exclude individuals who failed comprehension questions or attention checks. Online Appendix Table [A20](#) reports the estimated treatment effects among: the whole sample, while dropping individuals who missed both hard-garbling comprehension questions (i.e., the primary sample used in analysis), while dropping individuals who missed either hard-garbling comprehension question, and while dropping individuals who failed any of the three attention checks. The results are robust across these specifications.

We may be concerned that male participants are more likely to drop out of survey under hard-garbling if they are not perpetrators, which would artificially increase the treatment effects. Descriptively, attrition was not prevalent in the study. Of the total 1,658 responses from individuals who met the recruitment criteria for the study and answered the demographic questions, attrited individuals account for only 1% of responses.⁴⁷ To further verify

⁴⁷My primary sample size is 1,621, plus the 18 individuals who incorrectly answered both hard-garbling comprehension questions. Thus, there are a total of 19 individuals that did not complete the study. The majority of individuals who ultimately left the study did answer the experience questions prior to exiting. Of the 19 attrited responses, 14 individuals answered both sexual assault experience questions, and 1 individual

that the results are not driven by attrition, I run additional robustness checks that include all 1,658 respondents (Online Appendix Table A21). The treatment effects are robust.

As discussed in Section 2.2, the mathematical framework for hard-garbling requires the existence of a perceived non-zero leak probability $p > 0$. To provide a measure of this perceived leak probability, the survey contained questions regarding the perceptions of anonymity. On average, women (men) believe there is a 21 (25) percent chance that Prolific shares their identifying information with researchers and a 27 (29) percent chance that researchers can link their answers to their identities. These perceptions did not differ by hard-garbling treatment status. As an additional robustness check, I drop all respondents who indicate a perceived leak probability $p = 0$, i.e., those who answered that there is a 0 percent chance Prolific shares identifying information and a 0 percent chance that researchers could link their responses to their identity, and re-estimate Equation 12. The results of this exercise are reported in Online Appendix Table A22; the treatment effects are robust.⁴⁸

7 Discussion and Conclusion

Sexual assault experiences are under-reported, even in anonymous surveys designed to encourage honest disclosure. In this study, I use a between-subjects online experiment with 1,621 respondents to examine social judgment as a barrier to reporting. The experiment (1) quantifies (mis-)perceptions of social norms related to sexual assault, and (2) tests whether a survey technique designed to provide plausible deniability increases self-reports of victimization and perpetration.

I find that 50 percent of men and 28 percent of women agree that women who say they were sexually assaulted may be lying, though only 16 percent of men and 8 percent of women believe that women who were sexually assaulted may be to blame. However, despite low rates of agreement with victim-blaming, the norm is widely misperceived, as both men and women

answered the rape question but did not answer the other assault question.

⁴⁸In line with my pre-specified analysis plan, I also test for heterogeneous effects on the basis of p using the specification in Equation 13 (Appendix Tables A23 and A24). The results are consistent with hard-garbling being more effective among respondents who perceive higher leak probabilities for perpetration outcomes, but the pattern is not consistent for victimization. This may reflect limited power or that respondents are concerned with judgment from the researcher, as discussed in Section 7.

drastically overestimate the percent of others that agree with the victim-blaming statement. This may suggest that victims face high expected social costs when reporting.

I also find that hard-garbling increases reports of non-rape sexual assault victimization by 10.5 percentage points (23 percent) and increases reports of rape perpetration by 6.9 percentage points (111 percent), though it does not increase reports of rape victimization or non-rape perpetration. Overall, hard-garbling is most effective when the expected social costs of reporting are likely highest: for less severe victimizations, for more severe perpetrations, for individuals more sensitive to social judgment, and conditional on this sensitivity, for victims who expect more agreement with the social norm statements.

While we might expect hard-garbling to be most effective among women who expect the most agreement with the social norm statements, this pattern does not emerge in the full sample. A plausible explanation is that perceived social norms are broadly victim-unfriendly. On average, women expect about one-quarter of women and half of men to endorse victim-blaming and lying statements. As a result, hard-garbling may increase reporting regardless of perceived norms, as nearly all women anticipate substantial social costs.

The pattern of treatment effects suggests that respondents anticipate judgment-related utility costs from disclosure, even in an anonymous online setting. Several alternative mechanisms can be ruled out. A pure privacy-preference explanation, in which respondents simply dislike disclosing personal information regardless of judgment, is inconsistent with the results. First, this explanation would predict that hard-garbling has the largest effects for the most private information. However, I find a null effect on rape victimization, arguably the most private experience in the survey. Second, I find that the treatment effects of hard-garbling are most pronounced among individuals more sensitive to social judgment, which provides suggestive evidence against the privacy-preference explanation as a primary driver of the treatment effects. Other possible mechanisms, like emotional costs of recalling the experience or of labeling oneself as a victim or a perpetrator are mechanically unaffected by hard-garbling. These costs are determined by the intended answer, and hard-garbling, by construction, can only attenuate utility costs tied to the recorded output.

The remaining mechanisms, anticipated judgment from others or from the researcher, are variants of the same underlying channel: respondents' utility depends on how the act of

disclosure affects how others view their type. The experimental design cannot distinguish among these audiences, but the policy-relevant takeaway is robust to which is operative. Anonymous surveys systematically under-measure stigmatized experiences whenever disclosure activates judgment-related utility costs, and plausible deniability can attenuate these costs regardless of whose judgment respondents anticipate.

Further, estimates of sexual assault prevalence from large-scale surveys may be unreliable. This study used similar methods, including similar experience questions, an anonymous online format, and a trusted platform, yet still found under-reporting. My estimates suggest that standard anonymous survey techniques understate non-rape victimization by at least 19 percent and rape perpetration by at least 53 percent.⁴⁹ Because hard-garbling only attenuates social-cost-driven under-reporting and not other channels, these are likely lower bounds on total under-reporting. In the future, institutions can use hard-garbling to improve reporting, yielding more accurate data and ultimately better targeted prevention and support efforts. Hard-garbled perpetration reports could also serve as outcomes in evaluations of interventions designed to reduce sexual assault, where reliable outcome data are typically unavailable. Plausible-deniability designs could also improve measurement in a range of settings where anticipated social judgment biases honest reporting, including workplace climate surveys of harassment and discrimination, surveys of drug use and intimate partner violence, and program evaluations that rely on self-reports of stigmatized outcomes.

By directly measuring both sexual assault victimization and perpetration and identifying expectations of social judgment as a barrier to disclosure, this study advances our understanding of the decision (not) to report. Particularly in formal reporting contexts, other barriers to reporting operate alongside social judgment, including fears of retaliation, emotional and procedural burdens, and legal consequences. Understanding how these barriers interact with social judgment, and how they shape formal reporting to institutions and authorities, is an important direction for future research.

⁴⁹For non-rape victimization, hard-garbling raises the reporting rate from 45.6 percent under direct elicitation to 56.1 ($45.6 + 10.5$). The DE estimate is therefore 81.3 percent of the hard-garbling estimate, implying that DE understates non-rape victimization by at least $(1 - 0.813) = 18.7$ percent relative to hard-garbling.

References

- Anderson, R. E., Silver, K. E., Ciampaglia, A. M., Vitale, A. M., & Delahanty, D. L. (2021). The frequency of sexual perpetration in college men: A systematic review of reported prevalence rates from 2000 to 2017. *Trauma, Violence & Abuse, 22*(3), 481–495. <https://doi.org/10.1177/1524838019860619>
- Bedera, N. (2024). I can protect his future, but she can't be helped: Himpathy and hysteria in administrator rationalizations of institutional betrayal. *The Journal of Higher Education, 95*(1), 30–53. <https://doi.org/10.1080/00221546.2023.2195771>
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81*(2), 608–650. <https://doi.org/10.1093/restud/rdt044>
- Bermek, S., Matakos, K., & Unan, A. (2023). Victim-blaming social norms and violence against women: Correcting misperceptions or morality drive policy and behavior change? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4609063>
- Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association, 110*(511), 1304–1319. <https://doi.org/10.1080/01621459.2015.1050028>
- Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics, 131*(4), 1753–1794. <https://doi.org/10.1093/qje/qjw029>
- Boudreau, L. E., Chassang, S., Gonzalez-Torres, A., & Heath, R. M. (2025, March). *Monitoring harassment in organizations* (Working Paper No. 31011). National Bureau of Economic Research. <https://doi.org/10.3386/w31011>
- Bursztyn, L., Cappelen, A. W., Tungodden, B., Voena, A., & Yanagizawa-Drott, D. H. (2023, March). *How are gender norms perceived?* (Working Paper No. 31049). National Bureau of Economic Research. <https://doi.org/10.3386/w31049>
- Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review, 110*(10), 2997–3029. <https://doi.org/10.1257/aer.20180975>

- Bursztyn, L., & Yang, D. Y. (2022). Misperceptions about others. *Annual Review of Economics*, 14(1), 425–452. <https://doi.org/10.1146/annurev-economics-051520-023322>
- Cameron, M., Johnson, R., Lacy, T., Wu, J., Siegel, P., Holley, J., & Wine, J. (2023). *2019-20 National Postsecondary Student Aid Study: First Look at Student Financial Aid Estimates for 2019-2020* (tech. rep. No. NCES 2023-466). U.S. Dept of Education.
- Cantor, D., Fisher, B., Chibnall, S., Townsend, R., Lee, H., Bruce, C., Thomas, G., Harps, S., Kranz, V., Herbison, R., & Madden, K. (2020). Report on the AAU Campus Climate Survey on Sexual Assault and Sexual Misconduct.
- Chassang, S., & Miquel, G. P. I. (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies*, 86(6), 2530–2553. <https://doi.org/10.1093/restud/rdy075>
- Chassang, S., & Zehnder, C. (2024). Secure survey design in organizations: Theory and experiments. *American Economic Journal: Microeconomics*, 16(4), 371–405. <https://doi.org/10.1257/mic.20220187>
- Cheng, I.-H., & Hsiaw, A. (2022). Reporting sexual misconduct in the #MeToo era. *American Economic Journal: Microeconomics*, 14(4), 761–803. <https://doi.org/10.1257/mic.20200218>
- Chuang, E., Dupas, P., Huillery, E., & Seban, J. (2021). Sex, lies, and measurement: Consistency tests for indirect response survey methods. *Journal of Development Economics*, 148, 102582. <https://doi.org/10.1016/j.jdeveco.2020.102582>
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349–354. <https://doi.org/10.1037/h0047358>
- Dahl, G. B., & Knepper, M. (2026). Why is workplace sexual harassment underreported? The value of outside options amid the threat of retaliation. *American Economic Review*, 116(3), 897–933. <https://doi.org/10.1257/aer.20221703>
- Denti, D., & Iammarino, S. (2022). Coming out of the woods. Do local support services influence the propensity to report sexual violence? *Journal of Economic Behavior & Organization*, 193, 334–352. <https://doi.org/10.1016/j.jebo.2021.11.024>

- de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11), 3266–3302. <https://doi.org/10.1257/aer.20171330>
- Dhar, D., Jain, T., & Jayachandran, S. (2022). Reshaping adolescents’ gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, 112(3), 899–927. <https://doi.org/10.1257/aer.20201112>
- FBI. (2026). Crime Data Explorer. Available at: <https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/crime-trend>. Accessed: 2026-05-12.
- Fedina, L., Holmes, J. L., & Backes, B. L. (2018). Campus sexual assault: A systematic review of prevalence research from 2000 to 2015. *Trauma, Violence, & Abuse*, 19(1), 76–93. <https://doi.org/10.1177/1524838016631129>
- Fields, J. M., & Schuman, H. (1976). Public beliefs about the beliefs of the public. *Public Opinion Quarterly*, 40(4), 427. <https://doi.org/10.1086/268330>
- Fisher, B. S., Daigle, L. E., Cullen, F. T., & Turner, M. G. (2003). Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Criminal Justice and Behavior*, 30(1), 6–38. <https://doi.org/10.1177/0093854802239161>
- Gauthier, G. (2024). Measuring crime reporting and incidence: Method and application to #MeToo. <https://doi.org/10.2139/ssrn.4242506>
- Hoover, H., & Krupka, E. (2025). Different norms of sexual activity and consent seeking among college students: Social identity and statistical discrimination. *Journal of Economic Behavior & Organization*, 235, 107028. <https://doi.org/10.1016/j.jebo.2025.107028>
- Johnson, S. M., Murphy, M. J., & Gidycz, C. A. (2017). Reliability and validity of the sexual experiences survey–short forms victimization and perpetration. *Violence and Victims*, 32(1), 78–92. <https://doi.org/10.1891/0886-6708.VV-D-15-00110>
- Lee, F. X., & Suen, W. (2020). Credibility of crime allegations. *American Economic Journal: Microeconomics*, 12(1), 220–259. <https://doi.org/10.1257/mic.20180231>
- Levy, R., & Mattsson, M. (2024). The effects of social movements: Evidence from #MeToo. <https://doi.org/10.2139/ssrn.3496903>

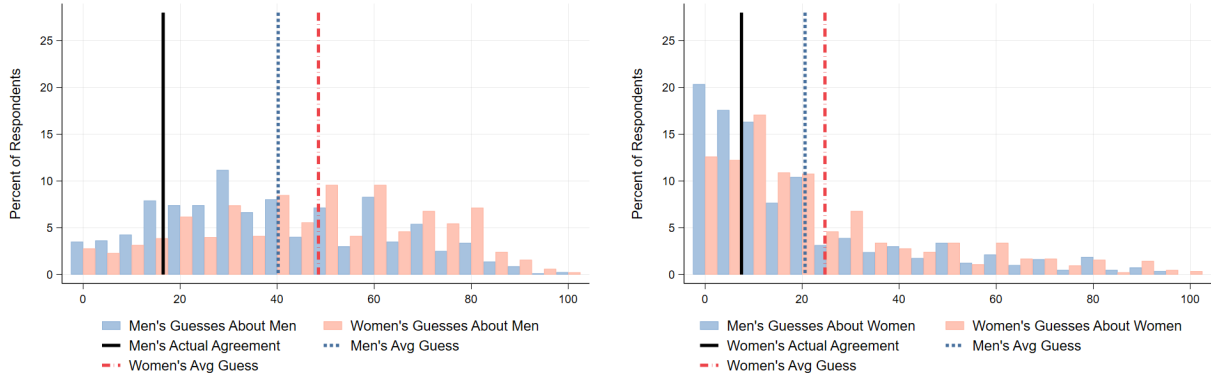
- Lindo, J. M., Siminski, P., & Swensen, I. D. (2018). College party culture and sexual assault. *American Economic Journal: Applied Economics*, 10(1), 236–265. <https://doi.org/10.1257/app.20160031>
- List, J. A. (2020, July). *Non est disputandum de generalizability? a glimpse into the external validity trial* (Working Paper No. 27535). National Bureau of Economic Research. <https://doi.org/10.3386/w27535>
- Lonsway, K. A., & Fitzgerald, L. F. (1994). Rape myths: In review. *Psychology of Women Quarterly*, 18(2), 133–164. <https://doi.org/10.1111/j.1471-6402.1994.tb00448.x>
- Miller, A., & Segal, C. (2019). Do female officers improve law enforcement quality? Effects on crime reporting and domestic violence. *The Review of Economic Studies*, 86(5), 2220–2247. <https://doi.org/10.1093/restud/rdy051>
- Miller, D. T., & Prentice, D. A. (1994). Collective errors and errors about the collective. *Personality and Social Psychology Bulletin*, 20(5), 541–550. <https://doi.org/10.1177/0146167294205011>
- Morgan, R., & Thompson, A. (2021, October). *Criminal Victimization, 2020* (tech. rep. No. NCJ 301775). U.S. Department of Justice. Bureau of Justice Statistics.
- Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2), 517–529. <https://doi.org/10.1017/S0003055418000837>
- National Center for Education Statistics. (2023). *Undergraduate Enrollment*. Retrieved November 24, 2025, from <https://nces.ed.gov/programs/coe/indicator/cha/undergrad-enrollment>
- Nix, E., Adams, A., Huttunen, K., & Zhang, N. (2026). *The Economic Costs of Rape* (Working paper).
- Postsecondary National Policy Institute. (2025, April). *First-Generation Students in Higher Education*. Retrieved November 24, 2025, from <https://pnpi.org/factsheets/first-generation-students/>
- Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282. <https://doi.org/https://doi.org/10.1111/j.1468-0262.2005.00615.x>

- Seabright, P., & Marti, J. H. (2022). *Alcohol, behavioral norms and sexual violence on U.S. college campuses* (CEPR Discussion Paper No. DP17147). Centre for Economic Policy Research. Rochester, NY.
- Spencer, C., Stith, S., Durtschi, J., & Toews, M. (2020). Factors related to college students' decisions to report sexual assault. *Journal of Interpersonal Violence*, 35(21-22), 4666–4685. <https://doi.org/10.1177/0886260517717490>
- Topper, M. (2026). The effect of fraternity moratoriums on alcohol offenses and sexual assaults. *Journal of Human Resources*. <https://doi.org/10.3368/jhr.0722-12422R1>
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63–69. <https://doi.org/10.1080/01621459.1965.10480775>

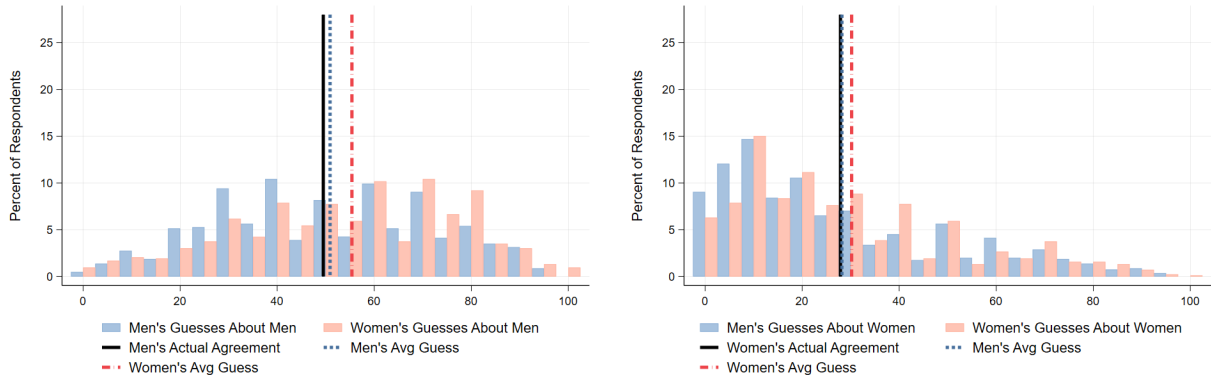
Figures

Figure 1: Perceptions of Agreement with Victim-Blaming and Lying Statements, by Gender

(a) Victim-Blaming: Perception of Men's Agreement (b) Victim-Blaming: Perception of Women's Agreement



(c) Lying: Perception of Men's Agreement (d) Lying: Perception of Women's Agreement



Notes: Panels (a)–(b) show the raw distribution of respondents' perceptions of agreement with the victim-blaming statement from all 825 female and 796 male respondents. Panels (c)–(d) show the raw distribution of respondents' perceptions of agreement with the lying statement for the same sample. Respondents were asked separately about the predicted percent of men and women in a prior study that they believe would have agreed with each statement. Panels (a) and (c) display respondents' perceptions of men's agreement; Panels (b) and (d) display respondents' perceptions of women's agreement. Men's guesses are displayed in blue; women's guesses are displayed in red. The black line shows actual agreement with the statement by men (in Panels (a) and (c)) and women (in Panels (b) and (d)). The blue line shows the average answer of male respondents' perceptions of men and women's agreement; the red line shows the average answer of female respondents' perceptions of men and women's agreement. For victim-blaming, Kolmogorov–Smirnov tests indicate 1) that the distribution of guesses by men about men is statistically different from the distribution of guesses by women about men (Panel (a); $p = 0.00$), 2) that the distribution of guesses by men about women is statistically different from the distribution of guesses by women about women (Panel (b); $p = 0.00$). For lying, Kolmogorov–Smirnov tests indicate 1) that the distribution of guesses by men about men is statistically different from the distribution of guesses by women about men (Panel (c); $p = 0.00$), 2) that the distribution of guesses by men about women is statistically different from the distribution of guesses by women about women (Panel (d); $p = 0.01$).

Tables

Table 1: Balance Table

	Panel A. Women			Panel B. Men			M=F (P-value)
	HG	DE	P-value	HG	DE	P-value	
Age	22.93 (2.50)	23.04 (2.53)	0.52	23.44 (2.68)	23.39 (2.69)	0.77	0.00
White	0.57 (0.50)	0.58 (0.49)	0.77	0.53 (0.50)	0.56 (0.50)	0.41	0.22
Current Undergrad	0.49 (0.50)	0.43 (0.50)	0.11	0.46 (0.50)	0.45 (0.50)	0.64	0.89
Greek	0.19 (0.39)	0.23 (0.42)	0.17	0.20 (0.40)	0.15 (0.36)	0.07	0.16
College Athlete	0.22 (0.41)	0.25 (0.43)	0.25	0.25 (0.43)	0.25 (0.44)	0.81	0.49
Public Institution	0.73 (0.45)	0.78 (0.42)	0.08	0.79 (0.41)	0.74 (0.44)	0.08	0.64
Large Institution	0.58 (0.49)	0.56 (0.50)	0.68	0.61 (0.49)	0.54 (0.50)	0.04	0.86
Low-Income	0.45 (0.50)	0.54 (0.50)	0.01	0.49 (0.50)	0.47 (0.50)	0.49	0.66
First-Gen	0.29 (0.45)	0.29 (0.45)	0.97	0.26 (0.44)	0.27 (0.45)	0.76	0.35
Non-Religious	0.31 (0.46)	0.33 (0.47)	0.51	0.31 (0.46)	0.33 (0.47)	0.57	0.86
Southern	0.44 (0.50)	0.41 (0.49)	0.39	0.41 (0.49)	0.31 (0.46)	0.01	0.01
Democrat	0.49 (0.50)	0.45 (0.50)	0.34	0.37 (0.48)	0.38 (0.49)	0.71	0.00
Observations	417	408		394	402		

Notes: Numbers are averages across female and male respondents by treatment, respectively. Standard deviations are in parentheses. Individuals who incorrectly answered both comprehension questions in the hard-garbling treatment have been excluded. Numbers in p-value columns are generated from tests of equality of means across Hard-Garbling (HG) and Direct Elicitation (DE) treatments. The numbers in the last column are generated from tests of equality of means across all male and all female respondents. All variables, excluding age, are 0/1 indicator variables. White takes a value of 1 if a respondent indicated they were white, and is 0 otherwise. Current Undergrad takes a value of 1 if a respondent indicated being a current undergraduate student. Greek and Athlete take a value of 1 if respondents indicated involvement in Greek life or collegiate athletics, respectively. Low-income takes a value of 1 if the respondent reported having ever been eligible for Pell Grants, and is 0 otherwise. Non-religious takes a value of 1 if the respondent indicated being non-religious (like Atheist or agnostic) and is zero otherwise. Public Institution takes a value of 1 if a respondent indicated that their higher education institution was public, and is 0 otherwise. Large Institution takes a value of 1 if a respondent indicated that their institution was large, and is 0 if else. Democrat takes a value of 1 if a respondent reported voting for Kamala Harris in the 2024 presidential election. Southern is an indicator for whether a respondent reported attending an undergraduate institution in the South, according to U.S. Census Bureau region designations.

Table 2: Hard-Garbling Treatment Effects

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling	-0.010 (0.037) [0.777]	0.105*** (0.038) [0.011]	0.069** (0.031) [0.055]	0.050 (0.031) [0.109]
Observations	825	825	796	796
Direct Elicitation Mean	0.343	0.456	0.062	0.090
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Each column is a separate regression. Columns (1) and (2) restrict to responses on victimization only, i.e., those from female respondents; columns (3) and (4) restrict to responses on perpetration only, i.e., those from male respondents. The experiment contained two binary questions on sexual assault experiences. The first question pertained to unwanted sexual penetration or oral sex (rape), and the second regarded all other types of unwanted sexual contact. The outcome variable in Columns (1) and (3) is reports of unwanted sexual penetration or oral sex, i.e., rape. The outcome variable in Columns (2) and (4) is reports of any other type of unwanted sexual contact. Outcome variables were normalized to account for garbled responses following Boudreau et al. (2025) and Blair et al. (2015). Hard-Garbling is an indicator equal to 1 if an individual answered the sexual assault experience questions using the hard-garbling survey technique, and is 0 if they answered using direct elicitation. Controls in all columns are selected using the Double Lasso CV method. Multiple hypothesis corrected p-values using the Romano and Wolf (2005) procedure are reported in square brackets, adjusted for victimization and perpetration outcomes separately, i.e., adjusted for Columns (1) and (2) together, and adjusted for Columns (3) and (4) together. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3: Hard-Garbling Treatment Effects – Sensitivity to Social Judgment

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling (HG)	-0.011 (0.053) [0.825]	0.165*** (0.053) [0.002]	0.096** (0.043) [0.056]	0.094** (0.044) [0.056]
Low Sensitivity to Judgment (Low SJ)	0.030 (0.045)	0.068 (0.047)	0.033 (0.025)	0.027 (0.029)
HG x Low SJ	0.002 (0.075)	-0.121 (0.075)	-0.054 (0.061)	-0.088 (0.063)
DE + High SJ Mean	0.33	0.42	0.05	0.08
HG + HGxLow SJ = 0 (P-value)	0.863	0.410	0.334	0.893
Observations	825	825	796	796
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Each column is a separate regression. Outcome variables were normalized to account for garbled responses following Boudreau et al. (2025) and Blair et al. (2015). Columns (1) and (2) restrict to responses on victimization only, i.e., those from female respondents; columns (3) and (4) restrict to responses on perpetration only, i.e., those from male respondents. The experiment contained two binary questions on sexual assault experiences. The first question pertained to unwanted sexual penetration or oral sex (rape), and the second regarded all other types of unwanted sexual contact. The outcome variable in Columns (1) and (3) is reports of unwanted sexual penetration or oral sex, i.e., rape. The outcome variable in Columns (2) and (4) is reports of any other type of unwanted sexual contact. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. Controls in all columns were selected using the Double Lasso CV method. “Low Sensitivity to Judgment” is an indicator equal to 1 if the respondent’s social desirability score was below the mean within their gender, and is zero if else. “HG x Low SJ” is an interaction term. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets, adjusted for pairs of hypotheses within victimization and perpetration outcomes separately, for victimization responses (Columns (1) and (2)) and for perpetration responses (Columns (3) and (4))

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4: Hard-Garbling Treatment Effects – Expectations of and Sensitivity to Social Judgment

	Victimization	
	Rape (1)	Other SA (2)
<i>Panel A: High Expectations of Judgment</i>		
Hard-Garbling (HG)	-0.007 (0.079)	0.193** (0.075)
Low Sensitivity to Judgment (Low SJ)	0.036 (0.065)	0.089 (0.065)
HG x Low SJ	-0.046 (0.108)	-0.247** (0.106)
DE + High SJ Mean	0.35	0.42
HG + HGxLow SJ = 0 (P-value)	0.470	0.480
Observations	413	413
<i>Panel B: Low Expectations of Judgment</i>		
Hard-Garbling (HG)	0.001 (0.075)	0.132* (0.074)
Low Sensitivity to Judgment (Low SJ)	0.050 (0.064)	0.050 (0.069)
HG x Low SJ	0.035 (0.106)	0.014 (0.107)
DE + High SJ Mean	0.32	0.42
HG + HGxLow SJ = 0 (P-value)	0.632	0.055*
Observations	412	412
Lasso Controls	Yes	Yes

Notes: OLS Regressions. Each column and panel is a separate regression. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses following Boudreau et al. (2025) and Blair et al. (2015). The table is restricted to victimization responses only. The experiment contained two binary questions on sexual assault experiences. The first question pertained to unwanted sexual penetration or oral sex (rape), and the second regarded all other types of unwanted sexual contact. The outcome variable in Column (1) is reports of unwanted sexual penetration or oral sex, i.e., rape. The outcome variable in Column (2) is reports of any other type of unwanted sexual contact. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. “Low Sensitivity to Judgment” is an indicator equal to 1 if the respondent’s social desirability score was below the mean within their gender, and is zero if else. “HG x Low SJ” is an interaction term. The experiment contained 4 social norm prediction questions. I aggregate and standardize these predictions around the control mean, and then create an indicator. Panel A restricts to only respondents who reported high expectations of social judgment, i.e., those with social norm guesses above or equal to the median within their gender. Panel B restricts to only respondents who report low expectations of social judgment, i.e., those with social norm guesses below the median within their gender. Controls in all columns and panels were selected via Double Lasso CV Method. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

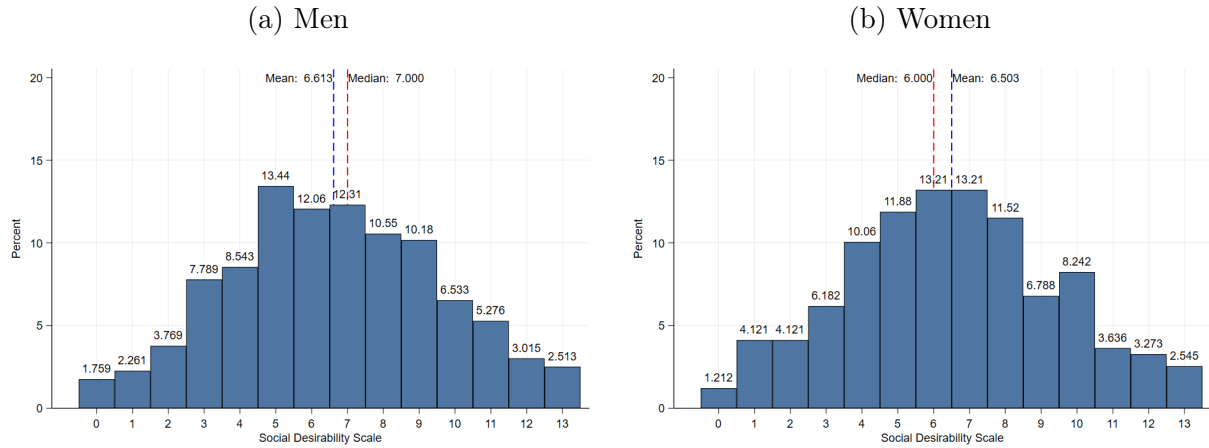
Table 5: Hard-Garbling Treatment Effects by Victimization Outcome

	Sexual Assault Outcomes		Theft Outcomes	
	Rape (1)	Other SA (2)	Over \$1,000 (3)	Under \$1,000 (4)
Hard-Garbling	-0.010 (0.037) [0.777]	0.105*** (0.038) [0.011]	0.045 (0.055) [0.397]	0.070 (0.057) [0.397]
Observations	825	825	306	306
Direct Elicitation Mean	0.343	0.456	0.359	0.438
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses following Boudreau et al. (2025) and Blair et al. (2015). The outcome variables are: reports of rape victimization (Column 1); reports of other unwanted sexual contact victimization (Column 2); reports of theft victimization worth over \$1,000 (Column 3); and reports of theft victimization worth under \$1,000 (Column 4). Hard-Garbling is an indicator equal to 1 if an individual answered the sexual assault or theft questions using the hard-garbling survey technique, and is 0 if direct elicitation. Controls in all columns were selected using the Double Lasso CV method. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets, adjusted for pairs of hypotheses within sexual assault and theft outcomes separately, for those in the whole sample (i.e., adjusting for cols (1) and (2), and for cols (3) and (4)). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

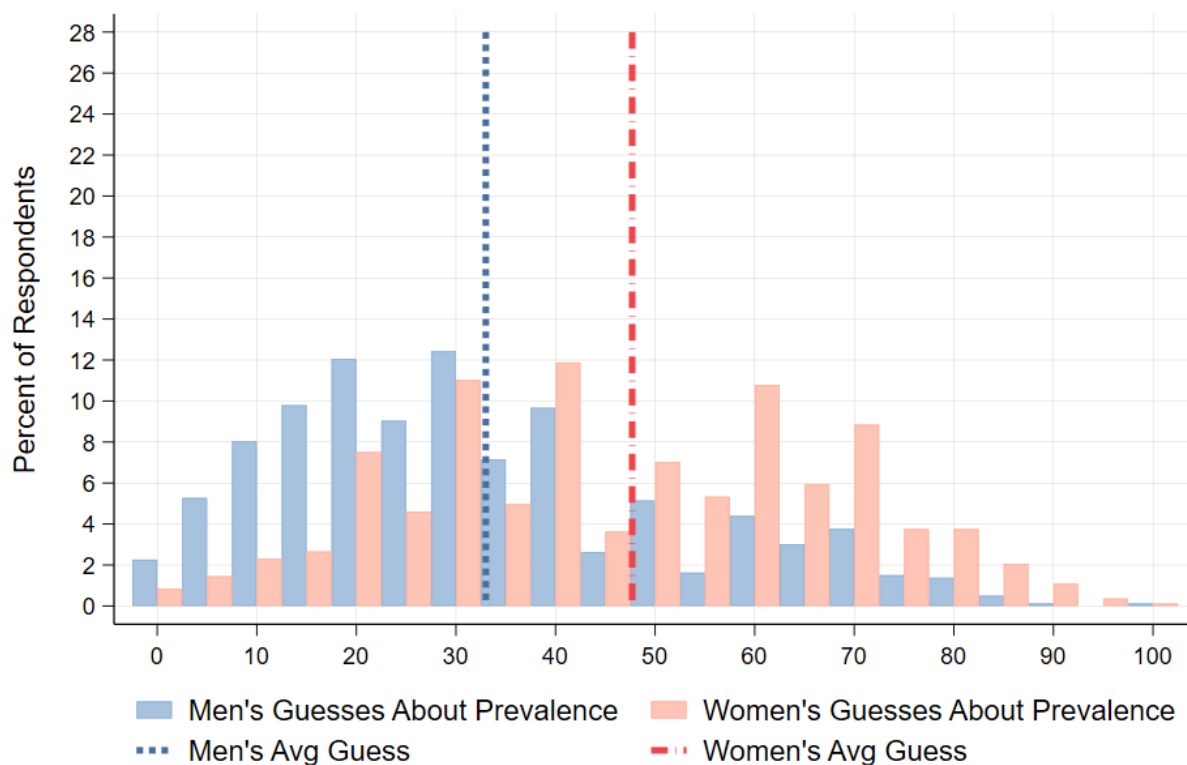
A Appendix Tables and Figures

Figure A1: Distribution of Sensitivity to Judgment Scores



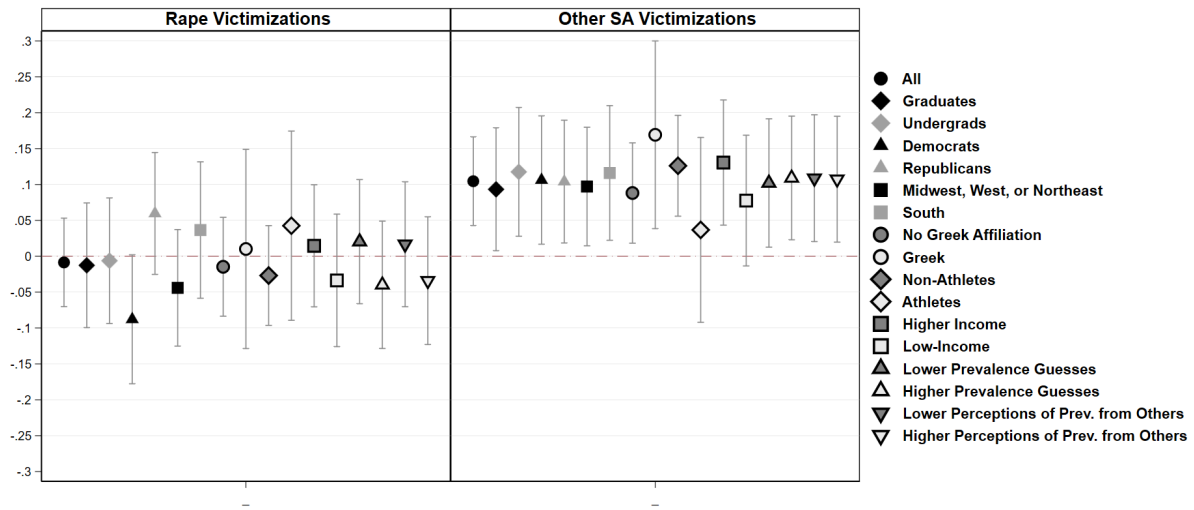
Notes: Following Dhar et al. (2022), I count the number of “too-good-to-be-true” traits each individual reports. This figure plots the raw distribution of these scores to the social desirability scale for all 825 female and all 796 male respondents. The mean is for each gender is plotted on the blue dotted line; the red dotted line represents the median within the gender.

Figure A2: Perceptions of College Sexual Assault Prevalence



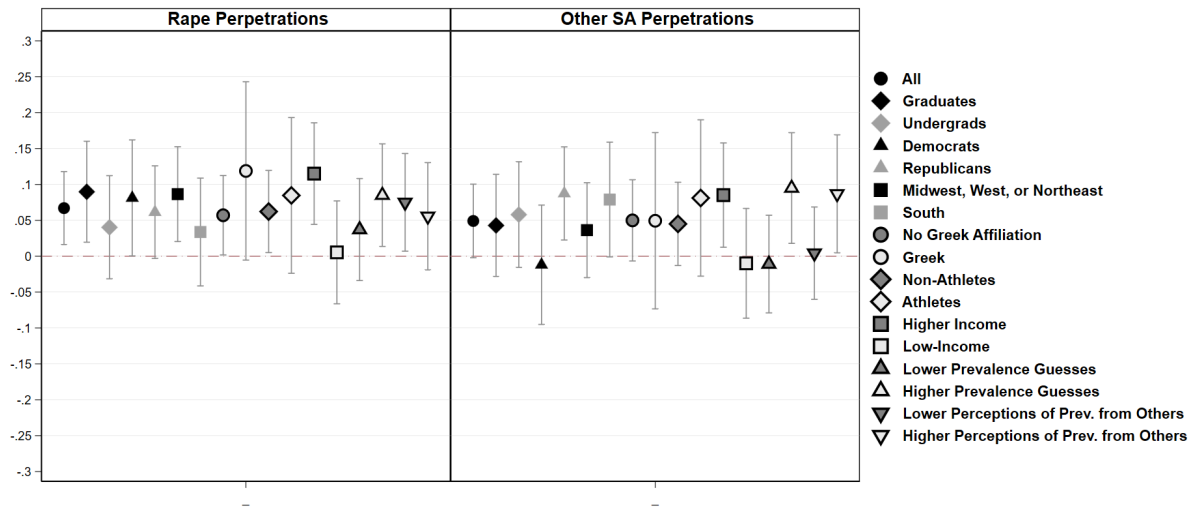
Notes: Figure shows the raw distribution of respondents' guesses regarding sexual assault prevalence from all 825 female and 796 male respondents. Men's guesses are displayed in blue; women's guesses are displayed in red. The blue line shows the average answer of male respondents' prevalence guesses. The red line shows the average answer of female respondents' prevalence guesses.

Figure A3: Hard-Garbling Heterogeneity: Victimization



Notes: Coefficients are estimated using OLS regressions, and outcomes were normalized to account for garbled responses. Regressions were restricted to victimization responses only. Each pair of characteristics are binary indicators and mutually exclusive (i.e., a respondent was either a current undergraduate student or a college graduate, but not both). Thus, each regression is run using a treatment indicator, an indicator for the characteristic of interest, and an interaction term. Coefficients plotted represent the total impact of hard-garbling for individuals with the characteristic of interest. Controls in all regressions were selected using the Double Lasso CV method. Individuals with “higher prevalence guesses” are those who had sexual assault prevalence guesses above the median within their gender. The experiment also had subjects guess the answers of prior male and female participants regarding prevalence. I created an indicator “Higher Perceptions of Prev. from Others” for these two social norm guesses using the Anderson method, and is equal to 1 if the index was above the median with the respondent’s gender, and is zero if else. Intervals shown represent the 90% confidence intervals.

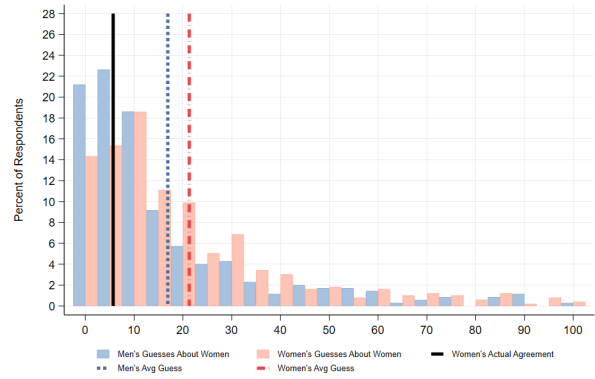
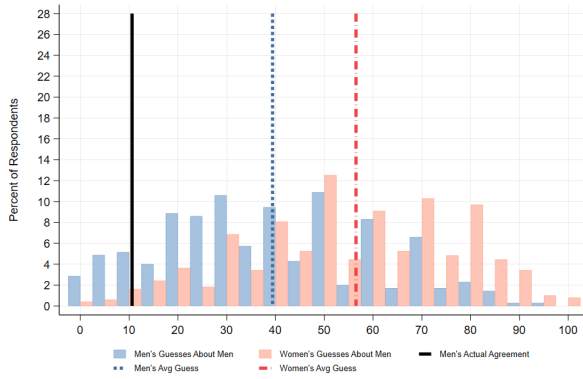
Figure A4: Hard-Garbling Heterogeneity: Perpetration



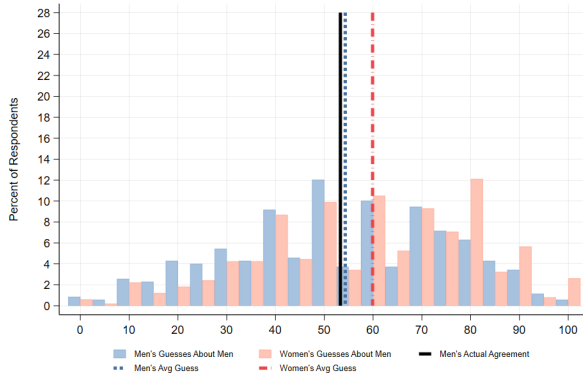
Notes: Coefficients are estimated using OLS regressions, and outcomes were normalized to account for garbled responses. Regressions were restricted to perpetration responses only. Each pair of characteristics are binary indicators and mutually exclusive (i.e., a respondent was either a current undergraduate student or a college graduate, but not both). Thus, each regression is run using a treatment indicator, an indicator for the characteristic of interest, and an interaction term. Coefficients plotted represent the total impact of hard-garbling for individuals with the characteristic of interest. Controls in all regressions were selected using the Double Lasso CV method. Individuals with “higher prevalence guesses” are those who had sexual assault prevalence guesses above the median within their gender. The experiment also had subjects guess the answers of prior male and female participants regarding prevalence. I created an indicator “Higher Perceptions of Prev. from Others” for these two social norm guesses using the Anderson method, and is equal to 1 if the index was above the median with the respondent’s gender, and is zero if else. Intervals shown represent the 90% confidence intervals.

Figure A5: Perceptions of Social Norm Agreement in TAMU Sample

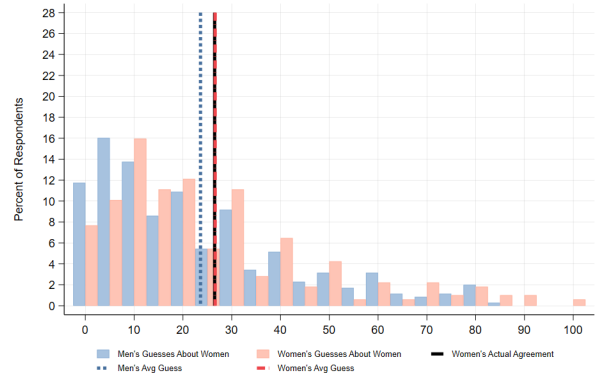
(a) Victim-Blaming: Perception of Men's Agreement (b) Victim-Blaming: Perception of Women's Agreement



(c) Lying: Perception of Men's Agreement



(d) Lying: Perception of Women's Agreement



Notes: Each panel shows the raw distribution of respondents' perceptions of agreement with either the victim-blaming or lying statements among 495 female and 349 male respondents in the TAMU sample. Respondents were asked separately about the predicted percent of men and women in a prior study that they believe would have agreed with the statement. Men's guesses are displayed in blue; women's guesses are displayed in red. The black line shows actual agreement with the statement by men (in left panels) and women (in right panels). The blue line shows the average answer of male respondents' perceptions of men and women's agreement; the red line shows the average answer of female respondents' perceptions.

Table A1: Existing Sexual Assault Survey Studies

Survey	Year	Modality	# of Participating Institutions	Sample Size
Higher Education Sexual Misconduct and Awareness Survey (HESMA)	2024	Online	10	42,133
Association of American Universities (AAU) Campus Climate Survey on Sexual Assault and Misconduct	2019	Online	33	181,752
AAU Climate Survey on Sexual Assault and Sexual Misconduct	2015	Online	27	150,072
The Campus Sexual Assault (CSA) Study	2007	Online	2	6,821
Historically Black College and University Campus Sexual Assault (HBCU-CSA) Study	2008	Online	4	3,951
Administrator-Researcher Campus Climate Collaborative (ARC3) Survey <i>Ongoing - Pretested in 2015 at 4 Institutions; Open-Source Instrument</i>	2015	Online	4	2,200
Campus Climate Survey Validation Study (CCSVS)	2015	Online	9	23,023

Notes: This table summarizes major campus sexual assault climate surveys conducted at U.S. institutions of higher education. Sample sizes reflect total respondents across all participating institutions. All listed surveys employ direct elicitation of sexual assault and misconduct experiences. The ARC3 survey is an open-source instrument developed by a consortium of researchers and student affairs administrators; it was pretested at four institutions in 2015 and has since been deployed at numerous additional schools, with sample sizes varying by deployment. Survey details are compiled from publicly available reports and documentation.

Table A2: Data Quality

	Victimization		Perpetration	
	Number of Respondents	% of Total	Number of Respondents	% of Total
<i>Panel A: Attention Checks</i>				
Passed All Attention Checks	766	92.00%	728	90.30 %
Failed One Attention Check	67	8.00%	77	9.60%
Failed Two Attention Checks	0	0.00%	1	0.10%
Failed All Attention Checks	0	0.00%	0	0.00%
Total	833	100.00%	806	100.00%
<i>Panel B: Comprehension Questions</i>				
Both Correct	386	90.82%	371	91.83%
One Correct	31	7.29%	23	5.69%
Both Incorrect	8	1.88%	10	2.48%
Total	425	99.99%	404	100.00%

Notes: The survey experiment contained 3 attention checks. Panel A shows the number and percent of respondents by gender in sexual assault treatments that: passed all 3 attention checks, failed one of the three checks, failed two of the three checks, and failed all three checks. Individuals in hard-garbling treatments answered 2 comprehension questions. Panel B displays the number and percent of respondents, by gender, in sexual assault treatments who answered: both comprehension questions correctly, answered only one of the comprehension questions correctly, and who answered both comprehension questions incorrectly, respectively.

Table A3: Descriptive Statistics

	Age	White	Current Undergrad	Greek	College Athlete	Public Institution	Large Institution	Low-Income	First-Gen	Non-Religious	Southern	Democrat	N
All Respondents	23.19 (2.60)	0.56 (0.50)	0.46 (0.50)	0.19 (0.39)	0.24 (0.43)	0.76 (0.43)	0.57 (0.49)	0.49 (0.50)	0.28 (0.45)	0.32 (0.47)	0.39 (0.49)	0.42 (0.49)	1621
Women	22.98 (2.51)	0.58 (0.49)	0.46 (0.50)	0.21 (0.40)	0.24 (0.42)	0.75 (0.43)	0.57 (0.50)	0.49 (0.50)	0.29 (0.45)	0.32 (0.47)	0.43 (0.49)	0.47 (0.50)	825
Men	23.42 (2.68)	0.55 (0.50)	0.46 (0.50)	0.18 (0.38)	0.25 (0.43)	0.76 (0.43)	0.58 (0.49)	0.48 (0.50)	0.27 (0.44)	0.32 (0.47)	0.36 (0.48)	0.37 (0.48)	796

Notes: Numbers are averages across all respondents, female respondents, and male respondents, respectively. Standard deviations are in parentheses. Individuals who incorrectly answered both comprehension questions in the hard-garbling treatment have been excluded. All variables, excluding age, are 0/1 indicator variables. White takes a value of 1 if a respondent indicated they were white, and is 0 otherwise. Current Undergrad takes a value of 1 if a respondent indicated being a current undergraduate student. Greek and Athlete take a value of 1 if respondents indicated involvement in Greek life or collegiate athletics, respectively. Low-income takes a value of 1 if the respondent reported having ever been eligible for Pell Grants, and is 0 otherwise. Non-religious takes a value of 1 if the respondent indicated being non-religious (like Atheist or agnostic) and is zero otherwise. Public Institution takes a value of 1 if a respondent indicated that their higher education institution was public, and is 0 otherwise. Large Institution takes a value of 1 if a respondent indicated that their institution was large, and is 0 if else. Democrat takes a value of 1 if a respondent reported voting for Kamala Harris in the 2024 presidential election. Southern is an indicator for whether a respondent reported attending an undergraduate institution in the South, according to U.S. Census Bureau region designations.

Table A4: Perceptions of Social Norms

	Victim-Blaming			Lying		
	Actual Agreement	Perception of Men	Perception of Women	Actual Agreement	Perception of Men	Perception of Women
Men	16.46 (37.10)	40.16 (22.65)	20.55 (21.85)	49.50 (50.03)	50.92 (22.27)	28.18 (23.09)
Women	7.52 (26.38)	48.47 (23.60)	24.71 (22.82)	27.88 (44.87)	55.42 (22.70)	30.21 (22.47)
P-value	0.00	0.00	0.00	0.00	0.00	0.07

Notes: Table shows the raw averages of answers from 796 men and 825 women. Agreement with each statement was asked on a 5-point Likert scale and was binarized to be equal to 1 if a respondent agreed or strongly agreed. Respondents were asked to separately predict the percent of men and women from a prior study that agreed with each statement; respondents were randomly assigned to answer regarding men or women first. P-values disclosed in the last row are generated from tests of equalities of means between male and female respondents.

Table A5: Social Norm Predictions

	Actual Agreement		Guesses About Women		Guesses About Men	
	VB (1)	Lying (2)	VB (3)	Lying (4)	VB (5)	Lying (6)
Female	-0.089*** (0.016)	-0.216*** (0.024)	4.161*** (1.109)	2.030* (1.132)	8.313*** (1.149)	4.505*** (1.117)
Male Mean	0.16	0.49	20.55	28.18	40.16	50.92
Observations	1,621	1,621	1,621	1,621	1,621	1,621
Controls	No	No	No	No	No	No

Notes: OLS regressions. Each column is a separate regression. Respondents were asked about their own agreement with each statement and were separately about the predicted percent of men and women in a prior study that they believe would have agreed with both the victim-blaming and lying statements. Individual agreement with each statement was asked on a 5-point Likert scale from strongly disagree and strongly agree, and was binarized to be equal to 1 if a respondent agreed or strongly agreed and is 0 if else. The outcomes are: respondent's own agreement with the victim-blaming statement (column 1), respondent's own agreement with the lying statement (column 2), respondents' guesses about women's agreement with the victim-blaming statement (column 3), respondents' guesses about women's agreement with the lying statement (column 4), respondents' guesses about men's agreement with the victim-blaming statement (column 5), and respondent's guesses about men's agreement with the lying statement (column 6). "Female" is an indicator equal to 1 if the respondent was a woman, and is 0 if the respondent was a man. "Male Mean" is the average answer for each outcome by male respondents. No controls were included. Robust standard errors are in parentheses.

Table A6: Hard-Garbling Treatment Effects by Expectations of Victim Social Judgment - Continuous Measure

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling (HG)	-0.008 (0.038)	0.104*** (0.038)	0.069** (0.031)	0.051 (0.031)
Expectations of Victim Judgment (V)	0.003 (0.033)	0.019 (0.033)	-0.002 (0.022)	0.006 (0.025)
HG x V	-0.028 (0.052)	-0.054 (0.053)	0.033 (0.042)	-0.013 (0.044)
DE Mean	0.34	0.46	0.06	0.09
Exp. of Victim Judgment DE Mean	0.00	0.00	0.00	0.00
HG + HGxV = 0 (P-value)	0.573	0.453	0.058*	0.503
Observations	825	825	796	796
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. I first aggregate and standardize all four social norm predictions around the control mean within each gender, and then create an index using the Anderson method. This index is “Expectations of Victim Judgment.” “DE Mean” is the rate of reporting for each outcome in the direct elicitation treatment, and “Exp. of Victim Judgment DE Mean” is the average value of the social norm index in the DE treatment within each gender, standardized to be zero. Controls in all columns are selected using the Double Lasso CV method. Columns (1) and (2) are restricted to only victimization responses; columns (3) and (4) are restricted to only perpetration responses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A7: Statements Included in Social Desirability Scale

Statement
I sometimes feel resentful when I don't get my way.
It is sometimes hard for me to go on with my work if I am not encouraged.
On a few occasions, I have given up on doing something because I thought too little of my ability.
There have been times when I felt like rebelling against people in authority even though I knew they were right.
No matter who I'm talking to, I'm always a good listener.
I'm always willing to admit when I make a mistake.
I sometimes try to get even rather than forgive and forget.
I am always courteous, even to people who are disagreeable.
I have never been irked when people expressed ideas very different from my own.
There have been times when I was quite jealous of the good fortune of others.
I am sometimes irritated by people who ask favors of me.
I have deliberately said something that hurt someone's feelings.
There have been occasions when I took advantage of someone.

Notes: Table displays the statements used in the social desirability scale following Dhar et al. (2022) and Crowne and Marlowe (1960). In the experiment, agreement is asked on a 2-point scale, i.e., respondents can either “Agree” or “Disagree” with each statement.

Table A8: Hard-Garbling Treatment Effects by Expectations of Victim Social Judgment

	Perpetration	
	Rape (1)	Other SA (2)
Hard-Garbling (HG)	0.031 (0.043) [0.489]	0.084* (0.043) [0.111]
High Expectations of Victim Judgment (V)	-0.035 (0.026)	-0.018 (0.029)
HG x V	0.075 (0.061)	-0.067 (0.062)
DE + Low V Mean	0.06	0.06
HG + HGxV = 0 (P-value)	0.016**	0.712
Observations	796	796
Lasso Controls	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. I first aggregate and standardize all four social norm predictions around the control mean, and then create an index using the Anderson method. Then, I set “High Expectations of Victim Judgment” equal to one for individuals with index values above or equal to the median within their gender, and set it equal to zero for individuals with index values below the median within their gender. Controls in all columns are selected using the Double Lasso CV method. Columns (1) and (2) are restricted to only perpetration responses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A9: Hard-Garbling Treatment Effects – Expectations of and Sensitivity to Social Judgment (Perpetration)

	Perpetration	
	Rape (1)	Other SA (2)
<i>Panel A: High Expectations of Judgment</i>		
Hard-Garbling (HG)	0.135** (0.066)	0.116* (0.067)
Low Sensitivity to Judgment (Low SJ)	0.032 (0.034)	0.019 (0.043)
HG x Low SJ	-0.076 (0.090)	-0.189** (0.091)
DE + High SJ Mean	0.05	0.11
HG + HGxLow SJ = 0 (P-value)	0.316	0.226
Observations	397	397
<i>Panel B: Low Expectations of Judgment</i>		
Hard-Garbling (HG)	0.038 (0.058)	0.076 (0.057)
Low Sensitivity to Judgment (Low SJ)	0.020 (0.035)	0.016 (0.037)
HG x Low SJ	-0.013 (0.085)	0.011 (0.086)
DE + High SJ Mean	0.05	0.04
HG + HGxLow SJ = 0 (P-value)	0.687	0.181
Observations	399	399
Lasso Controls	Yes	Yes

Notes: OLS Regressions. Each column and panel is a separate regression. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. Responses are restricted to perpetration responses only. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. “Low Sensitivity to Judgment” is an indicator equal to 1 if the respondent’s social desirability score was below the mean within their gender, and is zero if else. “HG x Low SJ” is an interaction term. The experiment contained 4 social norm prediction questions. I aggregate and standardize these predictions around the control mean, and then create an indicator. Panel A restricts to only respondents who reported high expectations of social judgment, i.e., those with social norm guesses above or equal to the median within their gender. Panel B restricts to only respondents who report low expectations of social judgment, i.e., those with social norm guesses below the median within their gender. Controls in all columns and panels were selected via Double Lasso CV Method. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A10: Hard-Garbling Treatment Effects by Expectations of Victim Social Judgment

	Victimization	
	Rape (1)	Other SA (2)
Hard-Garbling (HG)	0.015 (0.053) [0.767]	0.128** (0.052) [0.019]
High Expectations of Victim Judgment (V)	-0.006 (0.048)	-0.004 (0.048)
HG x V	-0.046 (0.076)	-0.046 (0.076)
DE + Low V Mean	0.32	0.44
HG + HGxV = 0 (P-value)	0.563	0.130
Observations	825	825
Lasso Controls	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. I first aggregate and standardize all four social norm predictions around the control mean, and then create an index using the Anderson method. Then, I set “High Expectations of Victim Judgment” equal to one for individuals with index values above or equal to the median within their gender, and set it equal to zero for individuals with index values below the median within their gender. Controls in all columns are selected using the Double Lasso CV method. Columns (1) and (2) are restricted to only victimization responses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A11: Balance Test: Sexual Assault and Theft Victimization

	Panel A. Sexual Assault			Panel B. Theft			SA=T
	HG	DE	P-value	HG	DE	P-value	(P-value)
Age	22.93 (2.50)	23.04 (2.53)	0.52	22.54 (2.69)	22.53 (2.70)	0.98	0.01
White	0.57 (0.50)	0.58 (0.49)	0.77	0.50 (0.50)	0.60 (0.49)	0.07	0.42
Current Undergrad	0.49 (0.50)	0.43 (0.50)	0.11	0.48 (0.50)	0.55 (0.50)	0.21	0.11
Greek	0.19 (0.39)	0.23 (0.42)	0.17	0.22 (0.41)	0.24 (0.43)	0.59	0.41
College Athlete	0.22 (0.41)	0.25 (0.43)	0.25	0.33 (0.47)	0.41 (0.49)	0.16	0.00
Public Institution	0.73 (0.45)	0.78 (0.42)	0.08	0.73 (0.44)	0.78 (0.42)	0.35	0.94
Large Institution	0.58 (0.49)	0.56 (0.50)	0.68	0.51 (0.50)	0.50 (0.50)	0.82	0.04
Low-Income	0.45 (0.50)	0.54 (0.50)	0.01	0.48 (0.50)	0.59 (0.49)	0.05	0.21
First-Gen	0.29 (0.45)	0.29 (0.45)	0.97	0.27 (0.45)	0.30 (0.46)	0.61	0.99
Non-Religious	0.31 (0.46)	0.33 (0.47)	0.51	0.22 (0.42)	0.22 (0.42)	1.00	0.00
Southern	0.44 (0.50)	0.41 (0.49)	0.39	0.45 (0.50)	0.42 (0.49)	0.57	0.81
Democrat	0.49 (0.50)	0.45 (0.50)	0.34	0.41 (0.49)	0.37 (0.49)	0.56	0.01
Observations	417	408		153	153		

Notes: Numbers are averages across respondents by treatment. Standard deviations are in parentheses. Panel A includes only individuals who completed the survey on sexual assault victimization. Panel B includes only individuals who completed the survey on theft victimization. Individuals who incorrectly answered both comprehension questions in the hard-garbling treatment have been excluded. Numbers in p-value columns are generated from tests of equality of means across Hard-Garbling (HG) and Direct Elicitation (DE) treatments. The numbers in the last column are generated from tests of equality of means across all sexual assault and all theft respondents. All variables, excluding age, are 0/1 indicator variables. White takes a value of 1 if a respondent indicated they were white, and is 0 otherwise. Current Undergrad takes a value of 1 if a respondent indicated being a current undergraduate student. Greek and Athlete take a value of 1 if respondents indicated involvement in Greek life or collegiate athletics, respectively. Low-income takes a value of 1 if the respondent reported having ever been eligible for Pell Grants, and is 0 otherwise. Non-religious takes a value of 1 if the respondent indicated being non-religious (like Atheist or agnostic) and is zero otherwise. Public Institution takes a value of 1 if a respondent indicated that their higher education institution was public, and is 0 otherwise. Large Institution takes a value of 1 if a respondent indicated that their institution was large, and is 0 if else. Democrat takes a value of 1 if a respondent reported voting for Kamala Harris in the 2024 presidential election. Southern is an indicator for whether a respondent reported attending an undergraduate institution in the South, according to U.S. Census Bureau region designations.

Table A12: Hard-Garbling Treatment Effects: Grouping Outcomes by Severity

	No Controls		Lasso Selected Controls	
	Severe (1)	Less Severe (2)	Severe (3)	Less Severe (4)
Hard-Garbling (HG)	0.028 (0.064)	0.047 (0.064)	0.052 (0.056)	0.075 (0.059)
Sexual Assault (SA)	-0.016 (0.045)	0.018 (0.047)	0.039 (0.044)	0.051 (0.046)
HG x SA	-0.060 (0.075)	0.035 (0.075)	-0.058 (0.068)	0.031 (0.070)
DE Theft Mean	0.359	0.438	0.359	0.438
HG + HGxSA = 0 (P-value)	0.398	0.033**	0.876	0.005***
Observations	1131	1131	1131	1131
Lasso Controls	No	No	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the experience questions using the hard-garbling survey technique, and is 0 if direct elicitation. “Sexual Assault” is an indicator equal to 1 if a respondent answered the experience questions on sexual assault, and is 0 if theft. “HG x SA” is an interaction term between the HG and SA indicators. The “Severe” outcome is either rape or theft of property over \$1,000. The “Less Severe” outcome is other types of sexual assault or theft of property under \$1,000. Columns 1 and 2 contain no controls. Columns 3 and 4 contain controls selected using the Lasso CV selection method. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A13: Comparison of Prolific and TAMU Respondents

	Age	White	Current Undergrad	Greek	College Athlete	Public Institution	Large Institution	Low-Income	First-Gen	Non-Religious	Southern	Democrat	N
<i>Panel A. Women</i>													
Prolific	22.98 (2.51)	0.58 (0.49)	0.46 (0.50)	0.21 (0.40)	0.24 (0.42)	0.75 (0.43)	0.57 (0.50)	0.49 (0.50)	0.29 (0.45)	0.32 (0.47)	0.43 (0.49)	0.47 (0.50)	825
TAMU	20.84 (1.60)	0.42 (0.49)	1.00 (0.00)	0.13 (0.34)	0.05 (0.22)	1.00 (0.00)	1.00 (0.00)	0.35 (0.48)	0.23 (0.42)	0.23 (0.42)	1.00 (0.00)	0.36 (0.48)	495
P-value (Prol=TAMU)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	
<i>Panel B. Men</i>													
Prolific	23.42 (2.68)	0.55 (0.50)	0.46 (0.50)	0.18 (0.38)	0.25 (0.43)	0.76 (0.43)	0.58 (0.49)	0.48 (0.50)	0.27 (0.44)	0.32 (0.47)	0.36 (0.48)	0.37 (0.48)	796
TAMU	20.83 (1.41)	0.48 (0.50)	1.00 (0.00)	0.11 (0.31)	0.10 (0.30)	1.00 (0.00)	1.00 (0.00)	0.25 (0.43)	0.21 (0.41)	0.26 (0.44)	1.00 (0.00)	0.21 (0.41)	349
P-value (Prol=TAMU)	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.00	

Notes: Numbers are averages across respondents. Standard deviations are in parentheses. Panel A includes only individuals who completed the survey on sexual assault victimization. Panel B includes only individuals who completed the survey on sexual assault perpetration. Individuals who incorrectly answered both comprehension questions in the hard-garbling treatment have been excluded. Numbers in p-value row are generated from tests of equality of means across Prolific and TAMU respondents. All variables, excluding age, are 0/1 indicator variables. White takes a value of 1 if a respondent indicated they were white, and is 0 otherwise. Current Undergrad takes a value of 1 if a respondent indicated being a current undergraduate student. Greek and Athlete take a value of 1 if respondents indicated involvement in Greek life or collegiate athletics, respectively. Low-income takes a value of 1 if the respondent reported having ever been eligible for Pell Grants, and is 0 otherwise. Non-religious takes a value of 1 if the respondent indicated being non-religious (like Atheist or agnostic) and is zero otherwise. Public Institution takes a value of 1 if a respondent indicated that their higher education institution was public, and is 0 otherwise. Large Institution takes a value of 1 if a respondent indicated that their institution was large, and is 0 if else. Democrat takes a value of 1 if a respondent reported voting for Kamala Harris in the 2024 presidential election. Southern is an indicator for whether a respondent reported attending an undergraduate institution in the South, according to U.S. Census Bureau region designations.

Table A14: Social Norm Predictions - TAMU

	Guesses About Women		Guesses About Men	
	VB (1)	Lying (2)	VB (3)	Lying (4)
Female	4.396*** (1.417)	3.054** (1.459)	17.184*** (1.511)	5.651*** (1.556)
Male Mean	16.91	23.57	39.35	54.27
Observations	844	844	844	844
Controls	No	No	No	No

Notes: OLS regressions. Each column is a separate regression. Respondents were asked separately about the predicted percent of men and women in a prior study that they believe would have agreed with both the victim-blaming and lying statements. The outcomes were: respondents' guesses about women's agreement with the victim-blaming statement (column 1), respondents' guesses about women's agreement with the lying statement (column 2), respondents' guesses about men's agreement with the victim-blaming statement (column 3), and respondent's guesses about men's agreement with the lying statement (column 4). "Female" is an indicator equal to 1 if the respondent was a woman, and is 0 if the respondent was a man. "Male Mean" is the average answer for each outcome by male respondents. No controls were included. Robust standard errors are in parentheses. Responses were restricted to only those in the TAMU sample.

Table A15: Perceptions of Social Norms - TAMU vs. Prolific Respondents

	Victim-Blaming			Lying		
	Actual Agreement	Perception of Men	Perception of Women	Actual Agreement	Perception of Men	Perception of Women
<i>Panel A. Men</i>						
Prolific (P)	16.46 (37.10)	40.16 (22.65)	20.55 (21.85)	49.50 (50.03)	50.92 (22.27)	28.18 (23.09)
TAMU (T)	10.60 (30.83)	39.35 (21.52)	16.91 (19.71)	53.30 (49.96)	54.27 (22.39)	23.57 (20.24)
P-value (P=T)	0.01	0.57	0.01	0.24	0.02	0.00
<i>Panel B. Women</i>						
Prolific (P)	7.52 (26.38)	48.47 (23.60)	24.71 (22.82)	27.88 (44.87)	55.42 (22.70)	30.21 (22.47)
TAMU (T)	5.66 (23.12)	56.53 (21.76)	21.30 (21.05)	26.46 (44.16)	59.92 (22.08)	26.62 (21.74)
P-value (P=T)	0.19	0.00	0.01	0.58	0.00	0.00

Notes: Table shows the raw averages of answers from 796 male Prolific respondents, 825 female Prolific respondents, 349 male TAMU respondents, and 495 female TAMU respondents. Agreement with each statement was asked on a 5-point Likert scale and was binarized to be equal to 1 if a respondent agreed or strongly agreed. Respondents were asked to separately predict the percent of men and women from a prior study that agreed with each statement; respondents were randomly assigned to answer regarding men or women first. P-values disclosed in the last rows of each panel are generated from tests of equalities of means between male Prolific and male TAMU respondents (in Panel A) and between female Prolific and female TAMU respondents (in Panel B).

Table A16: Balance Table - TAMU Students

	Panel A. Women			Panel B. Men			M=F
	HG	DE	P-value	HG	DE	P-value	(P-value)
Age	20.80 (1.46)	20.91 (1.77)	0.45	20.90 (1.39)	20.71 (1.46)	0.23	0.94
White	0.43 (0.50)	0.39 (0.49)	0.34	0.48 (0.50)	0.48 (0.50)	0.97	0.06
Current Undergrad	1.00 (0.00)	1.00 (0.00)	.	1.00 (0.00)	1.00 (0.00)	.	.
Greek	0.14 (0.35)	0.11 (0.31)	0.23	0.12 (0.33)	0.09 (0.28)	0.35	0.37
College Athlete	0.06 (0.24)	0.04 (0.20)	0.39	0.11 (0.31)	0.08 (0.27)	0.41	0.01
Public Institution	1.00 (0.00)	1.00 (0.00)	.	1.00 (0.00)	1.00 (0.00)	.	.
Large Institution	1.00 (0.00)	1.00 (0.00)	.	1.00 (0.00)	1.00 (0.00)	.	.
Low-Income	0.33 (0.47)	0.38 (0.49)	0.22	0.26 (0.44)	0.23 (0.42)	0.58	0.00
First-Gen	0.22 (0.41)	0.26 (0.44)	0.25	0.20 (0.40)	0.24 (0.43)	0.34	0.45
Non-Religious	0.23 (0.42)	0.23 (0.42)	0.99	0.26 (0.44)	0.24 (0.43)	0.63	0.33
Southern	1.00 (0.00)	1.00 (0.00)	.	1.00 (0.00)	1.00 (0.00)	.	.
Democrat	0.34 (0.47)	0.39 (0.49)	0.28	0.21 (0.41)	0.20 (0.40)	0.83	0.00
Observations	283	212		224	125		

Notes: Numbers are averages across female and male respondents by treatment, respectively. Standard deviations are in parentheses. Individuals who incorrectly answered both comprehension questions in the hard-garbling treatment have been excluded. Numbers in p-value columns are generated from tests of equality of means across Hard-Garbling (HG) and Direct Elicitation (DE) treatments. The numbers in the last column are generated from tests of equality of means across all male and all female respondents. All variables, excluding age, are 0/1 indicator variables. White takes a value of 1 if a respondent indicated they were white, and is 0 otherwise. Current Undergrad takes a value of 1 if a respondent indicated being a current undergraduate student. Greek and Athlete take a value of 1 if respondents indicated involvement in Greek life or collegiate athletics, respectively. Low-income takes a value of 1 if the respondent reported having ever been eligible for Pell Grants, and is 0 otherwise. Non-religious takes a value of 1 if the respondent indicated being non-religious (like Atheist or agnostic) and is zero otherwise. Public Institution takes a value of 1 if a respondent indicated that their higher education institution was public, and is 0 otherwise. Large Institution takes a value of 1 if a respondent indicated that their institution was large, and is 0 if else. Democrat takes a value of 1 if a respondent reported voting for Kamala Harris in the 2024 presidential election. Southern is an indicator for whether a respondent reported attending an undergraduate institution in the South, according to U.S. Census Bureau region designations.

Table A17: Hard-Garbling Treatment Effects - TAMU Replication

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling	0.036 (0.045)	0.008 (0.047)	0.011 (0.040)	0.042 (0.037)
Observations	495	495	349	349
Direct Elicitation Mean	0.184	0.278	0.024	0.008
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. Responses included are restricted to only those in the TAMU replication. Hard-Garbling is an indicator equal to 1 if an individual answered the sexual assault experience questions using the hard-garbling survey technique, and is 0 if they answered using direct elicitation. Controls in all columns are selected using the Double Lasso CV method. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A18: Hard-Garbling Treatment Effects – Sensitivity to Social Judgment (TAMU)

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling (HG)	0.080 (0.057)	0.113* (0.061)	0.059 (0.052)	0.042 (0.050)
Low Sensitivity to Judgment (SJ)	0.083 (0.053)	0.132** (0.058)	0.029 (0.032)	-0.016 (0.014)
HG x SJ	-0.087 (0.091)	-0.236** (0.094)	-0.115 (0.076)	0.001 (0.074)
DE + High SJ Mean	0.15	0.21	0.01	0.01
HG + HGxSJ = 0 (P-value)	0.922	0.088*	0.342	0.435
Observations	495	495	349	349
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. Responses included are restricted to only those in the TAMU replication. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. Controls in all columns were selected using the Double Lasso CV method. “Low Sensitivity to Judgment” is an indicator equal to 1 if the respondent’s social desirability score was below the mean within their gender among TAMU respondents, and is zero if else. “HG x SJ” is an interaction term. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A19: Hard-Garbling Treatment Effects with Different Vectors of Controls

	Rape			Other SA		
	No Controls (1)	Additional Controls (2)	Lasso Controls (3)	No Controls (4)	Additional Controls (5)	Lasso Controls (6)
<i>Panel A: Victimization</i>						
Hard-Garbling	-0.033 (0.039)	-0.008 (0.038)	-0.010 (0.037)	0.082** (0.039)	0.105*** (0.038)	0.105*** (0.038)
Observations	825	825	825	825	825	825
Direct Elicitation Mean	0.34	0.34	0.34	0.46	0.46	0.46
Controls	No	Yes	Yes	No	Yes	Yes
<i>Panel B: Perpetration</i>						
Hard-Garbling	0.053* (0.031)	0.069** (0.031)	0.069** (0.031)	0.032 (0.032)	0.050 (0.031)	0.050 (0.031)
Observations	796	796	796	796	796	796
Direct Elicitation Mean	0.06	0.06	0.06	0.09	0.09	0.09
Controls	No	Yes	Yes	No	Yes	Yes

OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. There are no controls included in Columns 1 or 4. Columns 2 and 5 contain the entire vector of possible controls used for Double Lasso selection (controls for: Southern institutions, large institutions, low-income student status, agreement with both social norms statements, perceptions of norms and sexual assault prevalence, sensitivity to judgment, Greek life, intercollegiate athletic involvement, public institutions, political affiliation, race, undergraduate student status, religion, first-gen status, age, beliefs that their survey responses were anonymous, and whether they completed the survey in the first round of data collection (December 2024)). Columns 3 and 6 use the preferred specification, i.e., controls selected using the Double Lasso CV method. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A20: Hard-Garbling Treatment Effects–Dropping Incorrect Comprehension Questions and Failed Attention Checks

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
<i>Panel A: Whole Sample</i>				
Hard-Garbling	-0.007 (0.037)	0.102*** (0.037)	0.071** (0.030)	0.051* (0.031)
DE Mean	0.343	0.456	0.062	0.090
Observations	833	833	806	806
<i>Panel B: Dropped 2 Wrong Comp. Qs</i>				
Hard-Garbling	-0.010 (0.037)	0.105*** (0.038)	0.069** (0.031)	0.050 (0.031)
DE Mean	0.343	0.456	0.062	0.090
Observations	825	825	796	796
<i>Panel C: Dropped 1+ Wrong Comp. Qs</i>				
Hard-Garbling	0.002 (0.039)	0.112*** (0.038)	0.077** (0.032)	0.048 (0.032)
DE Mean	0.343	0.456	0.062	0.090
Observations	794	794	773	773
<i>Panel D: Dropped 1+ Failed Attention Checks</i>				
Hard-Garbling	-0.018 (0.039)	0.087** (0.039)	0.088*** (0.032)	0.052 (0.032)
DE Mean	0.335	0.456	0.052	0.088
Observations	766	766	728	728
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. "Hard-Garbling" is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. Controls in all columns and panels selected using the Double Lasso CV method. Panel A contains the entire 1,639 respondents who completed the survey on sexual assault; Panel B drops those that incorrectly answered both hard-garbling comprehension question (this is the primary sample used for analysis). Panel C drops individuals that incorrectly answered either hard-garbling comprehension question, and Panel D drops individuals that missed one or more of the three total attention checks in the experiment. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A21: Hard-Garbling Treatment Effects – Attrition

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
<i>Panel A: All Recorded Responses</i>				
Hard-Garbling	-0.004 (0.037)	0.106*** (0.038)	0.070** (0.031)	0.035 (0.032)
DE Mean	0.342	0.455	0.062	0.089
Observations	836	836	818	817
<i>Panel B: Missing Attrited Answers Recoded as No</i>				
Hard-Garbling	-0.008 (0.037)	0.100*** (0.038)	0.069** (0.030)	0.033 (0.031)
DE Mean	0.342	0.455	0.062	0.089
Observations	839	839	819	819
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. Controls in all columns and panels were selected using the Double Lasso CV method from demographic controls. Panel A contains the survey responses for all individuals with non-missing responses for the outcome. Panel B contains all respondents who were successfully screened into the survey; if these respondents did not answer the primary outcome questions, they were recoded as “No” responses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A22: Hard-Garbling Treatment Effects - Dropping Respondents with $p = 0$

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling	-0.016 (0.041) [0.706]	0.104** (0.042) [0.028]	0.079** (0.034) [0.037]	0.075** (0.034) [0.037]
Observations	691	691	685	685
Direct Elicitation Mean	0.350	0.459	0.067	0.092
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Each column is a separate regression. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. The survey contained two questions regarding the respondent's belief that their answers were truly anonymous. Respondents who indicated that there was a 0 percent chance Prolific shares identifying information with researchers and a 0 percent chance that researchers could link their responses to their identity were excluded from these regressions. Hard-Garbling is an indicator equal to 1 if an individual answered the sexual assault experience questions using the hard-garbling survey technique, and is 0 if they answered using direct elicitation. Controls in all columns are selected using the Double Lasso CV method. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets, adjusted for victimization and perpetration outcomes separately. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A23: Hard-Garbling Treatment Effects by $p = 0$

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling (HG)	-0.010 (0.041) [0.804]	0.108*** (0.041) [0.022]	0.078** (0.034) [0.037]	0.081** (0.034) [0.037]
Perceived Zero Leak Probability (Zero LP)	-0.006 (0.062)	-0.007 (0.065)	-0.021 (0.027)	-0.005 (0.040)
HG x Zero LP	0.011 (0.102)	-0.013 (0.103)	-0.068 (0.075)	-0.167** (0.077)
DE + Non-Zero LP Mean	0.35	0.46	0.07	0.09
HG + HGxZero LP = 0 (P-value)	0.990	0.315	0.885	0.211
Observations	825	825	796	796
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. The survey contained two questions regarding the respondent’s belief that their answers were truly anonymous as a measure of this leak probability: 1) What is the probability that Prolific shares your identifying information, like name or birth date, with researchers?, and 2) What is the probability that researchers using Prolific to administer surveys are able to link your survey responses to your identity? “Zero LP” is an indicator equal to 1 if a respondent answered “0” to both anonymity questions, i.e., if they perceived a leak probability of zero, and is equal to 0 otherwise. There were 111 men and 134 women who indicated a perceived leak probability of zero. “HG x Zero LP” is an interaction term. “DE Mean + Non-Zero LP Mean” is the rate of reporting for each outcome in the direct elicitation treatment for individuals who perceived a leak probability greater than zero. Controls in all columns are selected using the Double Lasso CV method. Columns (1) and (2) are restricted to only victimization responses; columns (3) and (4) are restricted to only perpetration responses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets, adjusted for victimization and perpetration outcomes separately. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A24: Hard-Garbling Treatment Effects by Leak Probability

	Victimization		Perpetration	
	Rape (1)	Other SA (2)	Rape (3)	Other SA (4)
Hard-Garbling (HG)	-0.047 (0.053) [0.390]	0.095* (0.053) [0.120]	0.150*** (0.046) [0.005]	0.073 (0.044) [0.097]
Perceived Lower Leak Probability (Low LP)	0.013 (0.046)	0.036 (0.048)	0.004 (0.025)	0.018 (0.029)
HG x Low LP	0.075 (0.075)	0.016 (0.075)	-0.172*** (0.061)	-0.048 (0.062)
DE + High LP Mean	0.36	0.45	0.07	0.10
HG + HGxLow LP = 0 (P-value)	0.606	0.037**	0.580	0.564
Observations	825	825	796	796
Lasso Controls	Yes	Yes	Yes	Yes

Notes: OLS Regressions. Robust standard errors in parentheses. Outcome variables were normalized to account for garbled responses. “Hard-Garbling” is an indicator equal to 1 if an individual answered the sexual assault questions using the hard-garbling survey technique, and is 0 if direct elicitation. The survey contained two questions regarding the respondent’s belief that their answers were truly anonymous as a measure of this leak probability: 1) What is the probability that Prolific shares your identifying information, like name or birth date, with researchers?, and 2) What is the probability that researchers using Prolific to administer surveys are able to link your survey responses to your identity? “Low LP” is an indicator equal to 1 if a respondent perceived a leak probability below the median within their gender, and is equal to 0 otherwise. “HG x Low LP” is an interaction term. “DE Mean + High LP Mean” is the rate of reporting for each outcome in the direct elicitation treatment for individuals who perceived a leak probability greater than the median. Controls in all columns are selected using the Double Lasso CV method. Columns (1) and (2) are restricted to only victimization responses; columns (3) and (4) are restricted to only perpetration responses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets, adjusted for victimization and perpetration outcomes separately. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

B Additional Details on Framework

In this section, I discuss how the framework can be extended to consider formal reporting, the possibility of “false reports,” and the perpetrator’s decision to report.

B.1 Formal Reporting

While the framework included discusses only informal reports, the formal reporting decision is related. The costs and benefits of informal reporting likely generalize to formal reporting. Victims choosing a binary formal reporting decision against their perpetrator will still face intrinsic benefits, social benefits, intrinsic costs, and social costs – though these factors may be larger in magnitude for the formal reporting decision relative to the informal.

For example, victims who report their assaults to Title IX offices or law enforcement will likely face larger expected social costs. Unlike informal reporting, where a perpetrator is not identified, a victim may experience additional social backlash for “ruining” her perpetrator’s life as a result of the report. If the victim chooses to submit a formal report to law enforcement or Title IX, they also now face potential negative social reactions from law enforcement or Title IX administrators. In addition, a formal report to law enforcement or Title IX offices will result in less anonymity than informal reports. The perpetrator would likely be informed of the report, subjecting the victim to negative social reactions from the perpetrator or their friends and family. Victims may also expect their formal reports to result in news coverage of their assaults, which could result in additional negative social reactions from the broader public or additional reputational damage. Combined, these forces increase the size of expected social costs for the formal reporting decision.

There will also be additional costs and benefits for formal reporting. More specifically, formal reporting will be associated with extrinsic, or tangible, costs and benefits. These costs and benefits include any utility lost from procedures, time, or other material aspects of reporting, including costs of retaliation. For example, extrinsic costs include utility lost from the invasive nature of reporting sexual assault or from recounting the event to law enforcement. Extrinsic costs would also capture utility lost if a perpetrator retaliates against

their victim in the event of a positive report. Alternatively, extrinsic benefits would include any legal or academic benefits obtained as a result of a report. For example, this term would capture any utility gained from a perpetrator facing legal consequences or disciplinary action at their university and would also include any financial compensation received from a report.

Thus, the utility function for a student i from a formal reporting decision r_{is} given a victimization status v_{is} would be:

$$U_{is}(r_{is}|v_{is}) = IB_{is}(r_{is}|v_{is}) + SB_{is}(r_{is}|v_{is}) + EB_{is}(r_{is}|v_{is}) - IC_{is}(r_{is}|v_{is}) - SC_{is}(r_{is}) - EC_{is}(r_{is}|v_{is}), \text{ where} \quad (\text{B.1})$$

IB (intrinsic benefits), IC (intrinsic costs), SB (social benefits), and SC (social costs) follow the informal reporting decision, and EB and EC represent extrinsic costs and benefits as described above.

B.2 False Reports

Recall the assumptions discussed in Section 2.1. For simplicity, I focus on the case of victimization. Based on these assumptions, a non-victimized student will find it optimal to report $r_{is} = 0$. A non-victimized student has the following expected utilities from $r_{is} = 1$ and $r_{is} = 0$ (notation has omitted the expectations for simplicity):

$$\begin{aligned} U_{is}(1|0) &= IB_{is}(1|0) + SB_{is}(1|0) - IC_{is}(1|0) - p \times SC_{is}(1) \\ &= SB_{is}(1|0) - p \times SC_{is}(1) < 0 \end{aligned} \quad (\text{B.2})$$

and

$$\begin{aligned}
U_{is}(0|0) &= IB_{is}(0|0) + (1 - \pi) \times SB_{is}(0|0) \\
&\quad + \pi \times SB_{is}(1|0) - IC_{is}(0|0) - \pi \times (p \times SC_{is}(1)) \\
&= \pi \times (SB_{is}(1|0) - p \times SC_{is}(1)) \\
&= \pi \times U_{is}(1|0)
\end{aligned} \tag{B.3}$$

Thus, $U_{is}(0|0) > U_{is}(1|0)$, and a student with $v_{is} = 0$ will not find it optimal to report $r_{is} = 1$. The analogous math for perpetrators will similarly reveal that a non-perpetrator will not find it optimal to report a perpetration when one has not occurred.

However, there may be violations of the assumptions outlined in Section 2.1. One potential concern is the existence of “altruistic liars” who gain utility from reporting a victimization or perpetration that has not occurred. Such individuals might believe that true victims or perpetrators systematically under-report and view their own false reports as correcting the observed under-reporting gap. This behavior would violate the assumptions that individuals receive no intrinsic or social benefits from reporting a victimization or perpetration when one has not occurred, i.e., $IB_{is}(1|0) = 0$ and $SB_{is}(1|0) < 0$. If individuals derive positive utility from “correcting” under-reporting, the magnitude of that utility is likely increasing in the perceived degree of under-reporting: even altruistic liars should gain more from a false report when they believe under-reporting is especially severe.

Hard-garbling does reduce the perceived costs associated with a “Yes” report and could, in principle, make altruistic lying more attractive, as well as encourage true victims and perpetrators to report honestly. However, because hard-garbling also induces more true victims and perpetrators to submit “Yes” reports, the remaining scope for altruistic liars to “correct” under-reporting is reduced. In this sense, hard-garbling simultaneously weakens the main motive for altruistic lying (a large perceived under-reporting gap) while strengthening incentives for truthful reporting.

In addition, false reports for perpetration are likely rare given the potential high costs associated with admitting a sexual assault perpetration. While it may be seen as “irrational” to admit perpetration, and thus perhaps could imply all reports of perpetration are false

reports, there is prior evidence that men will self-report these behaviors. While there is currently no accepted best practice for eliciting perpetration experiences, many psychology studies use sexual experiences surveys. In these surveys, respondents see a list of specific items and respond with “yes” or “no” to denote whether they have taken the action. For example, one version of this survey (Johnson et al., 2017) includes the following item: “I put my penis into a woman’s vagina without her consent.” In a review study, Anderson et al. (2021) found that prevalence of rape perpetration ranged from 0% to nearly one third, with the average rate reported as 6.5%.⁵⁰

B.3 Informal Reporting: Perpetration

I again assume that the utility function for a student i from an informal reporting decision z_{is} given an assailant status a_{is} is a function of costs and benefits:

$$U_{is}(z_{is}|a_{is}) = IB_{is}(z_{is}|a_{is}) + SB_{is}(z_{is}|a_{is}) - IC_{is}(z_{is}|a_{is}) - SC_{is}(z_{is}), \quad (\text{B.4})$$

where:

- $IB_{is}(z_{is}|a_{is})$ denotes intrinsic or emotional benefits. These benefits include any internal utility gained from reporting. Perpetrators may experience alleviations of guilt for admitting their behavior. I assume that there are positive emotional benefits from a positive report of severity s given that a perpetration of severity s occurred and that otherwise emotional benefits are zero. Thus, I assume $IB_{is}(1|1) > 0$, and $IB_{is}(1|0) = IB_{is}(0|0) = IB_{is}(0|1) = 0$. These intrinsic benefits are likely larger for perpetrators of severity h .
- $SB_{is}(z_{is}|a_{is})$ denotes social benefits. Again, these benefits include any additional utility gained from informing policy.⁵¹ I assume that $SB_{is}(1|1) > 0$, $SB_{is}(1|0) < 0$, and

⁵⁰There are very few studies that use direct measures of rape or sexual assault perpetration rather than behavioral items like those from sexual experience surveys. However, in studies that do ask directly “Do you think you may have ever raped someone?”, the rates of self-reporting fall substantially, with the average rate of reporting in these studies below one percent (Anderson et al., 2021).

⁵¹These benefits will also include any social support experienced as a result of a positive report. This likely varies in magnitude for perpetrators relative to victims. Sociology and related fields have well documented a phenomenon known as “himpathy,” where perpetrators of sexual violence receive sympathy from society at

$SB_{is}(0|0) = SB_{is}(0|1) = 0$.⁵² Social benefits from informing policy are likely larger for perpetrators of severity h , as individuals may gain more utility from informing policies on more severe assaults. However, social benefits from social support are likely larger for perpetrators of severity l , as perpetrators who committed less severe assaults may be more likely to benefit from “himpathy.” Thus, we should not necessarily expect social benefits to be larger for perpetrators of either severity l or h .

- $IC_{is}(z_{is}|a_{is})$ denote intrinsic or emotional costs. Like victims, admitting perpetration may erode their self-identities or cause otherwise uncomfortable feelings. Perpetrators may also face a cost for not reporting an assault when one has occurred, due to unresolved feelings of guilt. Thus, I assume that $IC_{is}(1|1) \geq 0$, $IC_{is}(0|1) \geq 0$ and $IC_{is}(0|0) = IC_{is}(1|0) = 0$.
- $SC_{is}(z_{is})$ denotes social or reputation costs. These costs again broadly include any damage to a student’s reputation or social standing and any utility lost as a result of negative social reactions from others. I assume $SC_{is}(1) > 0$ and $SC_{is}(0) = 0$ and again discuss these costs in more detail below.

Similar to the decision to report for victims, the expected social costs term $SC_{is}(z_{is})$ captures lost utility from damage to a perpetrator’s reputation or negative social reactions from self-reporting. While both norms on assault and norms on reporting are relevant to a victim’s decision to report, we may expect only norms on assault to be relevant to a perpetrator’s decision to self-report. Social norms on reporting, like those regarding whether women should report or whether victims are perceived as lying, will not be relevant when self-reporting an act of perpetration and thus δ_{is} are excluded from the expected social cost

the expensive of their victims (Bedera, 2024). This could increase the size of social benefits (or decrease the size of social costs), relative to victims. However, an individual’s expectations of social support are likely the inverse of their expectations of social costs (as an individual expects higher social costs and/or more damage to their reputation, they likely also expect less social support). Because this dynamic is captured by the social costs term $SC_{is}(z_{is})$, I make a simplifying assumption $SB_{is}(z_{is}|a_{is})$ is utility from informing policy.

⁵²I assume people have negative social utility from reporting a perpetration when one has not occurred. For example, consider the case in which individuals receive utility from informing policy. Then, if $z_{is} = 1$ but $a_{is} = 0$, policy-makers are now developing policy based on incorrect information, which I assume causes negative utility.

term.⁵³ However, norms on the broader acceptability of assault will likely still be relevant to a perpetrator's decision to self-report their behavior. If a perpetrator thinks that sexual assault is more acceptable and thus that he is less likely to be socially condemned for the behavior, his expected social costs likely decrease. Recall that higher values of θ_{is} imply that social norms such that sexual assault is viewed as more socially acceptable. For perpetrators, this now implies that higher values of θ_{is} *decrease* expected social costs. To maintain that higher values of the social norm parameter imply that social costs are increasing, I let $SC_{is}(z_{is})$ be a function of $\psi_{is} = 1 - \theta_{is}$.

Closely following the victim's decision, we should also expect social costs for perpetrators to be larger for individuals who care more about social judgment. Thus, expected social costs for perpetrators are also a function of the social judgment sensitivity parameter $\phi_i > 0$. Therefore, I let:

$$E[SC_{is}(z_{is})] = \phi_i K(\psi_{is}, z_{is}), \quad (\text{B.5})$$

where ψ_{is} are student i 's perceptions of social norms on sexual assaults of severity s such that higher values imply sexual assault is *less* socially accepted, ϕ_i is student i 's sensitivity to social judgment, and K is a positive and strictly increasing function of ψ_{is} .

An individual with $a_{is} = 1$ will find it optimal to report $z_{is} = 1$ when the expected utility from a positive report is larger than the expected utility from not reporting. They have the following payoffs from $U_{is}(1|1)$ and $U_{is}(0|1)$ ⁵⁴:

$$U_{is}(1|1) = IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - SC_{is}(1) \quad (\text{B.6})$$

$$U_{is}(0|1) = -IC_{is}(0|1) \quad (\text{B.7})$$

Then, an individual with $a_{is} = 1$ will find it optimal to report $z_{is} = 1$ when:

$$IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - SC_{is}(1) + IC_{is}(0|1) \geq 0 \quad (\text{B.8})$$

⁵³For example, if a perpetrator is admitting that he sexually assaulted someone, the victim will not somehow be accused of lying, and thus these norms should be irrelevant to the perpetrator's decision to report.

⁵⁴The expectations have been omitted below for simplicity of notation

Recall that $E[SC(z_{is})] = \phi_i K(\psi_{is}, z_{is})$ such that K is a strictly positive and increasing function of ψ_{is} , which represent perceived social norms such that higher values of ψ_{is} indicate that assaults of severity s are less socially accepted. Then, Equation (B.8) becomes:

$$IB_{is}(1|1) + SB_{is}(1|1) - IC_{is}(1|1) - \phi_i K(\psi_{is}, 1) + IC_{is}(0|1) \geq 0 \quad (\text{B.9})$$

Equation (B.9) then demonstrates that as individuals perceive that assault is less socially accepted (i.e., as ψ_{is} increases in value), expected social costs also increase. Thus, holding all else equal, individuals with higher values of ψ_{is} are less likely to report.

B.4 Proofs of Predictions 2-4

This section will focus on the victim's decision for simplicity, though the math is analogous for perpetrators. Consider that hard-garbling is more effective when more true victims and perpetrators are induced to submit Yes responses. Define the “net benefit” of hard-garbling as the difference between a true victim's utility from reporting under HG and their utility from reporting under DE:

$$\Delta NB_{is} = NB_{is}^{HG} - NB_{is}^{DE}$$

For ease of notation, let $A = IB_{is}(1|1) - IC_{is}(1|1) + IC_{is}(0|1)$, and $SB = SB_{is}(1|1)$. Then, following Equation (9),

$$\begin{aligned} NB_{is}^{HG} &= A + (1 - \pi) \times [SB - p \times SC_{is}^{HG}(1)] \\ NB_{is}^{DE} &= A + SB - p \times SC_{is}^{DE}(1) \end{aligned}$$

Recall that, under multiplicative separability, $SC_{is}^{HG}(1) = \phi_i K(\delta_{is}, \theta_{is}) \times \text{prob}(r_{is} = 1 | \tilde{r}_{is} = 1)$ under HG. Under DE, $\text{prob}(r_{is} = 1 | \tilde{r}_{is} = 1) = 1$. For simplicity, let $\text{prob}(r_{is} = 1 | \tilde{r}_{is} = 1) = q$ under HG. In addition, under hard-garbling with $\pi > 0$ and $\text{prob}(r_{is} = 0) > 0$,

$q < 1$. Then,

$$\begin{aligned}
\Delta NB_{is} &= \left[A + (1 - \pi) \times [SB - p \times SC_{is}^{HG}(1)] \right] - \left[A + SB - p \times SC_{is}^{DE}(1) \right] \\
&= -p(1 - \pi) \left[\phi_i K(\delta_{is}, \theta_{is}) \times q \right] + p \times \phi_i K(\delta_{is}, \theta_{is}) - \pi SB \\
&= p\phi_i \left[(\pi - 1) [K(\delta_{is}, \theta_{is}) \times q] + K(\delta_{is}, \theta_{is}) \right] - \pi SB
\end{aligned} \tag{B.10}$$

B.4.1 Proof of Prediction 3: Sensitivity to Judgment

Prediction 3 states that hard-garbling is more effective for individuals more sensitive to social judgment, i.e., those with higher values of ϕ_i . For this to be true, ΔNB_{is} must be increasing in ϕ_i .

Taking the derivative of Equation B.10 with respect to ϕ_i yields the following:

$$\frac{\partial \Delta NB_{is}}{\partial \phi_i} = p \times \left[K(\delta_{is}, \theta_{is}) - (1 - \pi) [K(\delta_{is}, \theta_{is}) \times q] \right] \tag{B.11}$$

Assuming $p > 0$, this expression is always greater than 0:

$$\begin{aligned}
p \times \left[K(\delta_{is}, \theta_{is}) - (1 - \pi) [K(\delta_{is}, \theta_{is}) \times q] \right] &> 0 \\
\implies K(\delta_{is}, \theta_{is}) - (1 - \pi) [K(\delta_{is}, \theta_{is}) \times q] &> 0 \\
\implies K(\delta_{is}, \theta_{is}) &> (1 - \pi) [K(\delta_{is}, \theta_{is}) \times q]
\end{aligned} \tag{B.12}$$

Because $K(\delta_{is}, \theta_{is}) > K(\delta_{is}, \theta_{is}) \times q$ and $K(\delta_{is}, \theta_{is}) \times q > (1 - \pi) [K(\delta_{is}, \theta_{is}) \times q]$, this partial derivative is always positive. Thus, the increase in the incentive to report induced by hard-garbling is larger for individuals with higher ϕ_i , and hard-garbling is more effective in inducing reporting among individuals who are more sensitive to social judgment.⁵⁵

⁵⁵While the math here uses the notation from the victim's decision to report, the math is analogous for the perpetrator.

B.4.2 Proof of Predictions 2 and 4: Social Norms

Prediction 2 states that 2a) Hard-garbling is more effective at increasing reports of victimization when θ_{is} is larger, i.e., when $s = l$, and 2b) Hard-garbling is more effective at increasing reports of perpetrations when ψ_{is} is larger, i.e., when $s = h$.

Similarly, Prediction 4 states that hard-garbling is more effective for victims who expect relatively more negative social judgment, i.e., those with higher values of δ_{is} .

For simplicity, I prove only Prediction 2a and show that hard-garbling is more effective when θ_{is} is higher (i.e., that increase in the incentive to report induced by hard-garbling is larger for less severe victimizations⁵⁶), though the proofs for Prediction 2b and Prediction 4 are analogous.

Taking the derivative of Equation B.10 with respect to θ_{is} yields the following:

$$\begin{aligned}\frac{\partial \Delta N B_{is}}{\partial \theta_{is}} &= p\phi_i \left[(\pi - 1) \times q \times \frac{\partial K}{\partial \theta_{is}} + \frac{\partial K}{\partial \theta_{is}} \right] \\ &= p\phi_i \times \frac{\partial K}{\partial \theta_{is}} \left[(\pi - 1) \times q + 1 \right]\end{aligned}\tag{B.13}$$

Thus, assuming $0 < \pi < 1$, then $-1 < \pi - 1 < 0$. Then, because $q < 1$, this implies that $(\pi - 1) \times q + 1 > 0$, so the bracket is positive. Since $\phi > 0$, and K is strictly increasing and positive in θ_{is} , the entire expression is positive. Therefore, given $p > 0$, Equation B.13 demonstrates that the net benefit of hard-garbling relative to direct elicitation is increasing in θ_{is} . Hard-garbling is more effective in inducing reporting when θ_{is} is larger, i.e., for less severe victimizations.

⁵⁶Note that less severe victimizations have higher values of θ_{is} by assumption.

C Social Norm Misperception Mechanisms

To further investigate how individual mis-perceptions of the social norms relate to background characteristics, I also use the following specification:

$$Y_i = \beta_0 + \beta_1 Female_i + \beta_2 MaleGuess_i + \beta_3 Female_i \times MaleGuess_i + \delta X_i + \epsilon_i, \quad (C.14)$$

where Y_i is the outcome of interest (misperceptions of victim-blaming or lying, i.e., the “wedge” between the respondent’s guess about agreement and the actual level of agreement with that statement), $Female_i$ is an indicator for the respondent i being female, $MaleGuess_i$ is an indicator for whether the outcome Y_i is the respondent i ’s misperception of men’s agreement with the statement of interest (rather than their misperception of women’s agreement), and X_i is a vector of controls, including indicators for whether a respondent was shown the victim-blaming or lying statement first, whether they were asked about agreement among men or women first, and individual characteristics. Then β_1 estimates the difference between women and men’s misperception about women’s agreement; β_2 estimates the difference between men’s misperceptions of women’s agreement and men’s misperceptions of men’s agreement; and β_3 estimates how much more women over-perceive men’s agreement, compared to their over-perception of women’s agreement, relative to the same gap for men.

Table C1 reports the results. Firstly, men over-perceive women’s agreement with the victim-blaming statement by 12.55 percentage points. Thus, men guessed on average that 12.55 percentage points more women agreed with the victim-blaming statement than actually did. Women further over-estimated women’s agreement, guessing on average 5.14 percentage points higher than men, implying their average guess was approximately 17 percentage points higher than the truth. In addition, men over-perceive other men’s agreement with the victim-blaming statement 11.6 percentage points more than they over-perceive women’s agreement. Thus, men’s average guess about other men’s agreement was approximately 24 percentage points higher than the truth on average. Finally, women had an additional 4.15 percentage point misperceptions of men’s agreement, relative to their misperception of women’s agreement.

Importantly, there is little misperception on aggregate for the lying statement: men’s over-perception of women’s agreement with the lying statement is only .18 percentage points, and men are also fairly accurate regarding men’s agreement (a total misperception of $.18 + 0.741$ percentage points). Women’s misperception of women’s agreement is 4.3 percentage points larger than men’s misperception of women’s agreement, implying the total misperception by women regarding women is approximately 4.5 percentage points. Women misperceive men’s agreement by a total of approximately 7 percentage points more than men do ($4.33 + 2.48$).

Other characteristics are also strongly correlated with (mis)perceptions of the social norms. For example, an individual’s own agreement with the statement is correlated with higher levels of misperception of the norm. Other traits, like involvement in intercollegiate athletics, attending an undergraduate institution in the South, and religious affiliation are also correlated with misperceptions.

Closely following Bursztyn et al. (2023), I also descriptively explore how three mechanisms may be impacting misperceptions of these norms. Thus, I consider three forces: false consensus, minority overweighting, and gender stereotyping.

False Consensus. False consensus is defined as a positive correlation between an individual’s opinion and their expected prevalence of this opinion among other people (Bursztyn et al., 2023; Fields and Schuman, 1976). To explore this force, I run a regression in which the dependent variable is the misperception of each norm (pooled by gender), and include indicators for individual agreement with the statement, gender, and current undergraduate student status. Table C2 displays these results. On average, individuals who disagreed with the victim-blaming statement overestimated the percentage of others who agreed with the statement by nearly 20 percentage points. Individuals who disagreed with the lying statement had fairly accurate perceptions of the lying norm, underestimating the percentage of others who agreed with the statement by only 1.5 percentage points. For each statement, individuals who agreed overestimated agreement by others by approximately 12 percentage points more than those who disagreed.

Minority Overweighting. Minority overweighting is the extent to which individuals over-

Table C1: Misperceptions of Norms: Individual Characteristics

VARIABLES	(1) Victim- Blaming	(2) Lying
Female	5.148*** (1.094)	4.337*** (1.102)
Guess About Men	11.612*** (1.076)	0.741 (1.098)
Female x Guess About Men	4.152*** (1.557)	2.475 (1.528)
Agreed w/ VB	11.730*** (1.315)	0.998 (1.333)
Agreed w/ Lying	-0.268 (0.903)	10.802*** (0.875)
Southern Institution	-0.043 (0.813)	2.534*** (0.803)
Democrat	-0.119 (0.870)	-0.698 (0.855)
Athlete	3.803*** (1.014)	3.101*** (0.990)
Greek	1.745 (1.068)	0.781 (1.027)
Non-Religious	-1.951** (0.843)	-2.522*** (0.839)
Current Undergrad	-1.833** (0.785)	-0.344 (0.773)
White	-0.310 (0.802)	-0.430 (0.783)
Observations	3,242	3,242
Control Mean	12.55	0.18

Notes: OLS regressions on misperceptions of social norms, created by subtracting the “truth” (actual level of agreement) from each respondent’s guess about that gender’s agreement. In the social norm task, respondents were randomly assigned to see either the VB or lying statement first and to answer about either men or women’s answers first. Regressions include additional controls for both levels of randomization. “Female” is a binary indicator equal to 1 if the respondent was female, and is zero otherwise. “Guess About Men” is a binary indicator equal to 1 if the outcome was the respondent’s prediction of men’s agreement, and is zero if the outcome is the respondent’s prediction of women’s agreement. “Female x Guess About Men” is an interaction term.

Table C2: False Consensus

	Victim-Blaming (1)	Lying (2)
Statement Agreement	12.596*** (1.054)	12.016*** (0.664)
Female	7.371*** (0.642)	5.867*** (0.639)
Current Undergrad	-2.182*** (0.638)	-0.544 (0.634)
Observations	3,242	3,242
"Disagree" Mean	20.197	-1.414

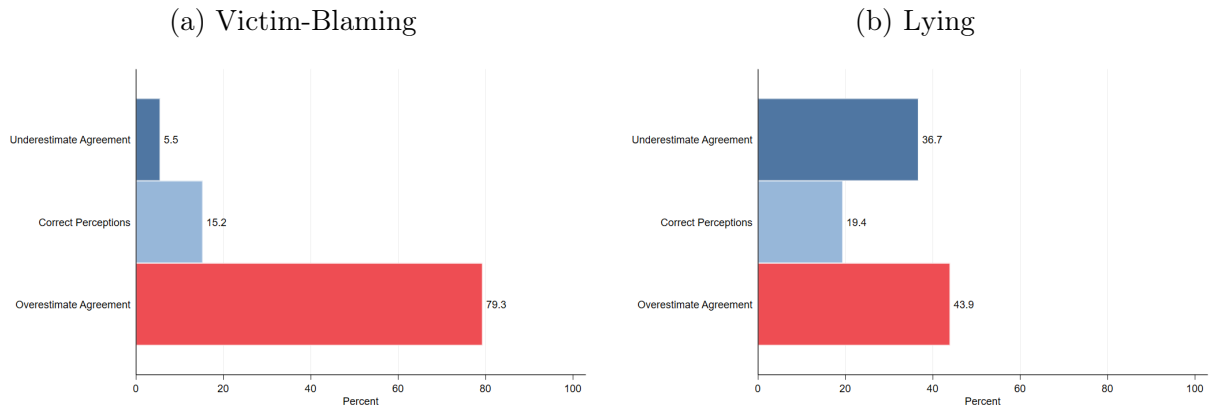
Notes: OLS regressions. The dependent variable in Column 1 is misperceptions of agreement with the victim-blaming statement (pooled across gender), and the dependent variable in Column 2 is misperceptions of agreement with the lying statement (pooled across gender). Misperception is defined as the perceived agreement with the statement minus actual agreement. "Statement Agreement" is equal to 1 if an individual agreed or strongly agreed with that social norm statement, and is 0 if else. "Female" and "Current Undergrad" are binary indicators equal to 1 if the respondent is female or a current undergraduate student, respectively.

estimate the share of the population that holds the minority view (Bursztyn et al., 2023; D. T. Miller and Prentice, 1994). In my context, the majority of respondents did not agree with the statements, implying that agreement with the statements is the minority view. To explore this force, I classify respondents into one of three mutually exclusive categories: 1) individuals with correct perceptions of agreement (within 5 percentage points); 2) individuals who overestimate actual agreement with the statement (i.e., individuals who overestimate the prevalence of the minority view); and 3) individuals who underestimate actual agreement with these statements (i.e., individuals who overestimate the prevalence of the majority view). By analyzing what percent of participants fall into these three categories, I am able to explore the relative extent of minority and majority overweighting.

Figure C1 provides strong evidence of minority overweighting for the victim-blaming statement. Nearly 80% of respondents overestimate agreement with the victim-blaming statement. However, for the lying norm, an approximately equal portion of the sample underestimates (36.7%) and overestimates agreement (43.9%) with the statement.

Gender Stereotyping. Gender stereotyping is the extent to which people overestimate gender differences in agreement with the statement (Bursztyn et al., 2023; Bordalo et al.,

Figure C1: Minority Overweighting

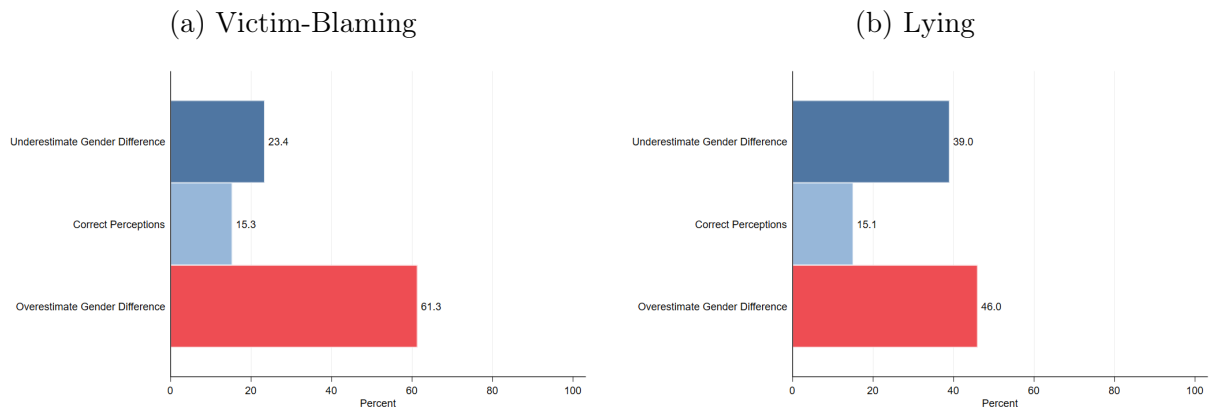


Notes: Minority overweighting is perceiving more agreement than actual agreement if a minority of the population actually agrees (or, perceiving less agreement than actual agreement if a minority of the population does not agree). The figure presents what percent of the sample underestimates agreement (majority overweights), has correct perceptions, or overestimates agreement (minority overweights) for both victim-blaming (graph A) and lying (graph B). Expectations of agreement and actual agreement was pooled by gender.

2016). I again classify respondents into one of three mutually exclusive categories: 1) individuals with the correct perceptions of the differences in agreement between men and women (within 5 percentage points); 2) individuals who overestimate the gender difference; and 3) individuals who underestimate the gender difference.

Figure C2 graphs the portion of respondents who fall into each category for the two norms. For victim-blaming, there is some evidence of gender stereotyping, as over 60% of the sample overestimates the difference in agreement between men and women. However, the lying norm is again relatively equal on this dimension: 39.0% underestimate and 46.0% overestimate the gender difference.

Figure C2: Gender Stereotyping



Notes: Gender stereotyping is defined as overestimating the gender difference in agreement. The figure presents what percent of the sample underestimates the gender difference, has correct perceptions, or overestimates the difference for both victim-blaming (graph A) and lying (graph B). Expectations of agreement and actual agreement was pooled by gender.