Pew Research Center

Online Panel Sample Quality

Andrew Mercer

Principal Methodologist

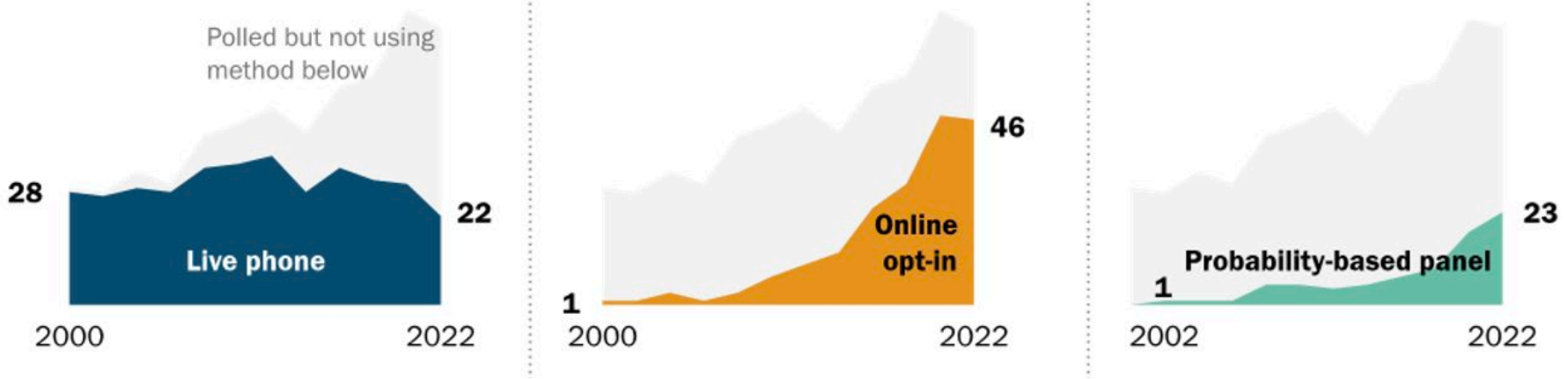
BACKGROUND

Two broad types of online survey panels

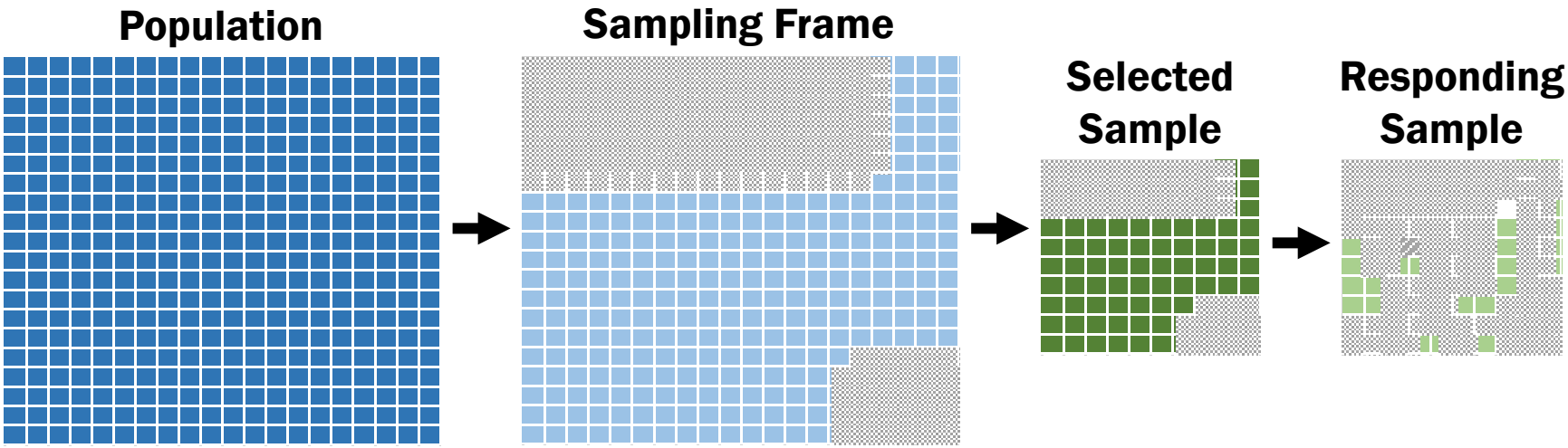
	Probability-based	Opt-in
Sources	Random sampling from U.S.P.S. residential address list	Convenience (e.g., social media ads, mobile apps, customer loyalty lists)
Examples	American Trends Panel, KnowledgePanel, AmeriSpeak, SSRS Opinion Panel	Cint, Qualtrics, Protégé, Prolific
Cost	Relatively high	Relatively low

Online opt-in now most common method for election polls

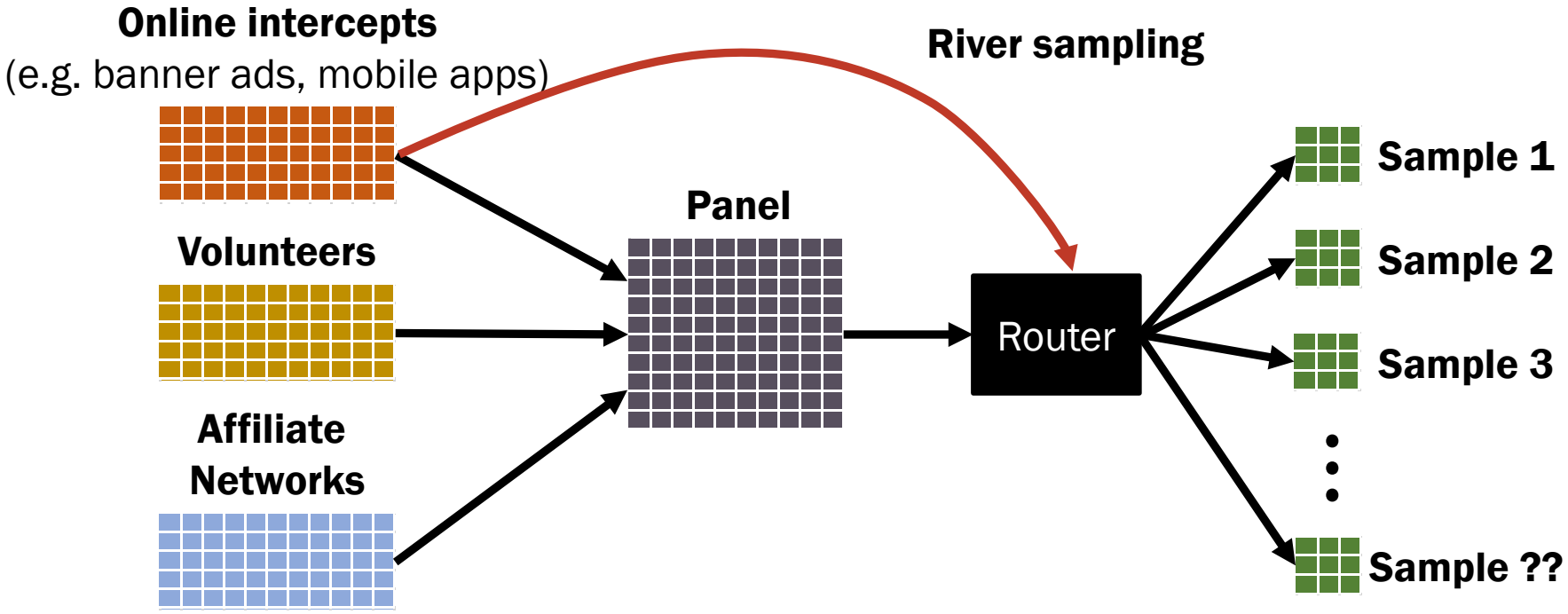
Number of U.S pollsters using method alone or in combination with other methods



Probability-based sampling...



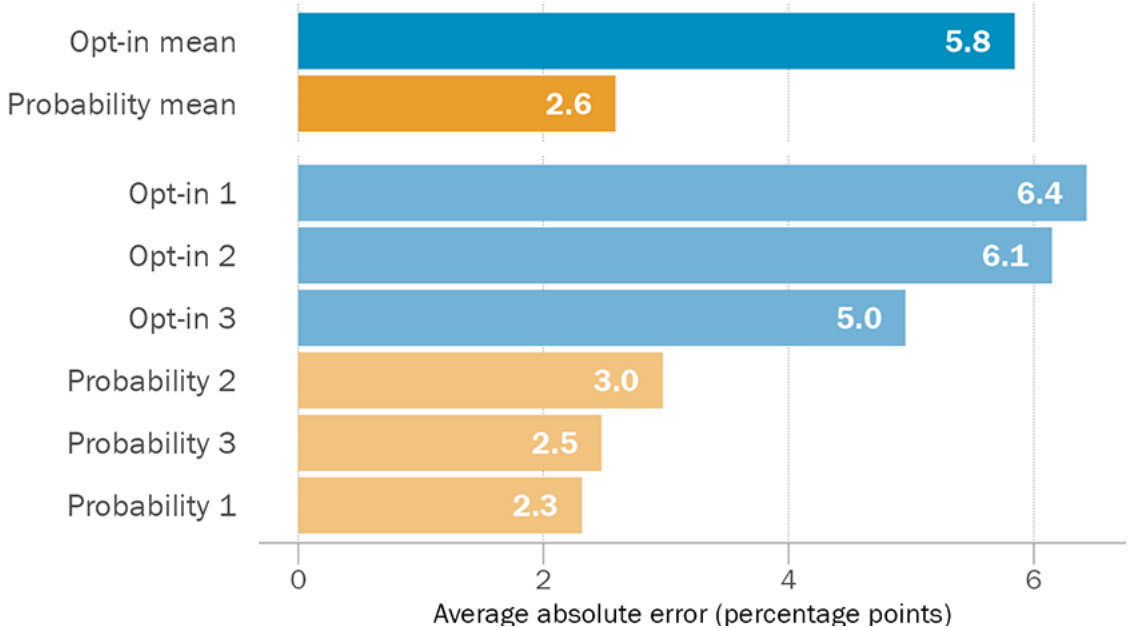
Online nonprobability sampling…

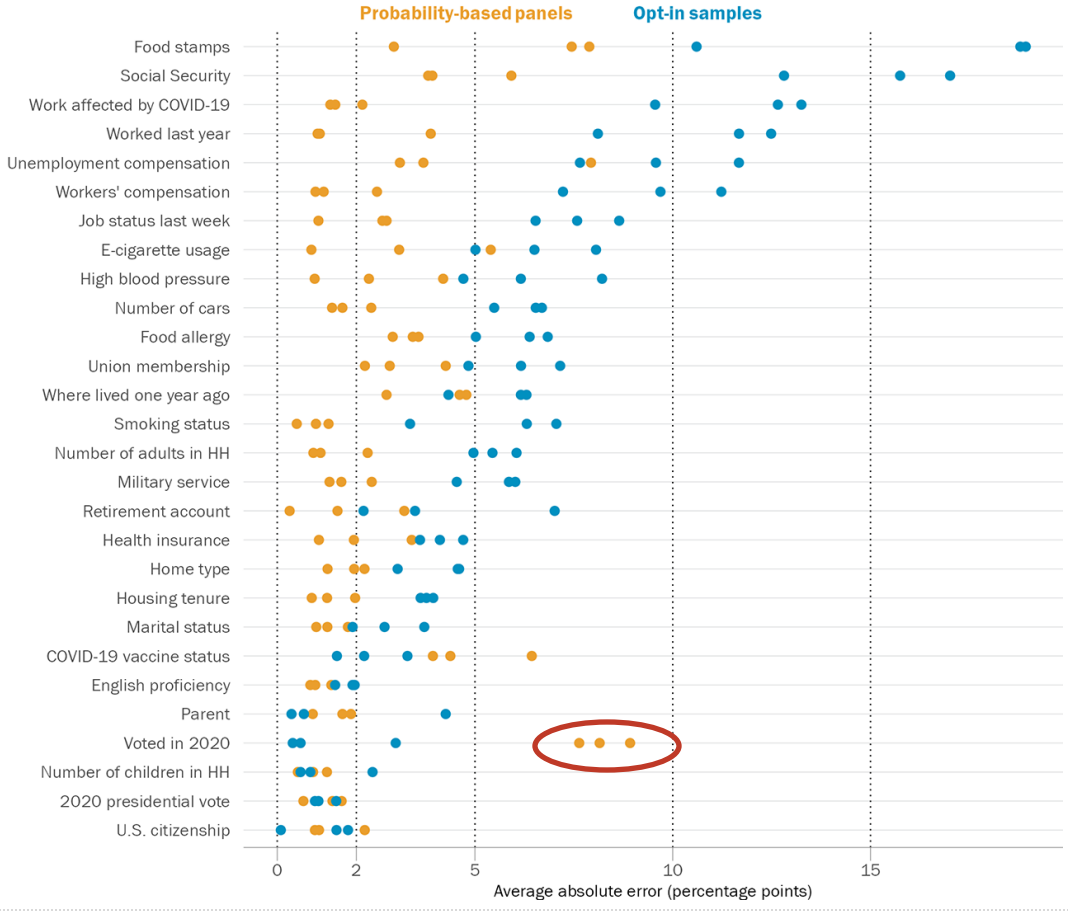


COMPARING DATA QUALITY

Average error on online opt-in samples was twice that of probability-based panels

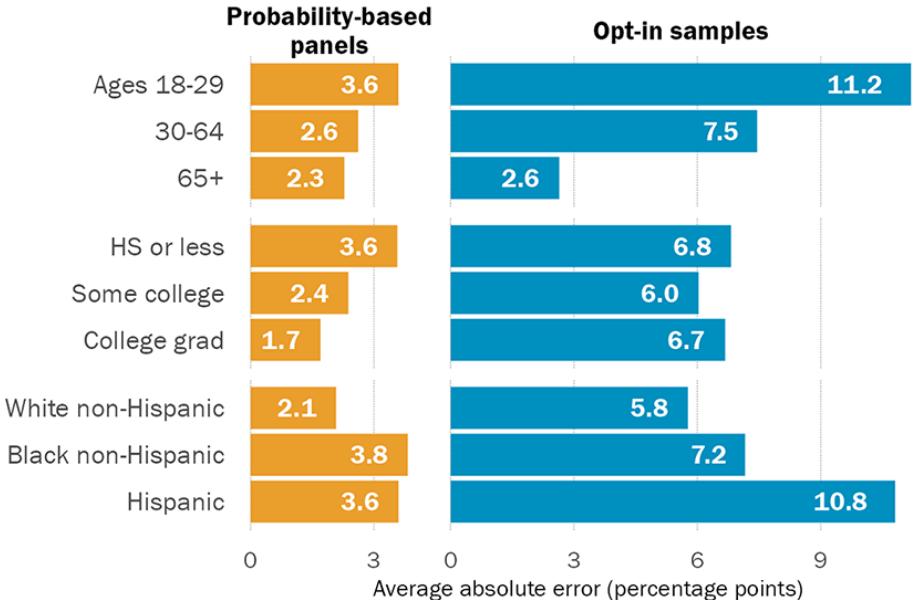
Average absolute error on 28 benchmark variables for estimates among U.S. adults




Probability-based panels
Opt-in samples
Food stamps
Social Security
Work affected by COVID-19
Worked last year
Unemployment compensation
Workers' compensation
Job status last week
E-cigarette usage
High blood pressure
Number of cars
Food allergy
Union membership
Where lived one year ago
Smoking status
Number of adults in HH
Military service
Retirement account
Health insurance
Home type
Housing tenure
Marital status
COVID-19 vaccine status
English proficiency
Parent
Voted in 2020
Number of children in HH
2020 presidential vote
U.S. citizenship
0
2
5
10
15
Average absolute error (percentage points)

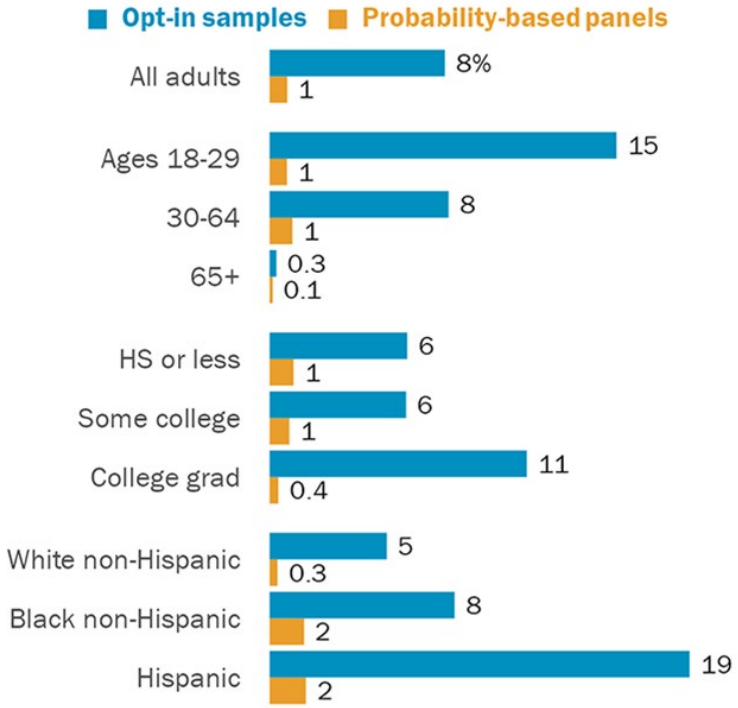
Large errors for 18- to 29-year-olds and Hispanic adults on opt-in samples

Average absolute error on 25 benchmark variables for estimates among demographic subgroups



Some opt-in respondents tend to answer 'Yes' regardless of the question

Average % of U.S. adults who answered "Yes" to at least 10 of 16 Yes/No questions



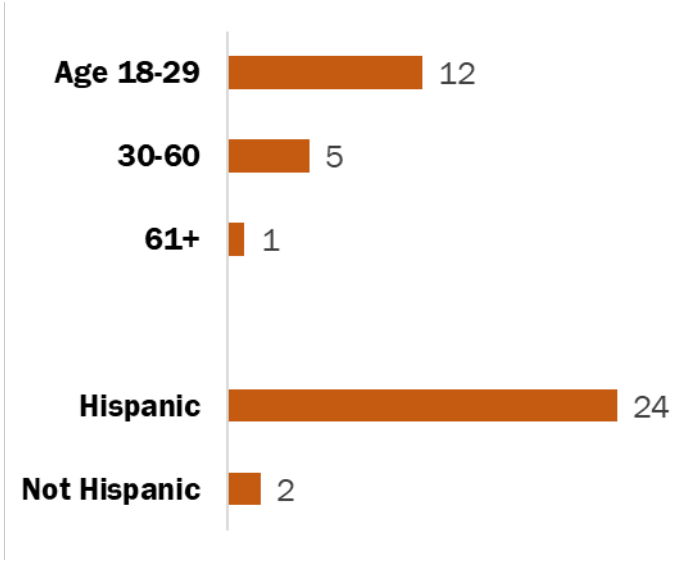
The problem of “bogus respondents” in opt-in samples

- Respondents who make little to no effort to answer survey questions truthfully.
- Disproportionately concentrated in key demographic subgroups.
- Results in a pattern of error that is dramatically larger within those groups.
- This is measurement error that’s not easily fixed through weighting.

- Does not manifest in probability-based samples.

Same pattern of results on a different opt-in sample

% of respondents saying they are licensed to operate a class SSGN submarine



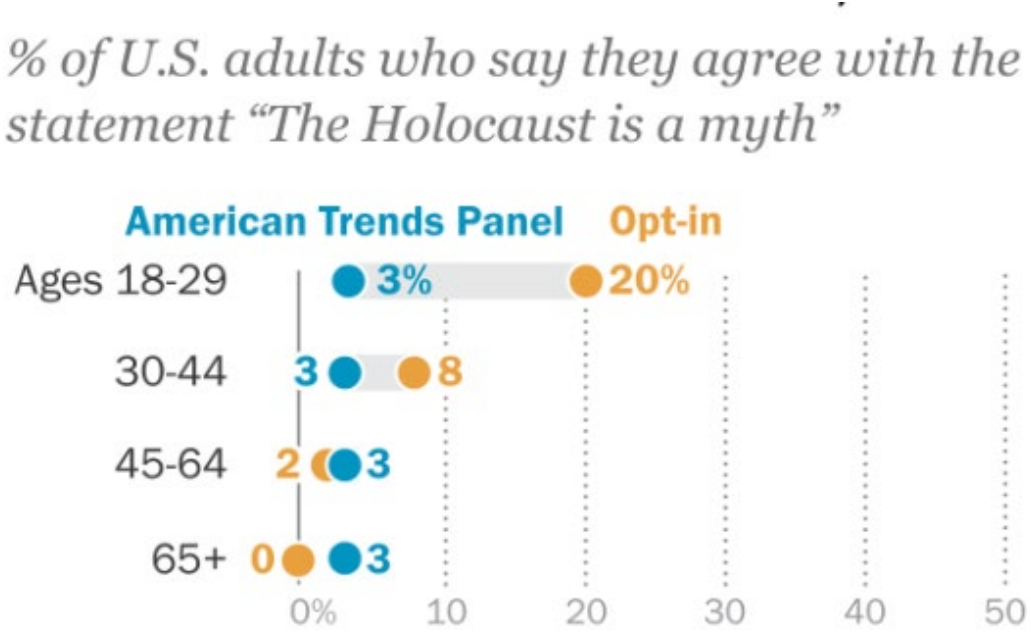
United States | Young opinions

One in five young Americans thinks the Holocaust is a myth

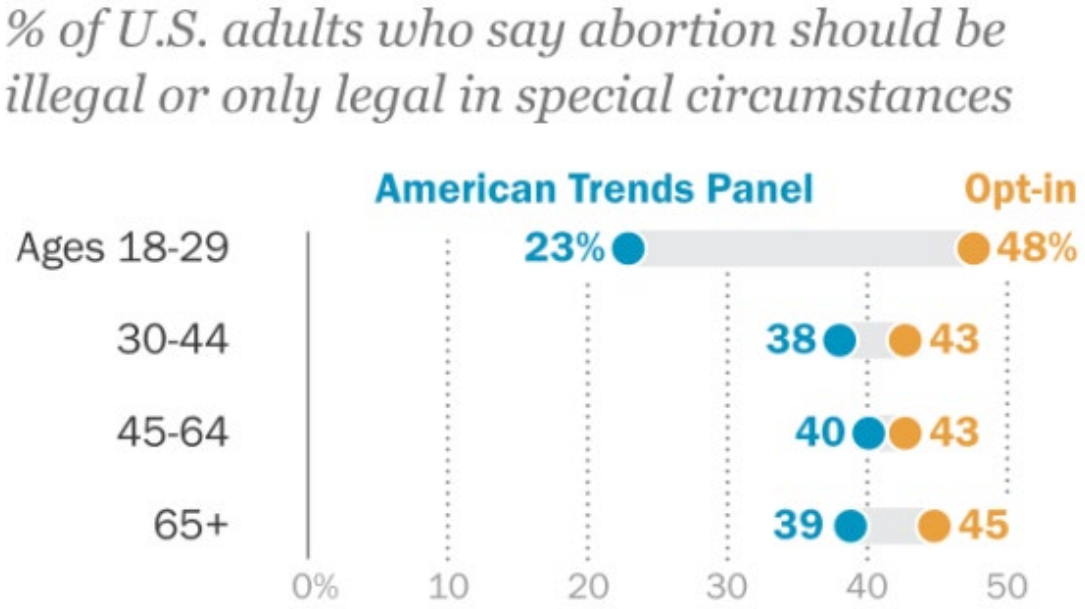
Our new poll makes alarming reading

Dec 7th 2023 | WASHINGTON, DC

We tried to replicate this result on the ATP



Findings on abortion attitudes also failed to replicate



FUTURE DIRECTIONS

Overarching goals

1. Improve accuracy
2. Reduce costs

These are often in conflict

Suggest different priorities for probability vs. opt-in

Opt-in: Bogus respondents and measurement error

Current methods fail to detect many bogus respondents:

- Current methods like trap questions or speeding checks fail to detect bogus respondents
- Asking open-ended text questions is more reliable but cumbersome

Opportunity for generative AI?

- Identifying suspicious open-ended text?
- Identifying other complex combinations of respondent behavior?

Other means of validation?

- Matching to registered voter lists
- Matching to other sources of administrative data?

Probability-based: Improving representation

Probability-based panels overrepresent:

- The politically and civically engaged

Probability-based panels underrepresent:

- Black and Hispanic adults (especially young men)
- Political conservatives
- “Web-hesitant” adults

High quality auxiliary data

Improving survey accuracy requires either:

- Improving the composition of samples
- Better models to correct biases

Requires knowing what the true population looks like:

- Federal government cannot produce benchmarks for things like partisanship or religious affiliation
- National voter registries useful for election polls but have poor coverage of the entire population (non-voters, non-citizens, etc…).
- What are other sources of administrative data that can inform sample design or statistical modeling?

High-quality reference surveys

National Public Opinion Reference Survey (NPORS)

- Launched by Pew Research Center in 2020
- Large scale address-based sample with both online and paper completion modes.
- Had a 32% response rate in 2024 (AAPOR RR1)
- Over half complete via paper
- Source of survey weighting targets related to partisanship, religion, internet use, and other non-demographic measures.

CNN and the Kaiser Family Foundation have launched a similar initiative

Are there opportunities for similar initiatives or research into best practices?

Some conclusions…

- Many of the problems seen in online opt-in samples are due to a race-to-the-bottom dynamic for generating as many survey completes as cheaply as possible.
- Now there is buzz about using LLMs to simulate survey respondents. This is fundamentally flawed.
- In recent years, the biggest improvements in data quality have involved *lower*, not higher, tech solutions (e.g. mail contact and paper questionnaires).
- In the next decade, it will be important to resist the appeal of techno-solutionism.
- As a field, we should refocus on methods for high-quality data and data infrastructure, even though it is difficult, expensive and less glamorous.

Thank you!

Andrew Mercer
Principal Methodologist

amercer@pewresearch.org