Continuous Survey Sample Optimization Using Ad Platform APIs

China India Insights Conference, HKU
June 21, 2024

Nandan Rao [1] Dante Donati [2]

[1]Virtual Lab and UAB

[2]Columbia Business School

Introduction

Importance of Survey Research

Surveys are essential for understanding trends and making informed decisions across various domains:

- **Academics**
  - Conduct experiments to validate hypotheses, train AI models.
- **Industry Practitioners**
  - Market research to informs product demand, development, and customer satisfaction.
- **Institutions**
  - Assesses program effectiveness and guide policy.
- **Political Polls**
  - Gauge public opinion and predict outcomes.

Sources for Sample Collection

Traditional Methods

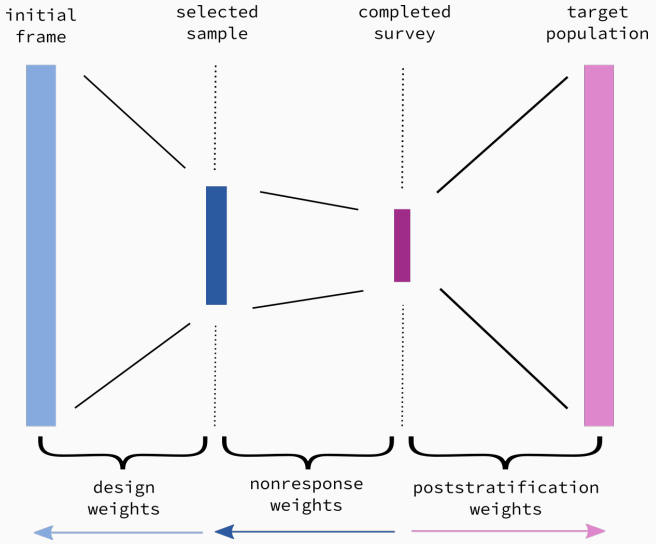
- Face-to-Face Interviews (GSS, lab studies, focus groups)
- Mail Surveys (census)
- Random Digit Dialing (phone-based opinion polls)

Internet-Based Methods

- Standard Online Panels (e.g., YouGov, Prolific)
- Platforms and Social Media (e.g., ads to recruit participants)

To some extent, all methods are subject to **selection** and/or **non-response** errors.

Weighting


initial frame
selected sample
completed survey
target population
design weights
nonresponse weights
poststratification weights

Social Media Sampling in Research

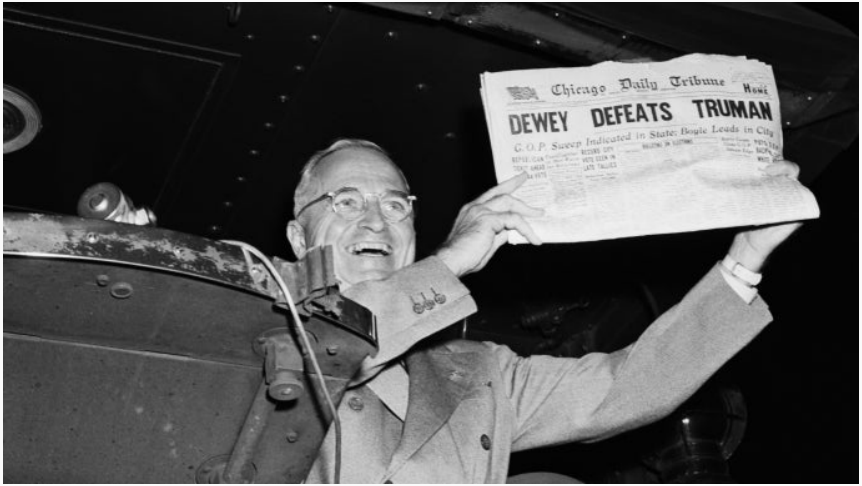
Researchers and companies have begun to use **Ads on Meta as a recruitment tool:**

- Reach wide and diverse populations.
- Relatively cheap compared to traditional methods and panels.
- Real-time data.

Problem: Selection Bias

- Not everyone uses social media.
- Social media ad delivery algorithms may skew the sample.

Skewed and unrepresentative samples can be misleading or dangerous



So what?

Existing solutions for social media samples:

- Quota sampling using **targeted advertising** (Zhang et al. 2020)
- **Poststratification weights** to recover population quantities (Zagheni et al. 2017)

This paper:

1. **Develops a methodology to optimally allocate ad budget across strata:**

   Take poststratification weights as a target, use ad platform API to dynamically adjust ad budget across strata (ad sets) to minimize variance subject to budget constraints.

2. **Validates it by collecting samples of Meta users in various countries:**

   Compare survey responses with benchmarks from nationally-representative face-to-face surveys (e.g., GSS, CPS, DHS).

Optimization

Setup

We consider the simplest case for poststratification weighting.

- We want to estimate a **population parameter** Y via sampling and measurement.
- The researcher wishes to use the stratified mean for a set of **strata** $h \in H$ (mutually exclusive cells), with a sample mean of $\bar{y}_h$ and an assigned **weight** W_h for each stratum:

$$\hat{Y} := \sum_h W_h \bar{y}_h$$

Idea

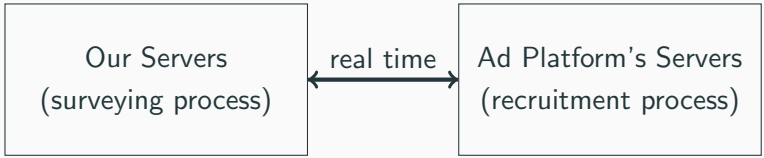
1. Commit to poststratification **weights** (e.g, take W_h from census).

Idea

1. Commit to poststratification **weights** (e.g, take W_h from census).
2. Run **recruitment ads** on Meta, targeting each stratum.

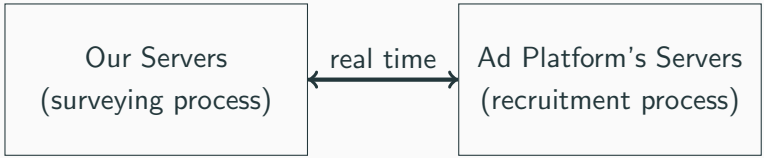
Idea

1. Commit to poststratification **weights** (e.g, take W_h from census).
2. Run **recruitment ads** on Meta, targeting each stratum.
3. **Measure** response rate during the surveying process and **adjust** the selected sample (ad budget) dynamically:



Idea

1. Commit to poststratification **weights** (e.g, take W_h from census).
2. Run **recruitment ads** on Meta, targeting each stratum.
3. **Measure** response rate during the surveying process and **adjust** the selected sample (ad budget) dynamically:



4. Make the adjustment to **minimize variance** subject to budget constraints.

Unconstrained optimization

- Assuming the variance of the outcome Y in each stratum is equal: $s_h^2 = s^2$.
- Then, the variance of the sample estimate is given by:

$$\mathbb{V}[\hat{Y}] = s^2 \sum_h \frac{W_h^2}{n_h}$$

  where n_h **is the sample size in stratum** h**.**

- This quantity is **minimized when** $\frac{n_h}{N} = W_h$, known as the Neyman (1934) allocation.

But...

We don't have infinite money...

Optimization under budget constraint

- Let P_h be the **cost to recruit** an individual from stratum h.
- Let B be the total **budget**, such that $P_h n_h \leq B_h$.
- Let N_d be the desired maximum sample size.

We can then frame the optimization problem of **finding the best allocation of budget to minimize the variance of the final estimate** as:

$$\begin{aligned} \underset{n_i,\ldots,n_h}{\operatorname{argmin}} \sum_h \frac{W_h^2}{n_h} \\ s.t. \sum_h P_h n_h \leq B \\ \sum_h n_h \leq N_d \end{aligned}$$

Price estimation Algorithm

Software

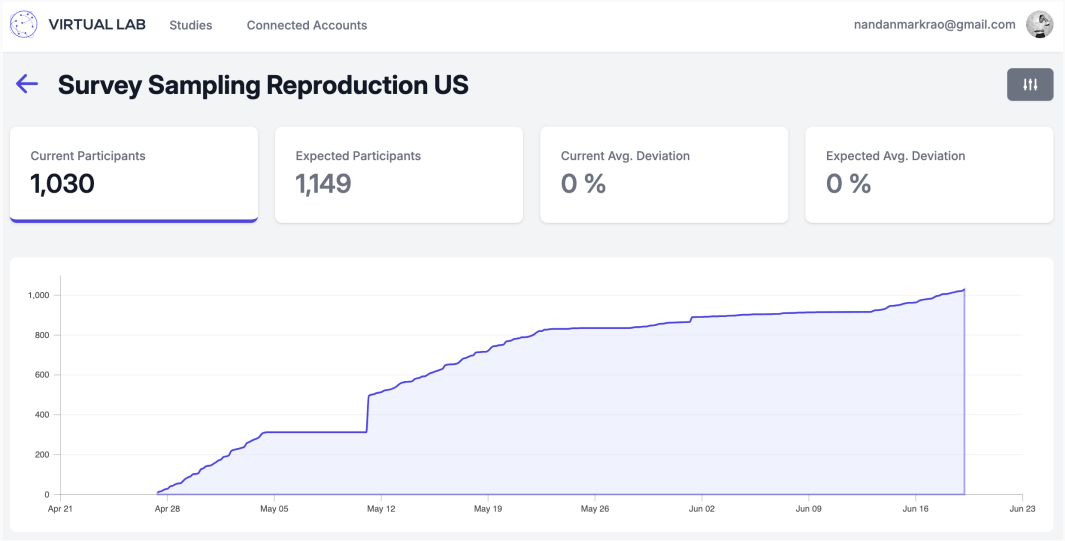
Software

Software is fully open source:

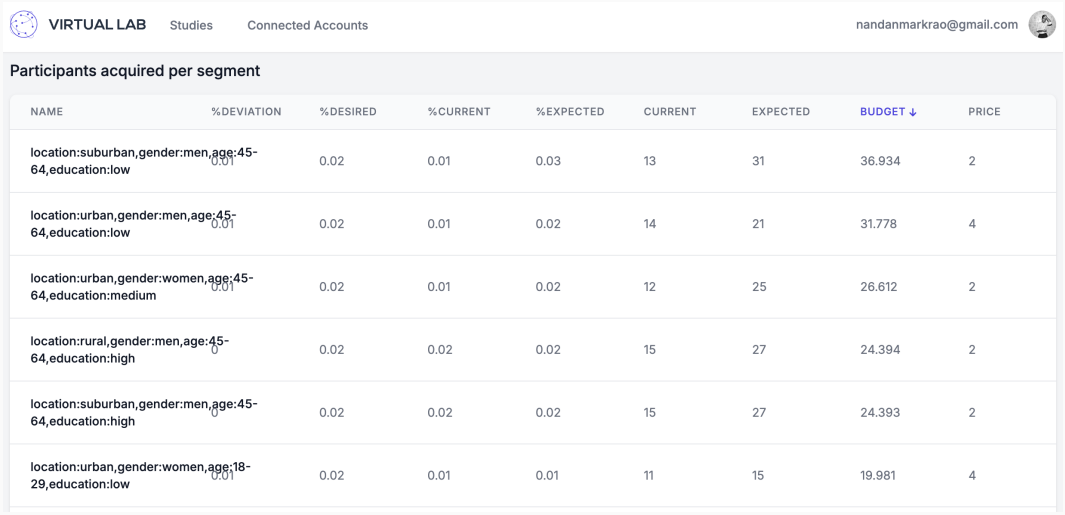
- **Input:** strata+weights, total budget, ad creatives, survey
- **Output:** csv file with responses

github.com/vlab-research

Dashboard: Progress over time



Dashboard: Stratum characteristics



NAME	%DEVIATION	%DESIRED	%CURRENT	%EXPECTED	CURRENT	EXPECTED	BUDGET ↓	PRICE
location:suburban,gender:men,age:45-64,education:low	0.01	0.02	0.01	0.03	13	31	36.934	2
location:urban,gender:men,age:45-64,education:low	0.01	0.02	0.01	0.02	14	21	31.778	4
location:urban,gender:women,age:45-64,education:medium	0.01	0.02	0.01	0.02	12	25	26.612	2
location:rural,gender:men,age:45-64,education:high	0	0.02	0.02	0.02	15	27	24.394	2
location:suburban,gender:men,age:45-64,education:high	0	0.02	0.02	0.02	15	27	24.393	2
location:urban,gender:women,age:18-29,education:low	0.01	0.02	0.01	0.01	11	15	19.981	4

Validation

Data

Meta sample:

- Collect samples of users using Ads on Facebook+Instagram
- 5 countries: US (ongoing) + India, Indonesia, Nigeria and Mexico (to be done)
- Offer Amazon gift cards or mobile top-ups as incentives

Nationally-representative benchmarks:

- US: General Social Survey (GSS, 2022), Current Population Survey (CPS, 2024), Pew Research (2023)
- Other countries: Demographic and Health Survey (DHS), Afrobaromoter, etc.

US study

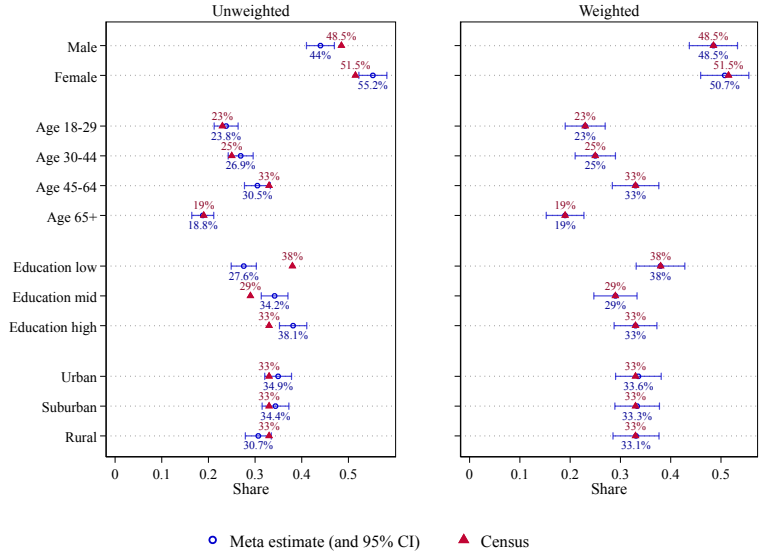
Recruitment

- N=1,059
- 72 strata: Gender (2), Age (3), Education (4), Settlement type (3)
- Partly exclude over-represented audiences (e.g., low income)
- Weights from the 2020 census

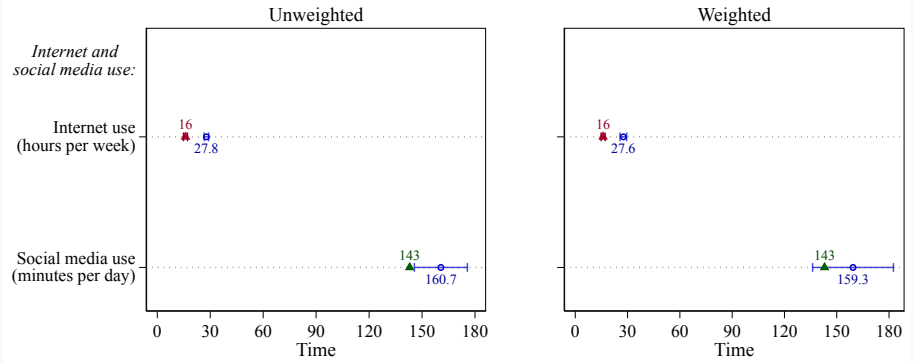
Survey outcomes

- Employment status and business ownership
- Health perceptions and satisfaction
- Use of technology and privacy concerns
- Trust and confidence
- Attitudes towards women and immigrants
- Political preferences

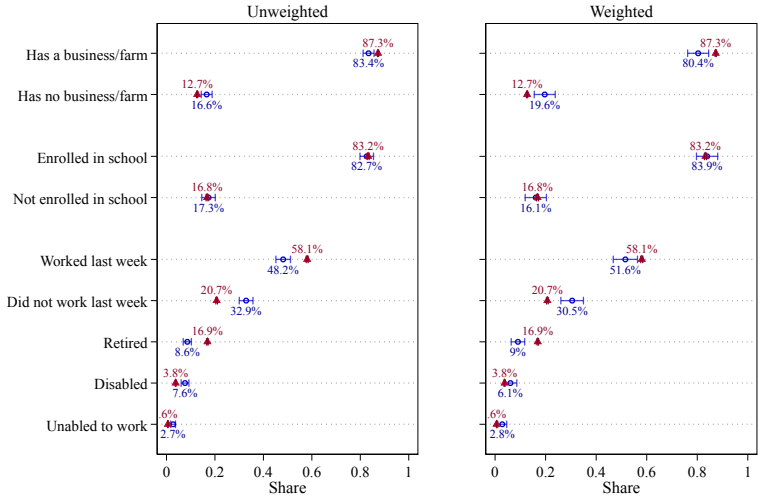
US Study Results: Stratification



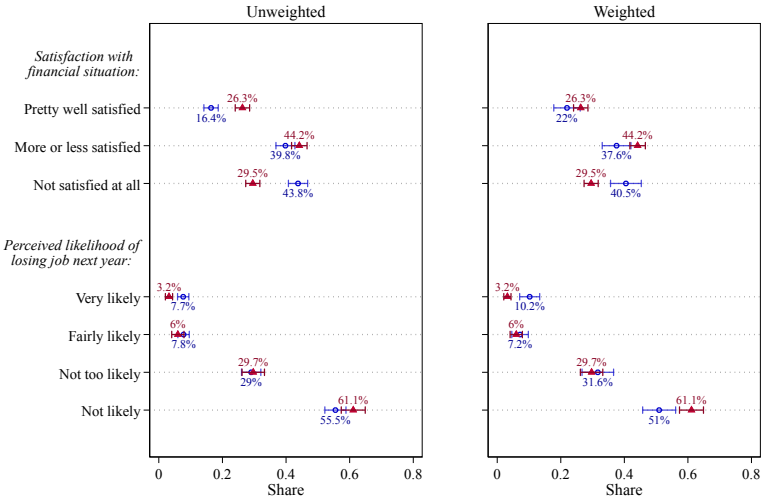
US Study Results: Internet/Social media



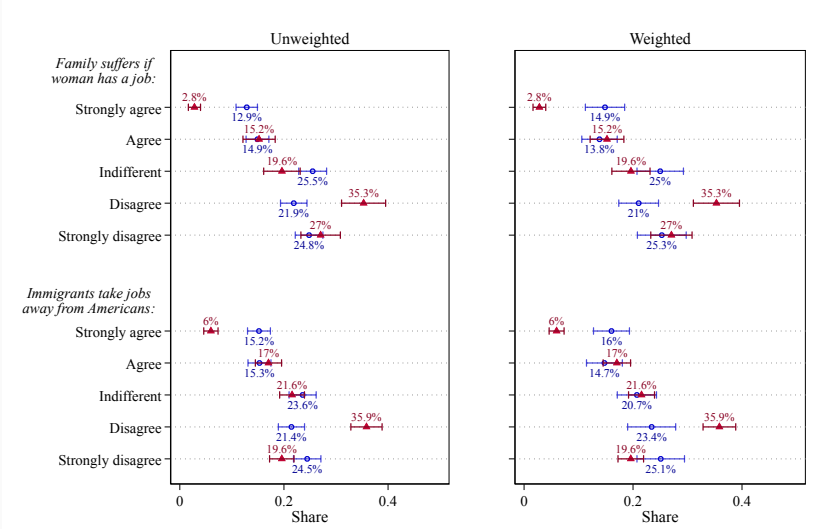
US Study Results: Employment/Education



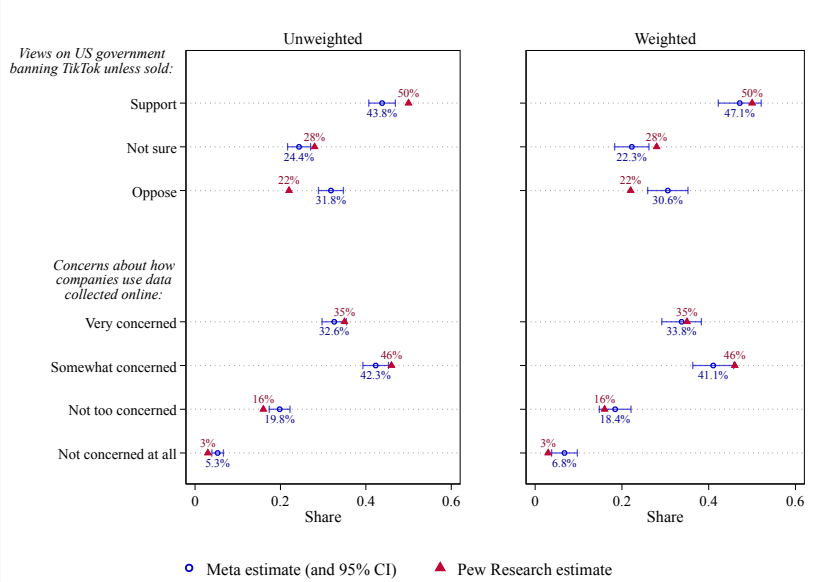
US Study Results: Satisfaction/Job security



US Study Results: Attitudes



US Study Results: Privacy Concerns



Conclusions

US study summary

- Compared to benchmark surveys, Meta respondents
  - Spend more time on the internet and social media
  - Report lower job security and financial satisfaction
  - Are more likely to be unemployed
  - Have more conservative attitudes towards women and immigrants
  - Are less concerned about privacy

US study summary

- Compared to benchmark surveys, Meta respondents
  - Spend more time on the internet and social media
  - Report lower job security and financial satisfaction
  - Are more likely to be unemployed
  - Have more conservative attitudes towards women and immigrants
  - Are less concerned about privacy

- Differences are generally small
  - **Mean Absolute Deviation (MAD): 6.2 percentage points (p.p.)**
  - MAD: MTurk vs. GSS/Pew: 7.3 p.p. (Goel, Obeng, and Rothschild 2015)
  - MAD: Pew online vs. phone-based surveys: 6 p.p. (Mercer, Lau, and Kennedy 2018)

Conclusions

- Costs per participant
  - US study:
    - Avg ad cost $6.6
    - $0.7 (urban young medium-education men) - $20 (urban mid-age low-education men)
    - $5 incentives
    - **$0.33 per question per respondent** (vs. $3 in the GSS)
  - 40+ studies (500K+ users worldwide): **avg ad cost $2.7**, varying incentives

⏩ Cost details

Conclusions

- Costs per participant
  - US study:
    - Avg ad cost $6.6
    - $0.7 (urban young medium-education men) - $20 (urban mid-age low-education men)
    - $5 incentives
    - **$0.33 per question per respondent** (vs. $3 in the GSS)
  - 40+ studies (500K+ users worldwide): **avg ad cost $2.7**, varying incentives

⏩ Cost details

- Sector-agnostic open-source software that can be applied to
  - Conduct survey research in **emerging markets**
  - Conduct **timely** surveys
  - Recruit **hard-to-reach** and **highly-targeted** audiences
  - Recruit from other ad platforms (TikTok, YouTube, etc.)

Thank you!

dd3137@gsb.columbia.edu

https://vlab.digital

Appendix

How to measure cost?

We can model the inverse cost ($\frac{1}{P_h}$), the number of respondents recruited n_{ht} given budget spend B_h, as a Poisson random variable:

$$n_{ht}|B_{ht} \sim Poisson(\lambda)$$
$$\lambda \sim Gamma(\kappa, \beta)$$

We can use closed-form Bayesian updating to obtain a MAP estimator of λ and the implied mean of the predictive distribution $(1/\lambda)$.

Back

Algorithm

How can we run this optimization problem?

We need an interface for recruitment that targets stratum h and allocates budget B_{ht} over a specific period of time t. We denote this interface Recruit(B_t) which accepts a budget allocation $B_t := \{B_{1t}, ..., B_{Ht}\}$.

Additionally, we require an interface GetResults(t) to collect information on the results of recruitment at time t given budget B_t. Results should be considered as the number of respondents recruited for each stratum h at time t and will be denoted n_{ht}.

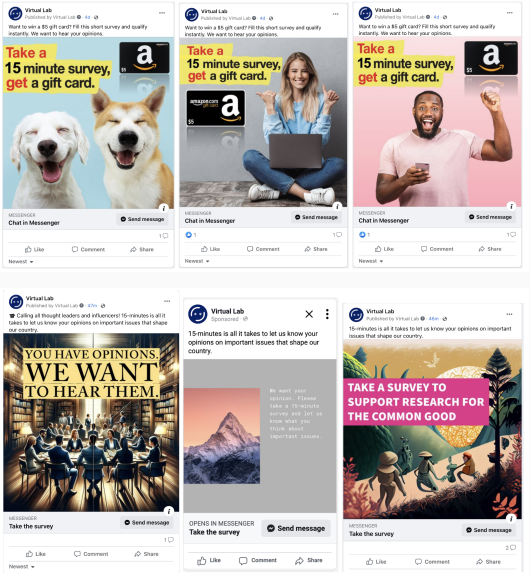
Algorithm

Algorithm 1 Optimizing Stratified Recruitment with Unknown Costs

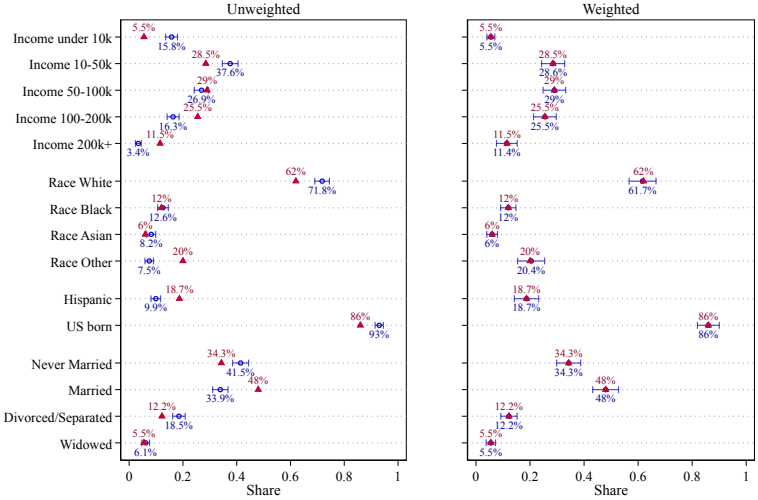
procedure AdOptimization($W, B, N_d, \kappa, \theta$)
  $B_0 := [1, ..., 1]$ ▷ Budget indexed by H strata
  $n := [0, ..., 0]$ ▷ Results indexed by H strata
  for $t \in T$ **do**
    $p_t := []$ ▷ Price estimates indexed by H strata
    for $h \in H$ **do**
      $n_{ht} :=$ GetResults(h, t)
      $n_h := n_h + n_{ht}$
      $p_{ht} :=$ EstimatePrice($\kappa, \theta, B_{ht}, n_{ht}$)
    end for
    $B_{t+1} :=$ Optimize(W, B, N_d, n, p_t)
    Recruit(B_{t+1})
  end for
end procedure

⏭ Back

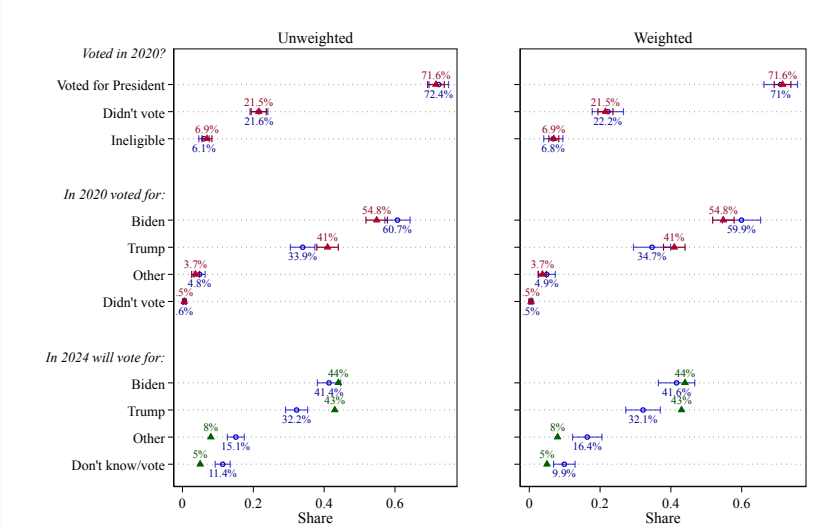
Ad creatives



US Study Results: Other Demographics



US Study Results: Political Preferences



Meta estimate (and 95% CI) ▲ Weighted GSS estimate (and 95% CI) ▲ MorningConsult (June 2024)

Average recruitment costs per country

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
0	India	24	5.09%	873653	9130	$0.24
1	Libya	176	27.1%	1000294	8338	$0.3
2	Labanon	48	15.03%	1370234	17399	$0.57
3	Jordan	192	10.05%	2223793	17223	$0.6
4	Iraq	144	8.58%	4280255	16015	$0.61
5	Serbia	48	13.58%	985109	13995	$0.67
6	Nigeria	23	6.42%	145437	2393	$0.75
7	US	10	13.61%	16849	1118	$0.85
8	US	4	6.9%	10616	316	$0.87
9	Haiti	16	10.0%	1218842	10491	$0.94
10	Honduras	144	7.4%	909597	4922	$0.95

Average recruitment costs per country

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
11	Lebanon	48	13.82%	1650052	14354	$1.03
12	Papa NG	64	8.71%	118980	1825	$1.46
13	Iraq	80	8.27%	5616663	6520	$1.55
14	US	14	6.14%	120647	2553	$1.66
15	Ukraine	64	5.17%	648512	2394	$1.69
16	Kyrgyzstan	24	9.63%	489054	3004	$1.76
17	Djibouti	16	10.84%	313493	2252	$2.19
18	Kosovo	56	19.93%	663059	6084	$2.46
19	Chad	32	7.59%	327048	2305	$2.64
20	Jamaica	16	15.03%	482097	4105	$2.72
21	Belize	24	9.1%	43684	264	$2.89

Average recruitment costs per country

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
22	Serbia	1	6.75%	342403	1737	$2.99
23	Macedonia	32	24.58%	384956	3156	$3.19
24	Romania	200	9.31%	785811	1863	$3.23
25	Macedonia	32	25.17%	538295	4565	$3.26
26	Jordan	96	7.14%	1304979	2146	$3.72
27	US	16	3.93%	241134	1429	$4.55
28	Nigeria	120	7.75%	563939	177	$5.55
29	Bulgaria	8	3.91%	170205	170	$6.16
30	Cameroon	80	11.85%	1427793	1712	$6.72
31	India	160	5.77%	3526970	1639	$7.85
32	Gambia	2	5.86%	279008	697	$11.07

Back