Perceptions of Workplace Sexual Harassment and Support for Policy Action: Pre-Analysis Plan

Sonia Bhalotra, Matthew Ridley [*]

September 2025

1 Analysis Plan

1.1 Data cleaning and integrity

We will drop from our analysis dataset any participants who do not pass all our attention checks, including attention checks near the end of the survey.

1.2 Descriptive analysis of benchmarks survey

Our benchmarks survey will produce estimates of several quantities which may be of independent interest, relating to the prevalence of sexual harassment, the harms from it, victims' reporting behaviour and perceptions of power imbalances that may affect reporting. We will report the estimates from our benchmarks survey, and compare them to existing estimates of prevalence, harms or reporting from the UK when comparable estimates exist.

[*] Note: this pre-analysis is also included as Section 5 of our Registered Report.

1.3 Descriptive analysis of beliefs

We will plot elicited beliefs using a histogram, in which we will mark the corresponding estimate from our benchmarks survey (which uses the same definition of sexual harassment and the same population) and the corresponding estimates from other studies, where these are available, as vertical lines.[1] We will report the median belief as well as the fraction of participants who over- and under-estimate each quantity relative to the estimate from our benchmarks survey.

To address concerns about potential biases due to participants' cognitive uncertainty or numeracy, we will also in robustness checks plot histograms separately for the subsamples of participants which:

- Report below-median average uncertainty in their prior beliefs
- Had answers to the questions about electric cars and happiness that were more accurate than median
- Saw a numerical anchor for their prevalence belief elicitation
- Have above-median numeracy

To provide further descriptive evidence on the patterns in over and underestimation of the sexual harassment problem, we will also regress beliefs on participant demographics as well as the variables measuring cultural background, prior harassment exposure, political, social and moral attitudes, and economic preferences described in our registered report.

To analyse participants' open-ended responses on the harms from sexual harassment, we will use large language models (LLMs) to code the types of harm mentioned and report their frequency. We will use human raters to develop an inductive coding scheme and validate the LLM output by coding a subset of responses (Haaland et al., 2024).

[1] Specifically, we will show on the corresponding graphs: estimated prevalence from the government survey report of Adams et al. (2020), the estimated proportion who would take a 10% pay cut to avoid sexual harassment from Folke and Rickne (2022), and the estimated proportions of victims who quit their jobs and have worse mental health from TUC (2016).

We will also analyse the extent to which beliefs predict outcomes, conditional and unconditional on observed characteristics.

1.4 Representativeness

We take steps to address the concern that our sample may be unrepresentative of the British public more broadly. A possible concern is that people with progressive views, particularly on gender, are more likely to take or complete our survey. To address this concern we use our data on political party support and on gender attitudes measured using questions from the UK Household Longitudinal Study (UKHLS). We will reweight our data to match the observed distributions of each variable as measured in, respectively, an average of recent UK opinion polls[2] and the most recent wave of the UKHLS.

In addition to this, we will test whether the averages of a standard battery of demographic dummy variables (gender, age over 45, Christian religion, university education, above-median income and parents being born in the UK) are significantly different in our sample to the overall UK population. If we find significant differences of more than 5 percentage points for any of these variables, we will include the distribution of that variable as an additional target to match when reweighting our data.

1.5 Attrition

Attrition can happen in one of three ways: firstly, participants may start the survey but not finish it; secondly, participants may fail to pass the later attention checks in our survey and therefore be dropped from our dataset (note that the early attention screeners that participants must pass to complete the survey come before treatment assignment), and thirdly, participants may complete the main survey but not return for the obfuscated follow-up. Our pilot data indicate that 97.8% of participants who start the survey on Prolific finish it successfully, and that, conditional on completing, 87% of participants pass all attention checks. Failure to finish or pass attention checks

[2] We intend to use Politico Europe's Poll of Polls for the UK Parliament voting intention.

was not differential by treatment condition. We found that 77% of participants who completed the main survey went on to complete the obfuscated follow-up survey but as there was no treatment in this pilot, we cannot report attrition at the follow-up stage by treatment status.

To test for differential attrition by treatment condition, we will regress an indicator for survey completion on indicators for treatment condition. If we find evidence of differential attrition we will use Lee bounds to bound the influence of this attrition on our estimated treatment effects (Lee, 2005).[3] We will also perform a robustness check in which we add back to the data all participants who completed the survey but failed an attention check.

We will regress attrition on demographic characteristics and gender attitudes to test for selection in terms of *who* completes the survey – for instance, whether people with more progressive gender views are more likely to complete. If so, we will correct for this by reweighting data by the inverse probability of survey completion given baseline characteristics (Little and Rubin, 2019). We discuss in section 1.6 how we will test for and address imbalances in characteristics between treatment and control groups (whether this is due to attrition or other factors).

1.6 Balance

We will test balance of control and treatment groups by regressing each demographic characteristic and gender attitudes on an indicator for each trial arm. If the standardized difference is greater than 0.25 then, following Imbens and Rubin (2015), we will adjust for all observable characteristics in the main analysis – this will account for any imbalances and increase efficiency. If we find substantial imbalance, we will use double-robust methods combining linear models with inverse probability weighting.

[3] As a robustness check we will also report Manski bounds (Manski, 1990), though we will not use these as our primary bounding estimates as they tend to be very conservative.

1.7 Regression Specifications: Treatment Effects

1.7.1 Average treatment effects

We will estimate overall treatment effects using the following specification:

$$Y_i = \alpha + \beta_1 PH_i + \beta_2 PH_i \times E_i + \eta' X_i + \epsilon_i$$

Where $PH_i = 1$ if i was assigned to see Prevalence and Harms information, and $E_i = 1$ if i was additionally assigned to see policy effectiveness information. $PH_i \times E_i$ is thus an indicator for assignment to the combined 'Prevalence, Harms and Effectiveness' treatment.

X_i is a vector of demographic controls, and Y_i denotes an outcome variable. The coefficients β_1 and β_2 give the treatment effect of each arm relative to the pure control group. Our hypotheses make the following predictions about the coefficients in this specification:

- Hypothesis 1 predicts that $\beta_1 \neq 0$
- Hypothesis 2 predicts that for our policy support outcomes, $\beta_2 > 0$ and $\beta_2 > \beta_1$

To test these predictions, we will use two-tailed t tests with a 5% significance.

1.7.2 Heterogeneous treatment effects by prior beliefs

Information treatments will tend to affect the relevant outcomes differently depending on the direction in which information moves participant beliefs.[4] We estimate heterogeneous treatment effects by whether the priors of the individual tend to over- or under-estimate the information provided. This provides a test of whether beliefs causally affect outcomes. We will use the following specification:

[4] Indeed, Coffman et al. (2025) note that average effects of information treatments are difficult to interpret because they depend on the beliefs of the marginal actor. Restricting to a subsample of over- (under-) estimators, as we do in this subsection, ensures that any marginal actors in the subsample are also over- (under-)estimators meaning that treatment effects can be interpreted as the effect of decreasing (increasing) beliefs.

$$Y_i = \alpha + \beta PH_i + \gamma Over_i + \delta PH_i \times Over_i + \eta' X_i + \epsilon_i \qquad (1)$$

We estimate this specification in the subsample of participants who can be regarded either as overall ‘overestimators’ or overall ‘underestimators’ of the information in our prevalence and harms treatment condition. In our main analysis, we define an over-(under-)estimator as one who over- (under-) estimates *both* our prevalence statistic *and* a majority (at least 3 out of 5) of our harms statistics.[5] $Over_i$ is an indicator variable for being an over- rather than an underestimator within this sample.

The sample for the above regression excludes those who receive the effectiveness information, as it is less clear how to define over- and underestimation in this case. At least some of the effectiveness information we provide is qualitative in nature, such as the observation that the existing employment tribunal process for victims to seek redress is complex and cumbersome, and (as explained above) direct quantitative estimates of the effectiveness of different policies is lacking.

We preregister robustness checks using different definitions of over and under-estimation – in particular, a narrower definition considering only those who either overestimate or underestimate both prevalence and all of our harms statistics, and a broader definition that includes participants who over-(under-) estimate both prevalence and at least two of the harms statistics.

Hypotheses 1 predicts that $\beta > 0$ and $\delta < 0$ (because of how we define our outcome variables, described in our registered report). We will again test these predictions using two-tailed t tests with a 5% significance level. We will report both point estimates and 95% confidence intervals for all coefficients in all regressions.

This exercise also enables us to verify that the effects of our information treatments are not just due to salience. If salience were at play, we would expect that individuals change their beliefs and their demand for SH policy (or related outcomes) in the same direction, irrespective of their initial beliefs. However, if information drives the results then participants whose initial beliefs were lower than the information

[5] This means that the sample excludes those who, for instance, underestimate prevalence but overestimate all harms information as they cannot clearly be classified as over- versus under-estimators.

provided will increase their beliefs and demand; while those who initially held beliefs higher than the value we present as "true" will decrease their beliefs and demand. This asymmetry holds under information but not under salience.

A potential confound when conducting this exercise is that over- and under-estimators may differ in other ways besides their prior beliefs that correlate with the response to treatment. To address this confound as far as possible, when estimating equation (1) we control for interactions between the treatment and other observables that may predict treatment effects. Starting from interactions between all potential control variables and an indicator for treatment, we will use a post-double-selection LASSO procedure to select interactions to control for (Belloni et al., 2014).

Investigating the relative importance of prior beliefs about prevalence versus harms. We will also provide suggestive evidence on whether beliefs about prevalence or harms appear to play a relatively more important role in driving our outcomes. To do this we will estimate a specification in which indicators for treatment are interacted separately with indicators for overestimating either prevalence or harms:

$$Y_i = \alpha + \beta PH_i + \gamma_1 OverP_i + \gamma_2 OverH_i + \delta_1 PH_i \times OverP_i + \delta_2 PH_i \times OverH_i + \eta' X_i + \epsilon_i$$

In this specification, $OverP_i$ equals one if i overestimated our prevalence statistic and $OverH_i$ equals one if i overestimated a majority of our harms statistics. δ_1 measures the difference in treatment effects between over- and under-estimators of prevalence (conditional on whether they overestimated harms) and δ_2 measures the difference in treatment effects between over- and under-estimators of harms (conditional on whether they overestimated prevalence). If δ_1 is larger in absolute value than δ_2 for a given outcome, we will interpret this as evidence that beliefs about prevalence appear to be more important in driving that outcome.

1.8 Outcome variables

We divide our outcome variables into three families:

Policy Support Outcomes

- The amount of the donation chosen to go to Rights of Women UK.
- The amount of donation chosen to go to the Survivors' Trust in our obfuscated follow-up survey.
- An Anderson Index (Anderson, 2008) which combines stated support for each of the policy changes described in our registered report and pre-registration with participants' ranking of sexual harassment relative to other policy issues. This includes the policy ranking and policy support outcomes from our obfuscated follow-up survey.
    - Respondents will be given a five-point answer scale for each policy ranging from 'strongly oppose' to 'strongly support'. We will encode these with the integers 1 through 5. We will combine these answers with participants' rankings of sexual harassment as a policy issue, coded so larger numbers are a higher rank, and construct an inverse-covariance-weighted index (Anderson, 2008).
- The number of petitions calling for our policy changes that respondents sign. For this outcome, we cannot include controls in our regression because petition signatures are anonymous (we only observe the total number of signatures by petition and treatment condition).
- The estimated proportion agreeing with the sexual harassment-related statement in our list experiment, by treatment condition. Again, for this outcome we cannot include individual-level controls.

Job choice outcomes

- An indicator for choosing the lower-paying, female-dominated job in our hypothetical job choice (for this outcome, the sample will include female respondents only).
- An indicator for recommending a hypothetical daughter/female family member to choose the lower-paying, female-dominated career path.

Reporting outcomes

- Stated willingness to report a sexual harassment incident.

In addition, we will examine treatment effects on the secondary outcomes described in our pre-registration and registered report.

1.9 First-stage effects

We will estimate first-stage effects on posterior beliefs using the above regression specifications. We will additionally estimate specifications with uncertainty in posterior beliefs as the outcome variable.

1.10 Robustness checks

Experimenter demand. We will assess the robustness of our results to experimenter demand effects in several ways. Firstly, we will estimate the above regressions using only the outcomes from our obfuscated follow-up survey (combining the policy support outcomes from this survey into an Anderson index as above). Secondly, we will estimate the above regressions among the sample who report in debriefing questions that they did not perceive experimenter demand during the survey. Thirdly, we will estimate the treatment effect of our demand manipulation treatments (De Quidt et al., 2018) in the control group and use these to create lower and upper bounds for our 'true' treatment effects by respectively subtracting and adding them to our estimated treatment effects.

Other data quality concerns. To check the robustness of our results to data quality issues that may be caused by participant inattention or response noise, we will perform the following subgroup analyses:

- The above regressions among the subsample who remembered all key statistics from their treatment information. This uses the fact that the control group is also shown the treatment information and given memory questions on it at the very end of the study.
- The above regressions, dropping from the sample those whose average subjective uncertainty in their priors was above the 90th percentile.

- The above regressions, dropping from the sample those whose average survey completion time was below the 10th percentile.
- The above regressions, dropping from the sample those whose average survey completion time was above the 90th percentile.
- The above regressions including only the intersection of the above four samples.

1.11 Heterogeneity analyses

We will perform and pre-register additional subgroup analyses as follows:

- The above regressions among the subsamples whose answers to the questions on electric cars and happiness were (on average) more versus less accurate than the median.
- The above regressions among the subsamples whose answers to the questions on domestic violence and crime victimization were (on average) more versus less accurate than the median.
- The above regressions among the subsamples in which participants were versus were not randomly assigned to a numerical anchor for their prevalence belief elicitation.
- The above regressions among the subsample of participants who remember at least one piece of information accurately.
- The above regressions among the subsamples with above and below median social desirability bias according to the Marlowe-Crowne scale.
- Analysis allowing for heterogeneous treatment effects by the following dimensions:
    - Whether participants have experienced sexual harassment themselves
    - An index of masculinity
    - An index of traditional gender norms

- An index of perceived credibility of the information (derived from participants' answers to questions about whether women over-report due to being 'sensitive' or under-report due to harassment being normalized)
- Gender, age, religion, education, political affiliation, income, immigration status (parents born in the UK or not), and whether participants have a daughter

In addition to the pre-specified heterogeneity discussed here, we will use causal random forest to identify variables that are relatively important in determining heterogeneous treatment effects.

1.12 Multiple hypothesis testing

To address concerns related to multiple hypothesis testing, we will utilize the Romano-Wolf procedure to adjust p-values within each family of outcome variables (Romano and Wolf, 2005).

For each of the subgroup or heterogeneity analyses above, we will similarly adjust p-values within outcome families when restricting to a particular subsample or testing the equality of effects across subsamples.

1.13 Control variables

In all regressions, we will use the post-double-selection Lasso method of Belloni et al. (2014) to select control variables from all available controls (including the heterogeneity and mechanism variables described above as well as demographics).

References

Adams, L., L. Hilger, E. Moselen, T. Basi, O. Gooding, and J. Hull (2020). 2020 sexual harassment survey. Working Paper, UK Government Equalities Office.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association 103*(484), 1481–1495.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Coffman, L. C., C. R. Featherstone, and J. B. Kessler (2025). A model of information nudges.

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

Folke, O. and J. Rickne (2022). Sexual harassment and gender inequality in the labor market. *The Quarterly Journal of Economics 137*(4), 2163–2212.

Haaland, I., C. Roth, S. Stantcheva, and J. Wohlfart (2024). Understanding economic behavior using open-ended survey data.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press.

Lee, D. S. (2005). Training, wages, and sample selection: Estimating sharp bounds on treatment effects.

Little, R. J. and D. B. Rubin (2019). *Statistical analysis with missing data.* John Wiley & Sons.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Romano, J. P. and M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

TUC (2016). Still just a bit of banter?