Appendix H. Extended methods for design-based supervised learning implementation.

Analyses	Implemented Approach	Description
Characteristic; contingency tables	Using the DSL R package by Egami et al. (2025)	This uses the DSL package, regressing each dummy-encoded interrogative category i on each demographic feature d using linear regression without an intercept. This approach is mathematically equivalent to computing subgroup means. Although it is less interpretable, this was necessary because the DSL estimator cannot simply be applied to aggregated count data. As this results in category proportions by demographic group, these are less interpretable than counts in terms of the estimands of interest; e.g., P(interrogative type — demographic characteristic) requires bayes rule. Therefore, the main report includes traditional contingency tables with counts, but the cells of plots are greyed out when the analysis with the DSL package indicated relatively high uncertainty according to the criteria described.
Association; over-representation factors	Manual implementation of the DSL	To manually implement the DSL, I created design-adjusted outcomes for each dummy-encoded category, and used these adjusted outcomes to compute over-representation factors (using bootstraps to get uncertainty estimates). This formula stems from Egami et al. (2023): For observations that are expert-labelled: $$\tilde{Y} = \hat{Y} - \frac{\hat{Y} - Y}{\pi},$$ where $\hat{Y}$ = the label from the LLM, Y = expert label π = proportion of observations that are expert-labelled. For observations that are not expert-labelled: $$\tilde{Y} = \hat{Y},$$ where $\hat{Y}$ = LLM label This leads to conservative uncertainty estimates, as shown in the magnitude of the standard errors for some of the standard errors in Appendix J. A supervised learning model (g) to improve the LLM predictions was not implemented for simplicity and because this is not strictly necessary, particularly if the LLM labels already have high accuracy (as they do in this case).
Conditional association; logistic regressions	Not implemented	The current version of the DSL R package (0.1.0; 12.08.2025) does not support clustered standard errors when specifying the 'logit' model. To address this limitation, I attempted to implement a linear approximation using the felm model. Please see . . ./02_code/07_descriptive_log_regs.Rmd for a detailed description of the errors encountered, and the steps attempted to address them. I then considered using the design-adjusted columns from my previous manual implementation, and while these work for descriptive analyses, in the case of the current logistic regressions, the formula by definition introduces numbers that are not 0 or 1, which makes the logit model break. I was not able to find an indication of how to address this in the relevant DSL papers. For these reasons, I am not implementing the DSL for this step of my descriptive analyses.

Note. DSL = design-based supervised learning. All methods stem from Egami et al. (2023) and Egami et al. (2024). The DSL R package was developed by Egami et al., (2025).