Improving the Catalan Citizen Panel through Adaptive Survey Design

Jordi Muñoz[1] Joel Ardiaca[2] Raül Tormos[3]

[1]Universitat de Barcelona

[2]Universitat de Barcelona and Centre d'Estudis d'Opinió

[3]Centre d'Estudis d'Opinió

11/07/2024

Overview

In this presentation we will...

1. Briefly introduce the Catalan Citizen Panel
2. Provide evidence on the response rate and nonresponse biases of the Panel
3. Discuss the plan for the adaptive survey design
4. Present the recruitment randomized experiments and their results
5. Present the models that predict (non)response under each treatment condition
6. Discuss the optimization strategy and present the first results based on simulations
7. Discuss further steps

The Catalan Citizen Panel

In 2023 we decided to move all the Center's surveys (except the f2f Barometer) from telephone and face-to-face to the Catalan Citizen Panel, based on a **push-to-web+paper** mode of administration. The decision was partly motivated by the challenges of the existing modes of administration:

1. Declining response rates for telephone-based surveys (5-7%) and face-to-face (11-13%)
2. Increasing costs and quality concerns

The Catalan Citizen Panel

The decision was also based on some opportunities that the context offered us.

1. Generalization of the use of smartphones that allows for wide coverage of online surveys
2. Access to the population register allows for a true random sampling
3. The cost per survey is lower, so we can increase our sample size and number of studies
4. Possibility to run a longitudinal study and accommodate more and more complex survey experiments

Sampling and administration

The sampling strategy and mode of administration of the Catalan Citizen Panel are as follows:

1. We obtain a random sample of individuals from the population register (all residents, including nonnationals, over 16) from the Statistics Institute. The sample includes information on address, age group, gender, place of birth and citizenship.
2. We send invitation letters in which we ask selected individuals to answer a survey with a link to our landing page. Eventually, we follow up with (1 or 2) reminders and/or phone calls/email/SMS.
3. Some respondents receive the paper questionnaire (either in the first letter or in the reminders).
4. After answering a survey, respondents are invited to join the panel.

The Catalan Citizen Panel

The Panel is conceived as a research infrastructure that allows us to conduct several cross-sectional and longitudinal studies.

	1	2	3	4	5
Date	12/2022	06/2023	10/2023	12/2023	04/2024
Fresh sample	40,000	30,000	19,964	-	12,000
Panel members	-	7,705	13,382	16,939	17,010
Responses	11,037	11,548	11,930	6,813	9,425
Response rate	27.6%	26.1%	28.4%	-	25.9%
RR panelists	-	48.1%	47.3%	40.2%	38.4%
Paper	22%	23.3%	18%	20.6%	19.1%
Recruitment	18.9%	19%	19.2%	-	17.6%

Response rate, by sex

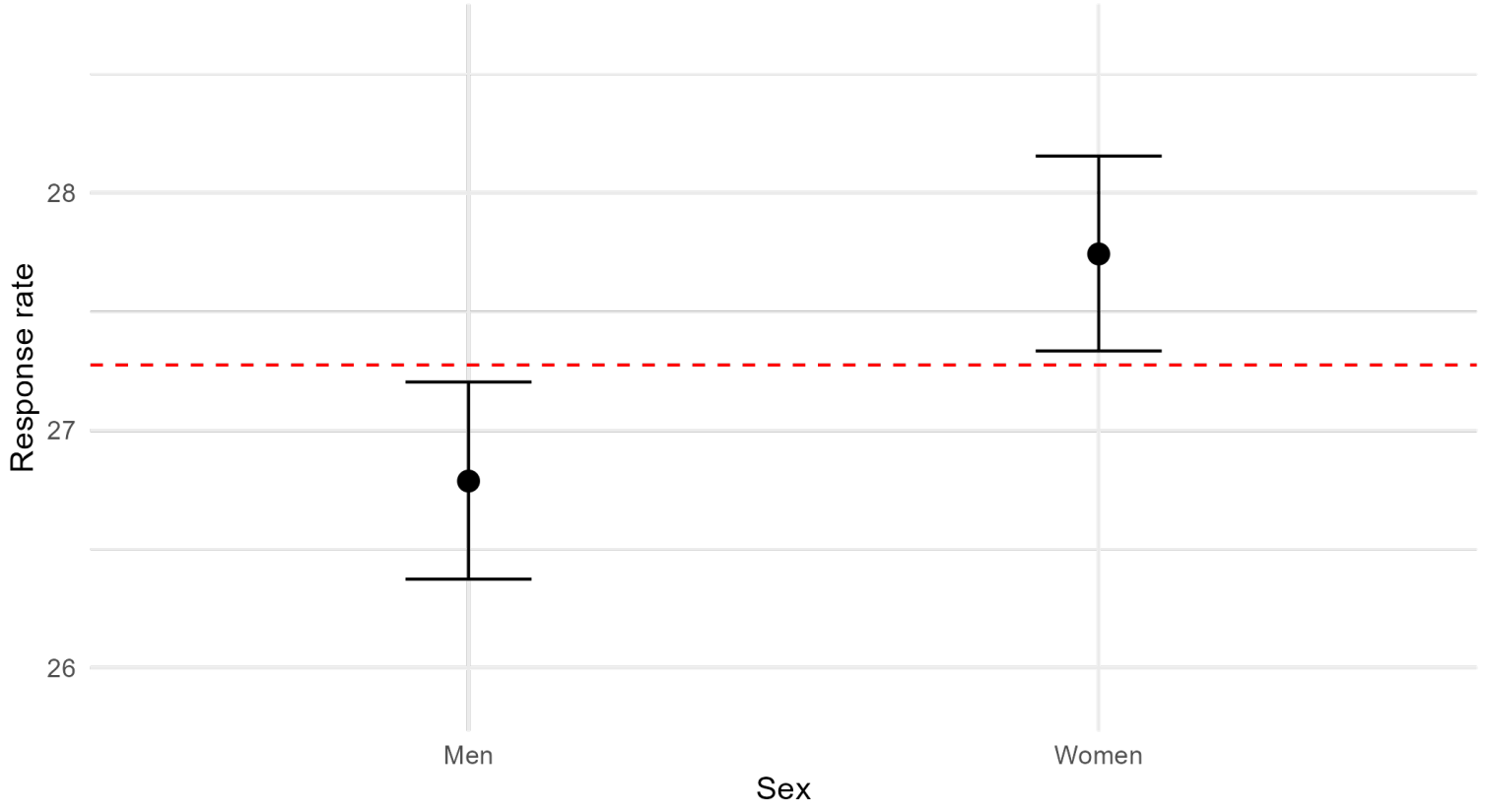


Figure 1: Response rate by sex, across first three waves for new respondents, with 95% confidence intervals

Response rate, by age

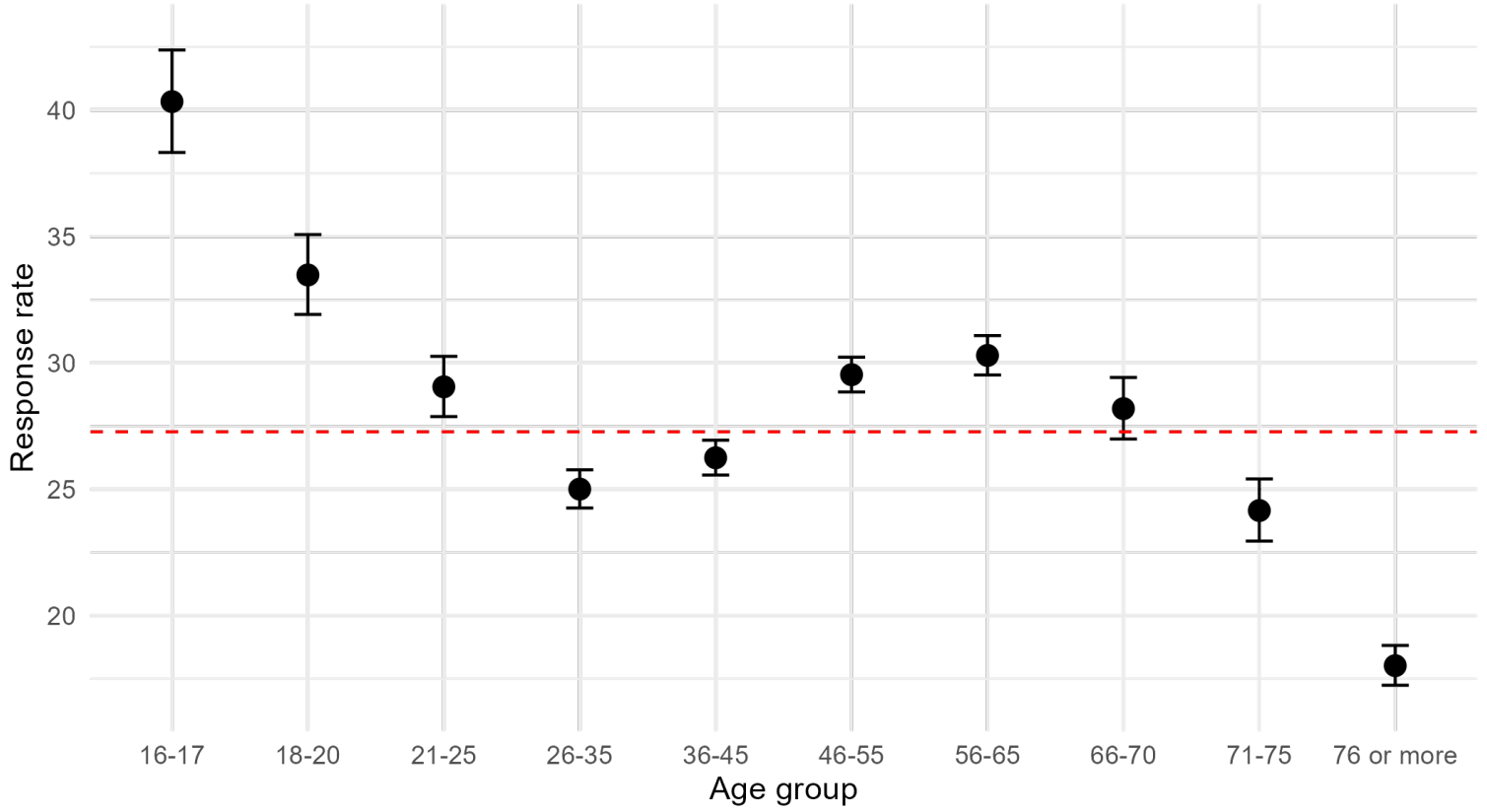


Figure 2: Response rate by age group, across first three waves for new respondents, with 95% confidence intervals

Response rate, by place of birth

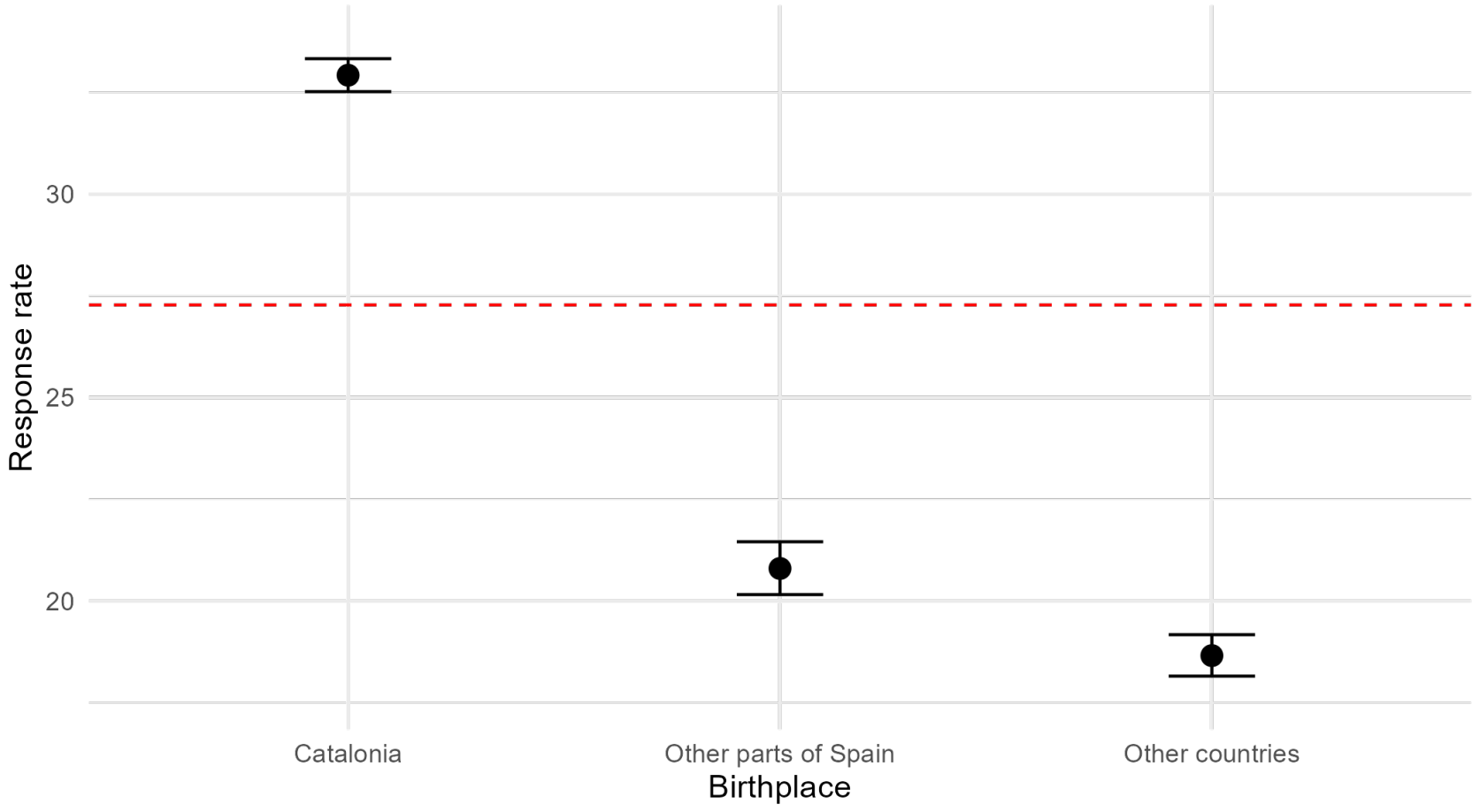


Figure 3: Response rate by birthplace, across first three waves for new respondents, with 95% confidence intervals

Response rate, by income of census tract

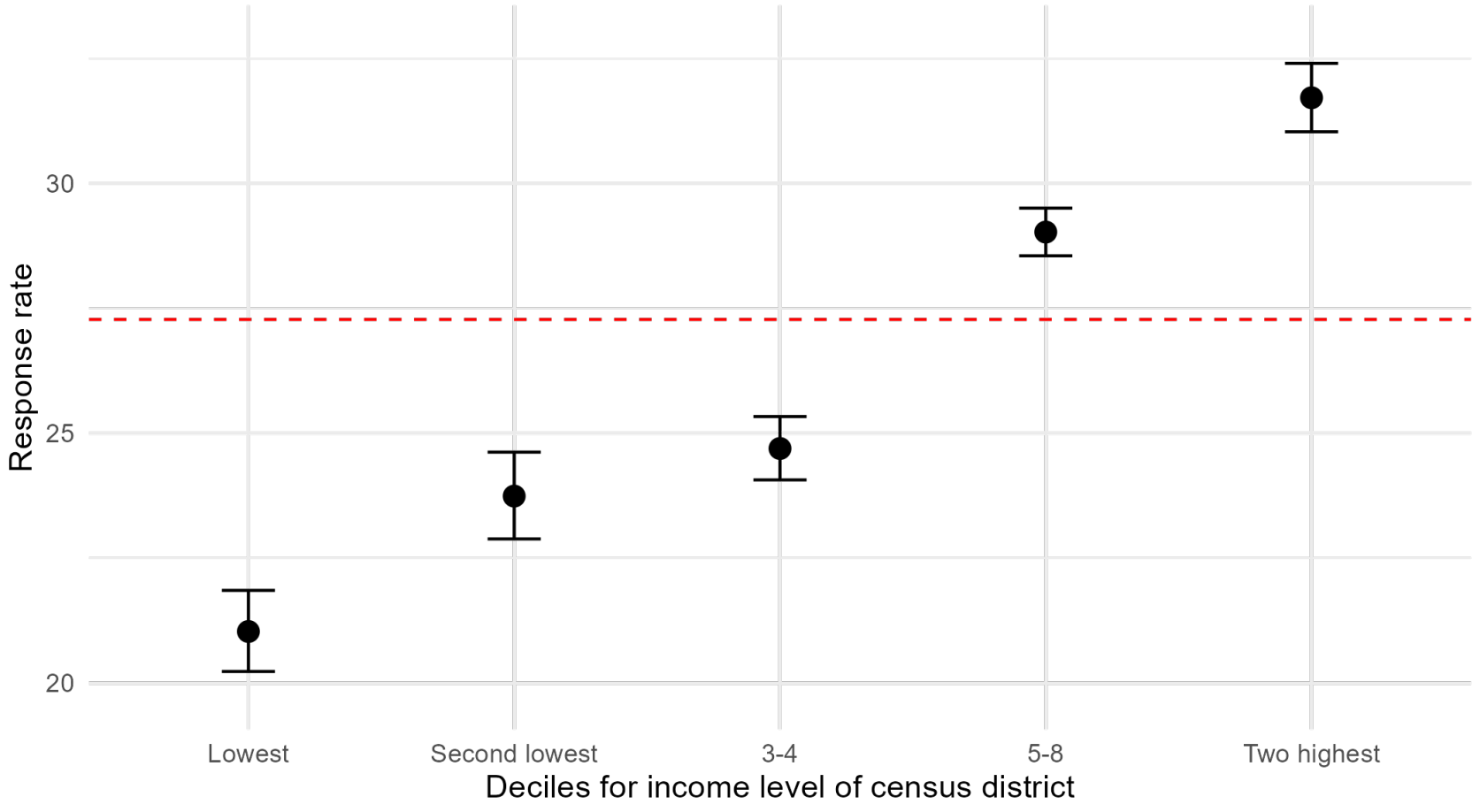


Figure 4: Response rate by income level of census tract, across first three waves for new respondents, with 95% confidence intervals

Addressing non-response bias

Despite the comparatively high response rate, we face challenges related to non-response bias.

- We address them mostly through **weighting**. We use large samples to minimize sampling error despite the design effect introduced by our weighting schemes. All our weighting procedures are publicly available on our github repository
- We deliberately avoid the strategy of oversampling underrepresented areas or individuals, because it may worsen the biases on unobservables (and should be corrected with design weights).
- We are working on the incorporation of the **adaptive survey design**: Optimize the allocation of resources/recruitment efforts to maximize **representativity** (over response rate), subject to the logistic, economic, legal and ethical constraints.

Adaptive Survey Design

ASD (Wagner, 2008) is based on the idea that the optimal survey protocol (contact, incentives, mode) to maximize response probability may not be the same for each individual in the sample.

- Protocols are then tailored to potential respondents in order to **optimize aggregate survey outcomes (response rates and/or sample balance), subject to budget restrictions** (Beaumont et al. 2014, Särndal 2011, Schouten et al. 2009)
- Responsive designs (Groves & Heeringa, 2006) include a dynamic component in which protocols are adjusted for each sampled unit during the fieldwork.
- Static adaptive designs (Bethlehem, Cobben, and Schouten 2011), also called **targeted designs** only vary between, and not within units (Lynn 2017)

Adaptive Survey Design

The ASD can be expressed as an optimization problem with the following elements:

Objective function

In order to obtain the most representative sample, we aim to maximize the R indicator: the inverse of the standard deviation of the likelihood of responding

$$max \quad R(\rho) = 1 - 2S(\rho)$$

Data

$i = \{1, 2, ..., n\}$, where n =number of individuals in the sample

$j = \{1, 2, ..., k\}$, where k = number of possible treatments

P_{ij} = Response probability of individual i under treatment j

Adaptive Survey Design

$$X_{ij} = \begin{cases} 1, & \text{if individual } i \text{ is assigned treatment } j \\ 0, & \text{otherwise} \end{cases}$$

We have a set of **restrictions**. Mainly:

1. Each individual can be assigned to one, and only one treatment

$$\sum_{j=1}^{k} X_{ij} = 1 \quad \forall_i \in \{1, 2, ...n\}$$

2. We have a budgetary restriction B that limits the number and type of treatments based on their cost c_j

$$\sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij} \cdot c_j \leq B$$

Adaptive Survey Design

Hence, we need to:

1. Collect information about the response probability of each type of individual under each treatment, based on the random assignment of sampled units to treatments in the first waves
2. Model the response probability based on this information, as well as the individual information available in the sample, to produce a predicted probability of responding for each individual under each treatment
3. Solve the optimization problem based on this information to optimally assign treatments to sampled units

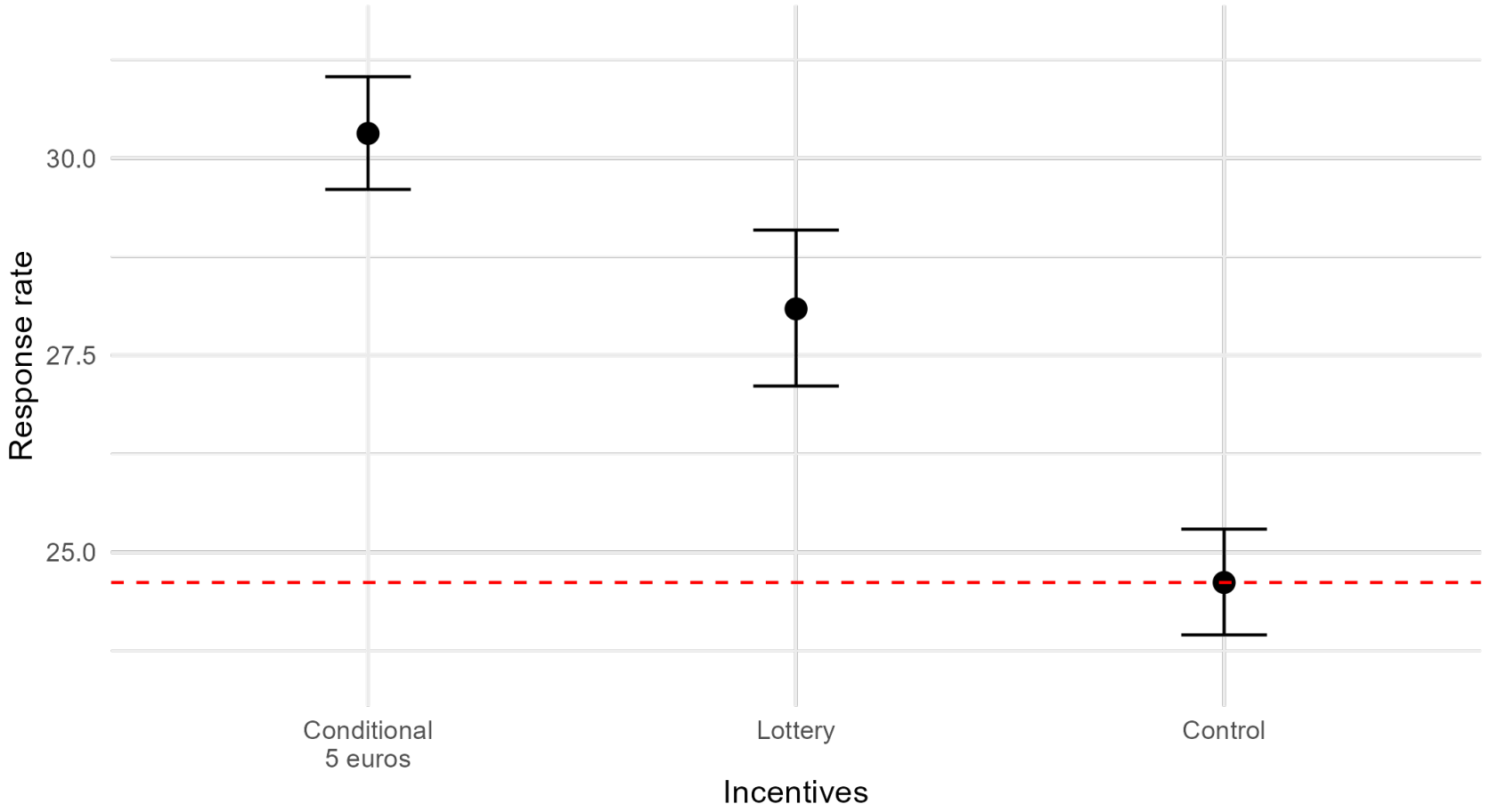
Recruitment Experiments

In order to obtain the necessary information for the ASD, during the first waves of the Panel we have run a series of recruitment experiments:

1. Experiments with incentives: randomizing the type and amount of incentives offered
2. Experiments with reminders: randomly varying the number and type of reminders
3. Other, less intensive experiments: letter design, lottery prize, targeting and framing the letter...

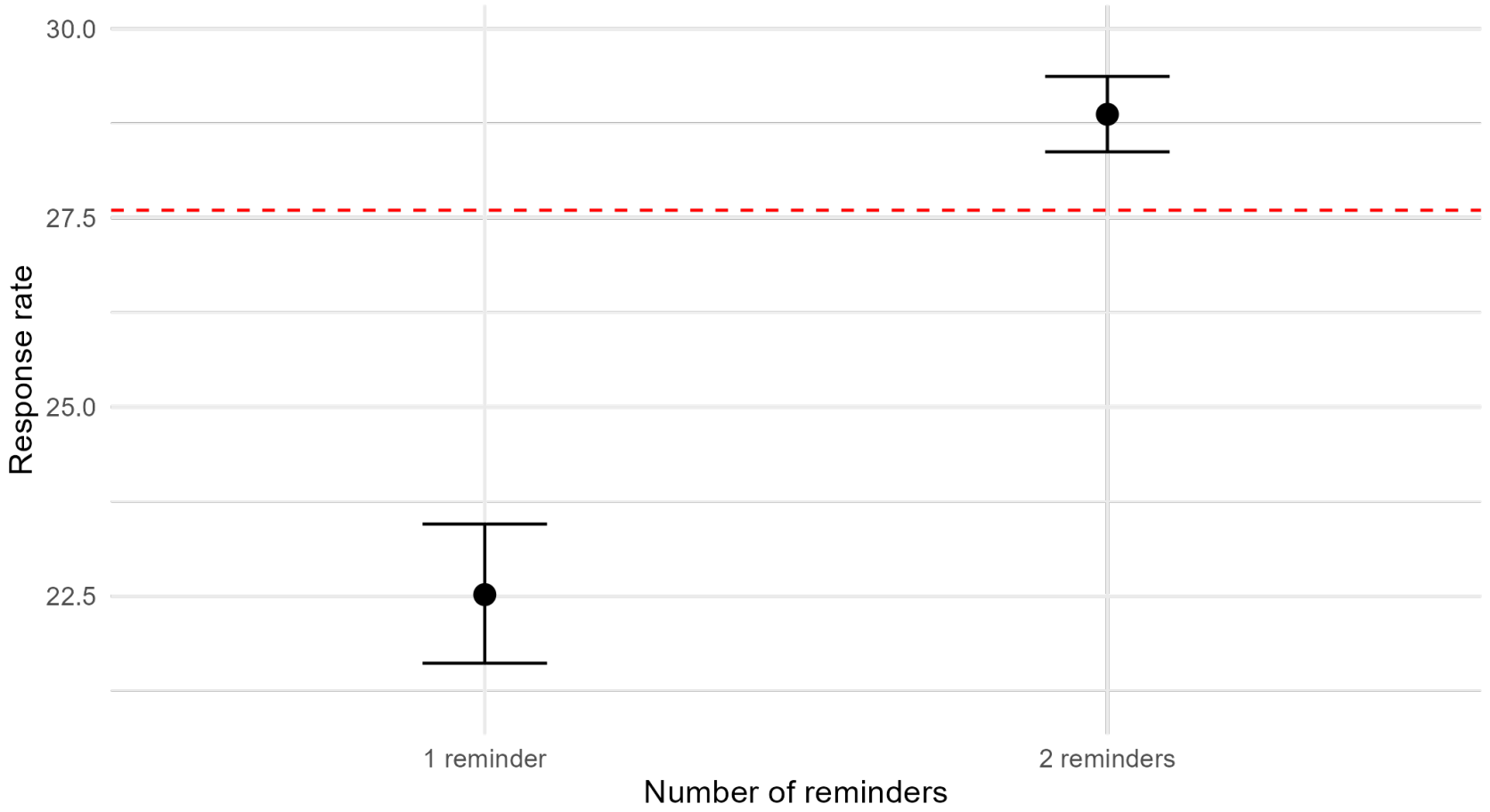
Experiment with incentives

In the first wave we randomized the incentives offered to survey respondents.



Experiments with reminders

In wave 0 and 2, we run an experiment with the number of reminders sent.



Modelling response

- Model based on the data of first three waves (N $\approx 90,000$).
- Logistic regression with the following variables:
  - Individual: sex, age, nationality and birthplace
  - Census tract: income decile, sociopolitical cluster, turnout...
  - Treatment: number of reminders & incentives

Modelling response

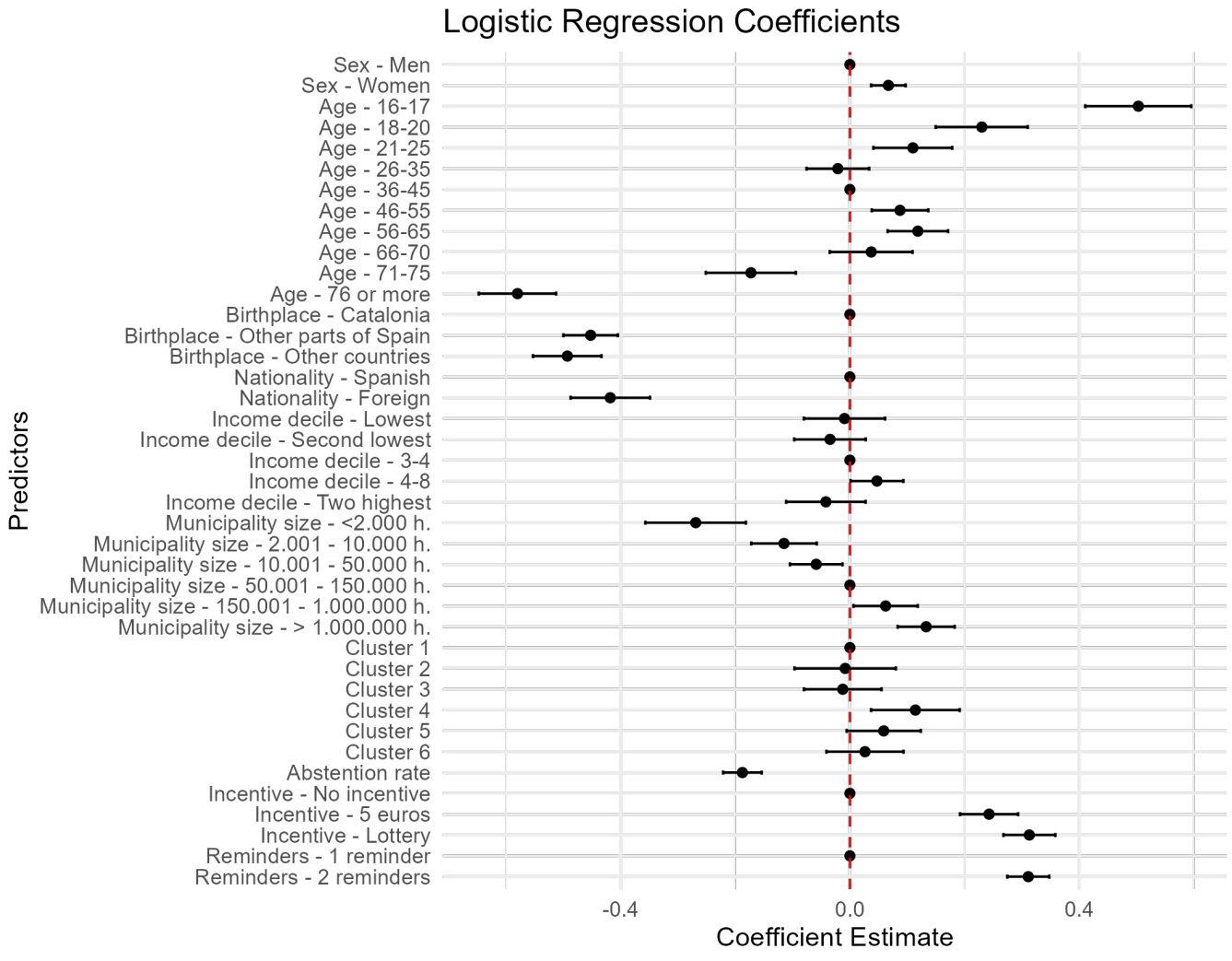


Figure 7: Model coefficients

Evaluation of the model (simulation)

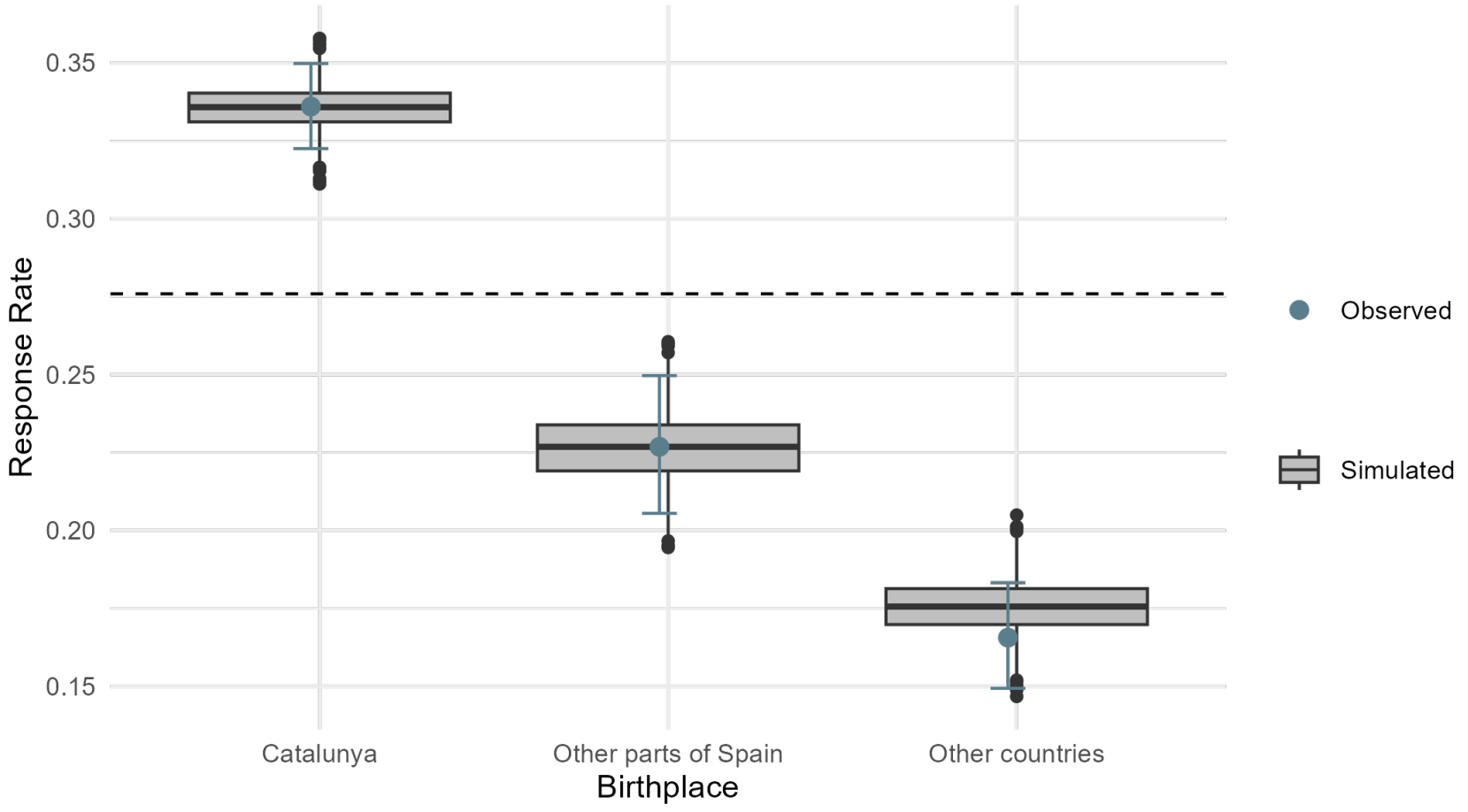


Figure 8: Observed and simulated response rate for model evaluation, by place of birth

Evaluation of the model (simulation)

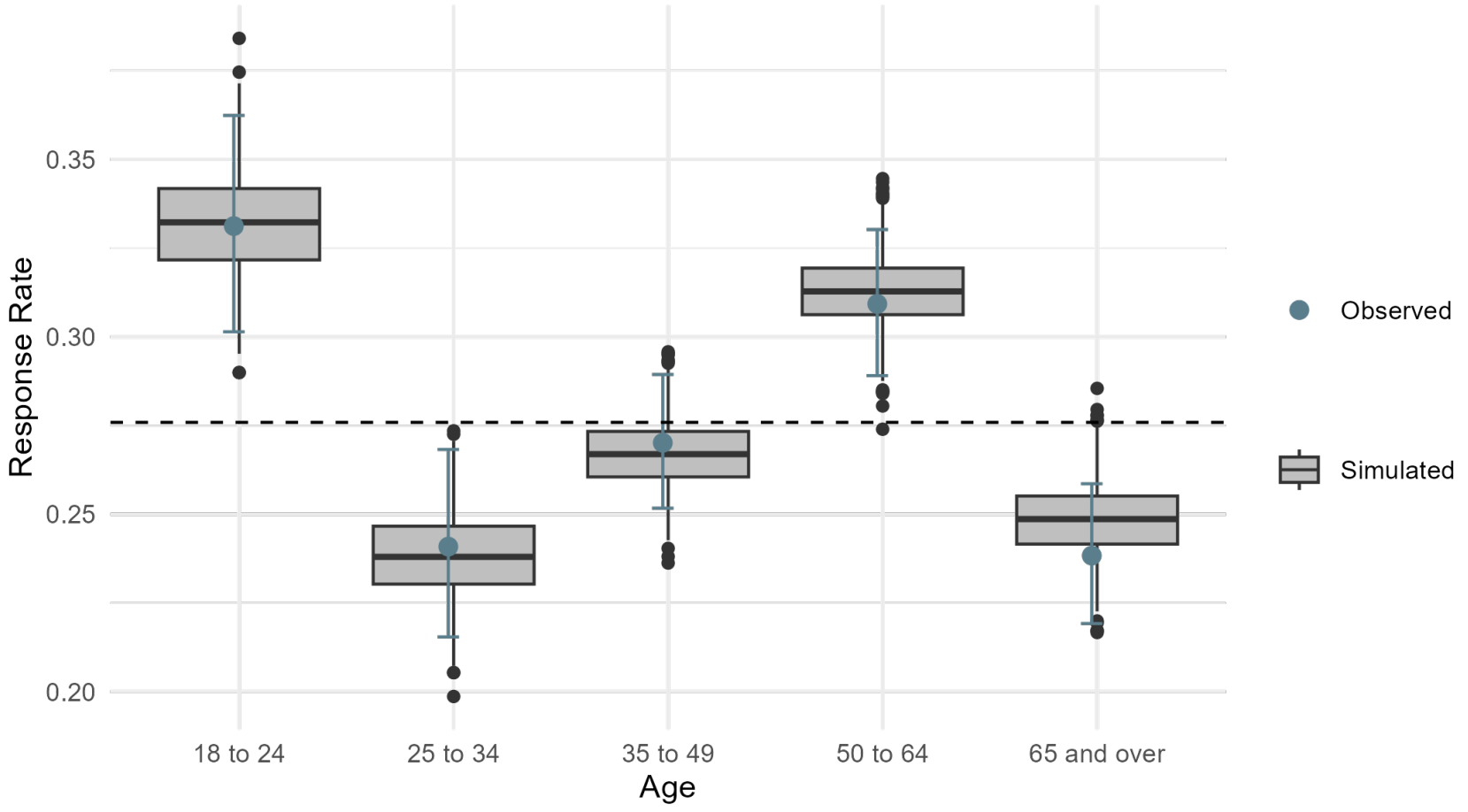


Figure 9: Observed and simulated response rate for model evaluation, by age

Optimization: Proof of concept

As a proof of concept, we run the optimization model for 8,000 individuals and 4 treatments:

1. 5€ conditional incentive
2. 5€ conditional incentive in 2nd letter if nonresponse to the 1st letter
3. Lottery
4. No incentive

Optimization

The optimization problem rapidly escalates: with 8,000 individuals and 4 treatments, the amount of combinations is of the order 10^{4800} , so we need to simplify the problem. There are several possible strategies:

1. [Simple] Sort the incentives according to the degree of intensity, and give the most intensive ones to the people least likely to respond, based only on their probability in the control group. Doing so ignores the heterogeneous effects of incentives
2. [Complex] Find formulas to simplify the optimization problem. Some options (which can be combined)
    1. Randomly separate the problem into pieces, solve them mathematically and then combine them
    2. Rely on heuristics as decision rules

Simple solution: simulation of results

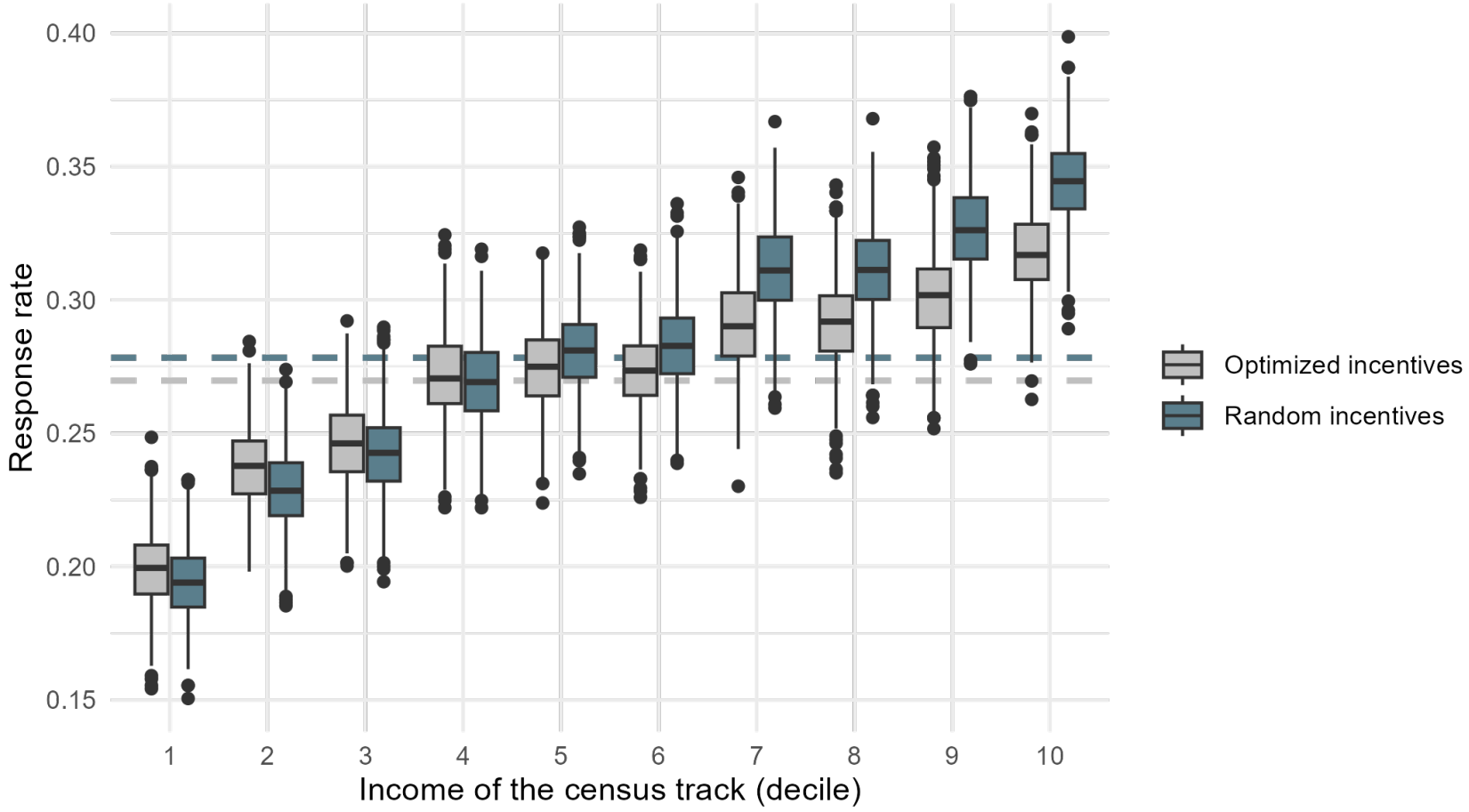


Figure 10: Simulated response rate by income level of census district.

Simple solution: Evaluation

- This first simple optimization, which does not take into account the heterogeneous effects of the incentives, achieves a certain improvement in the representativeness of the sample (at the cost of a slightly lower response rate).
- Not an ideal solution, since it improves R mostly by lowering the response rate of the groups with higher likelihood of responding, and not so much increasing the probability of those groups that respond less.
- This is due to the fact that those groups with lower response rate are also less responsive to incentives and reminders.

More complex solution 2: heuristics

Given that optimizing the problem exactly (although in a simplified version) is complex, we also evaluate another option based on the use of heuristics or simple rules to allocate incentives.

- People with a $P_i > 27.5\%$ (without incentives), do not receive incentives
- People for whom the incentives have no effect $\Delta P_i < 1\%$ for $j = 1, 2, 3$ are assigned to the control group
- If the lottery works better than the 5€ (by more than 2%), they are assigned to the lottery
- The rest are split by P_i: those below the median receive the 5€ in the first letter and the rest in the reminder.

Complex solution 2: simulation of results

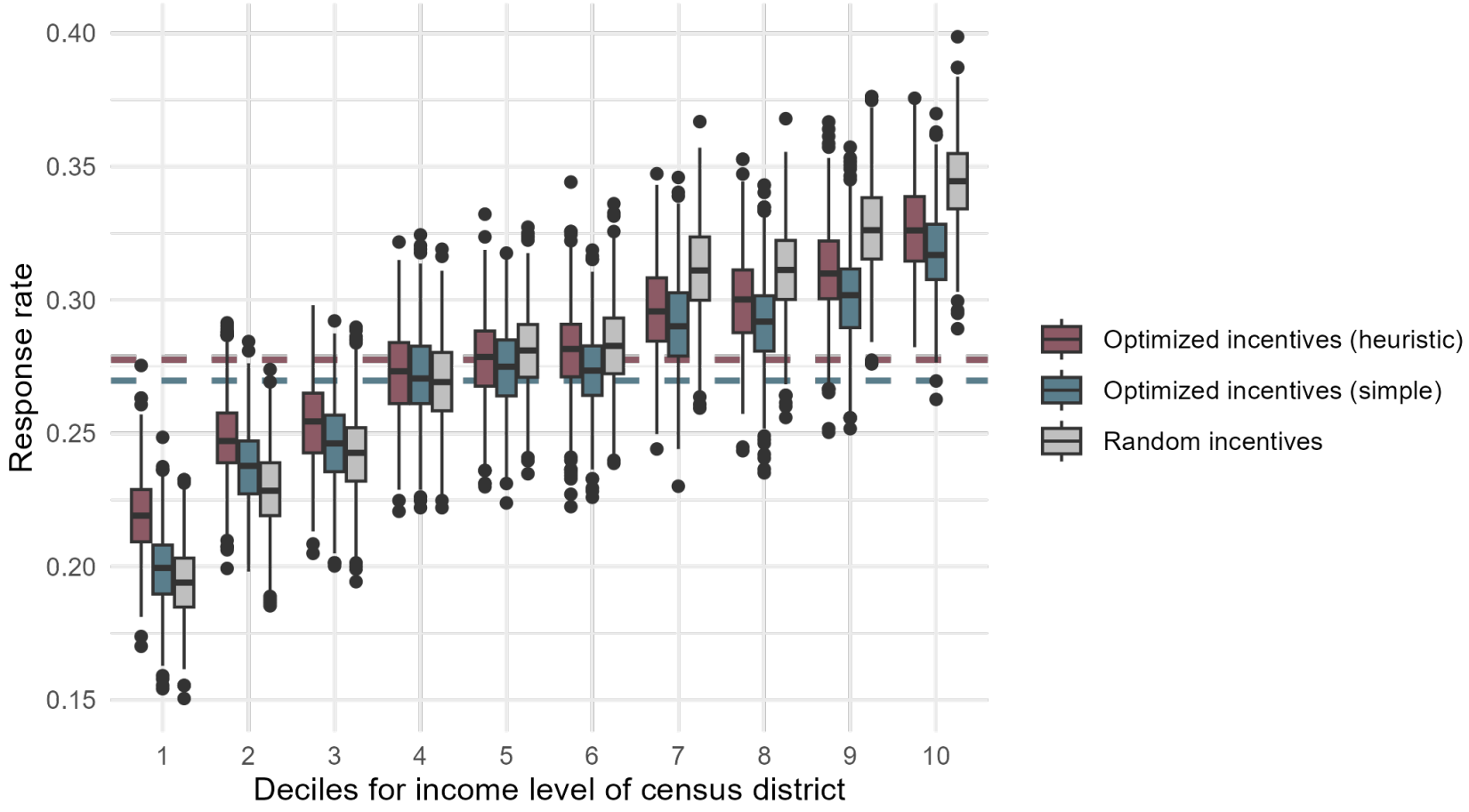


Figure 11: Simulated response rate by income level of census district

Complex solution 2: simulation of results

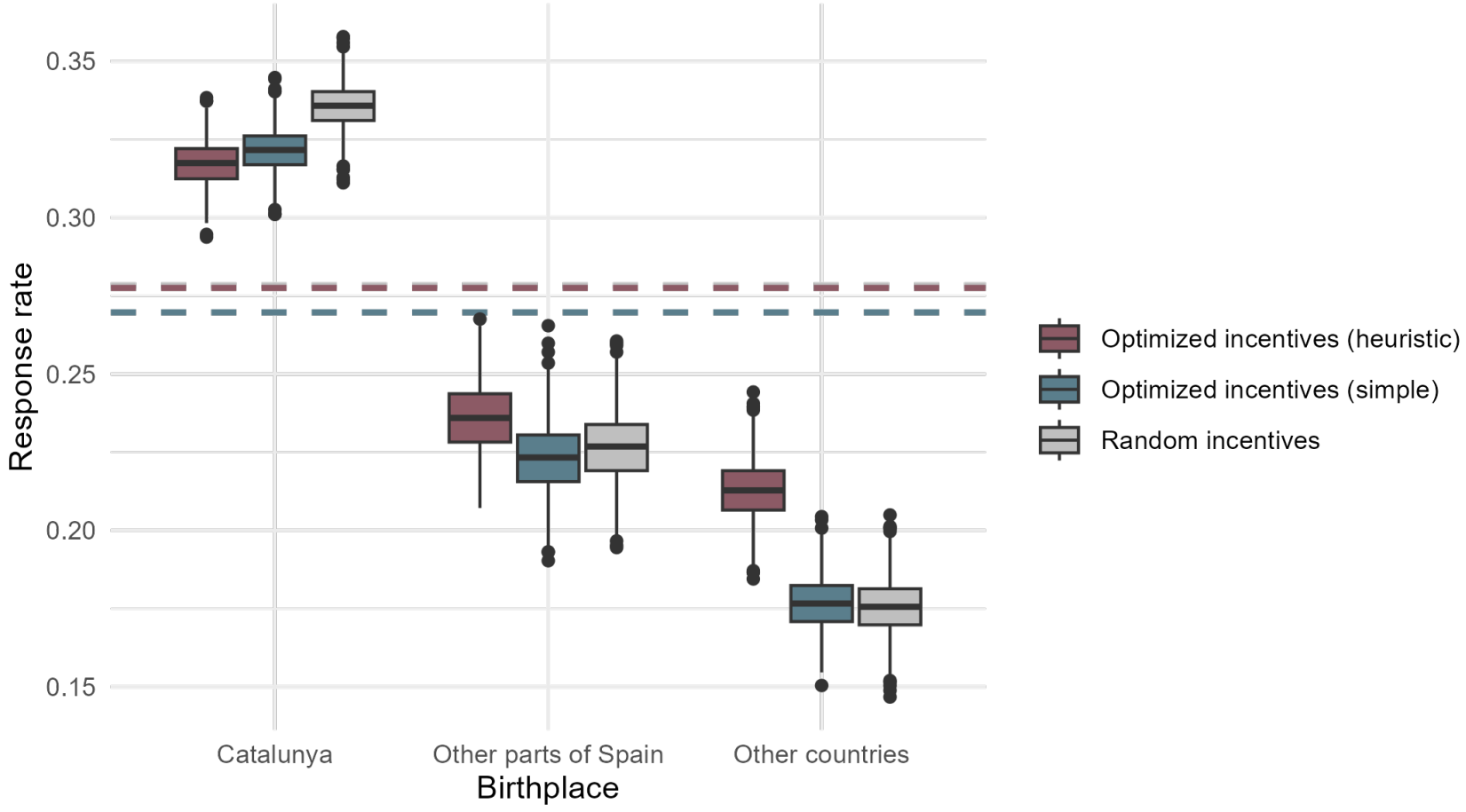


Figure 12: Simulated response rate by place of birth

Comparison of results

	Response rate	R-indicator
No incentives	25.0%	0.84
Random incentives	27.8%	0.81
Optimized incentives (simple)	27.0%	0.84
Optimized incentives (heuristics)	27.7%	0.87

Conclusion and further steps

1. The proof of concept shows that we can improve the representativity of the panel with a version of the ASD
2. In further iterations we will:
    1. Run the ASD to improve the R of panel members in subsequent waves. In this case the models will work better, because they will use the information provided in the first survey
        1. Provided in previous surveys
        2. Their individual nonresponse history
    2. Adapt the treatment conditions, to exclude the "no incentive" condition and include the treatments based on the number of reminders
    3. Further explore potential treatments that have effects on the units with lower P_{ij}