# Summary of Voter Match Research at AAPOR 2018

**Edward Paul Johnson, RN-SSI**

June 29, 2018

# Presentations Summarized

**The Truth is Out There: Using Voter Files to Improve Election Polls**

Assessing Coverage Bias in Registration-Based Sample
(Seth Brohinsky, Abt Associates & Scott Clement, Washington Post)

Examining Coverage and Response Bias Between Data Collection Modes in Voter Files
(Edward Paul Johnson, RN-SSI & Nathan LaCombe, The Data Trust)

Examining the Accuracy of Likely Voter Models
(Scott Clement, Washington Post)

Relational validity: A new approach to evaluating political surveys
(Jonathan Robinson, Catalist & Kevin Collins, ChangePoint Analytics)

Verifying Voter Registration Records
(Michael McDonald, University of Florida)

**Panel or Wallpaper?: How to Cover your Survey Needs and other advice on…Online Panels**

Creating a Probability Framework for Online Panels by Matching to Voter Files
(Edward Paul Johnson, RN-SSI & Kori Bishop, Aristotle)

# Assessing Coverage Bias in Registration-Based Sample: Overall Research

**Problem:**

Phone RDD interviewing of voters is getting expensive.

- Decreased Response Rate
- Increased Cell Phone Cost

Voter File Sampling offers solutions:

- Extensive frame information
- Efficiently target registered voter population
- Improved contact rates
- Access to out-of-area cell phones for state and local surveys
- Lower cost per interview

However:

- Missing phone numbers in the voter file can be up to 30%
- Can't use census data to develop reliable weighting targets

**What bias is associated with the missing phone numbers?**

**Research:**

Used Aristotle Voter File in Virginia

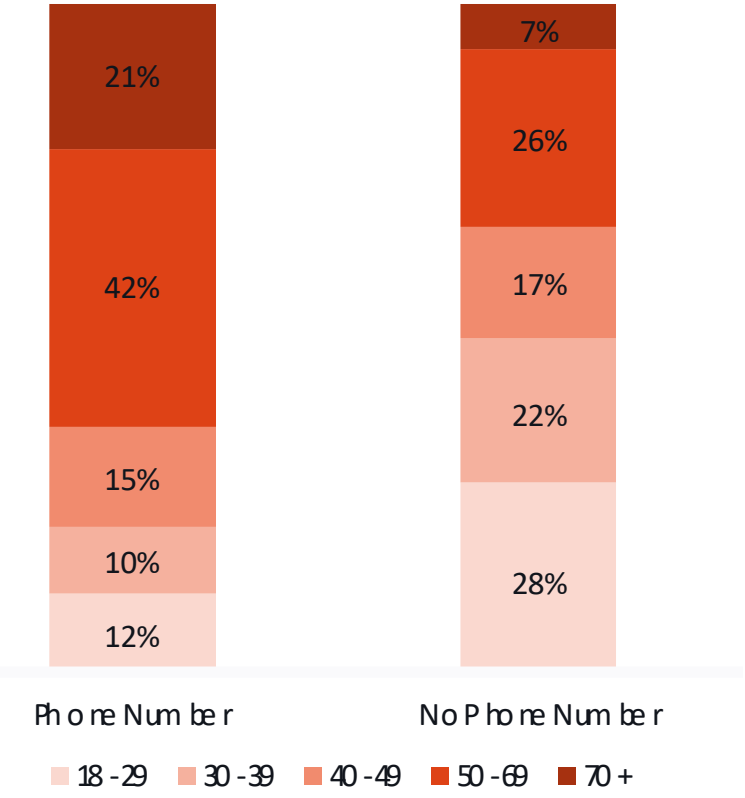
**Found important differences in frame with and without phone numbers**

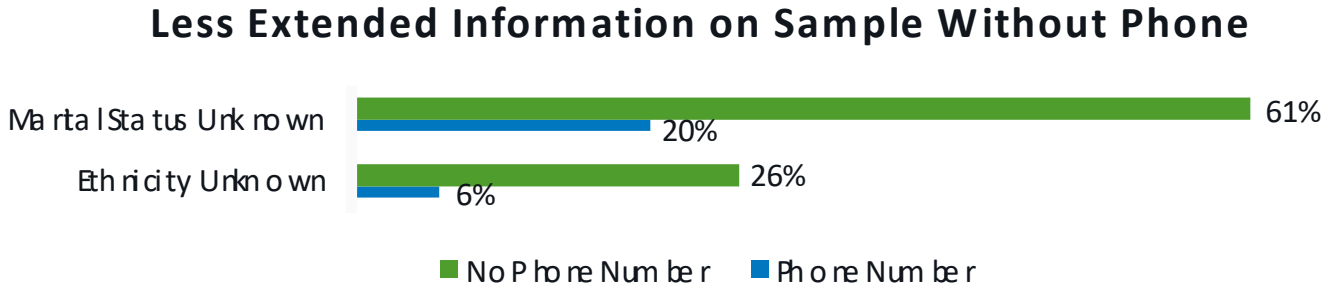
Dual Frame RDD/Voter File Survey

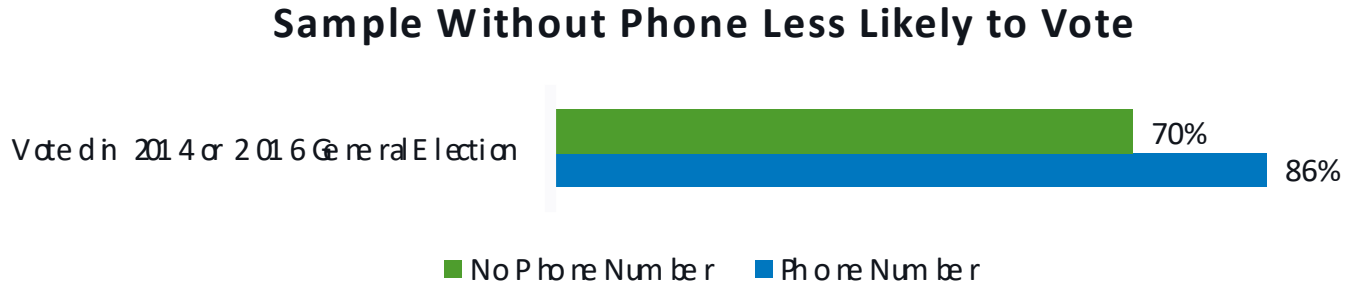
- Two waves (combined N=2,359)
- Tried to match RDD sample back to voter file on first/last name, zip, age.

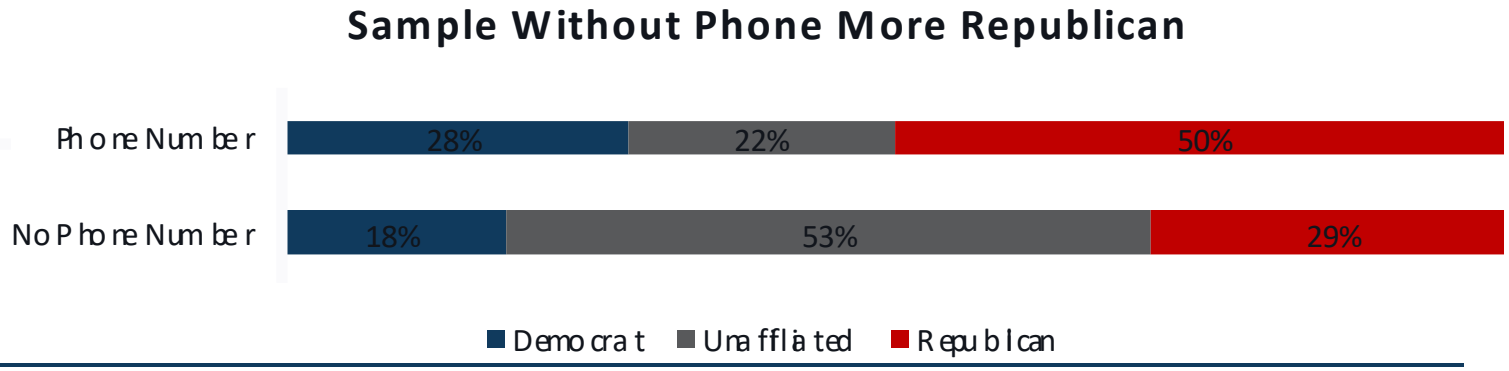
- Very little matched back to voter file records with no phone numbers (N=96).

# Assessing Coverage Bias in Registration-Based Sample: Skews in Aristotle VA Voter File Only with Phone



# Examining Bias in Voter File Between Collection Modes: Overall Research

**Problem:**

Phone interviewing of voters is getting expensive and has limitations.

- Decreased Response Rate
- Increased Cell Phone Cost
- Hard to show multi-media (Ad-testing)

Online Voter File Match offers solutions:

- Extensive frame information
- Efficiently target registered voter population
- Improved participation rates
- Lower Cost

**How should we start doing surveys on the voter file in both modes?**

**Research:**

Used National Data Trust file

N=500 Online Matched

N=500 CATI Non-Matched

- 300 Landline, 200 Cellphone

Topic was Tax Reform

Match Criteria

Individual Matches

Full Name, Address (Zip+4), Year of Birth* (I)
Full Name, Address (Zip+4), Year of Birth (S)

Full Name, Zip Code, Year of Birth (M)

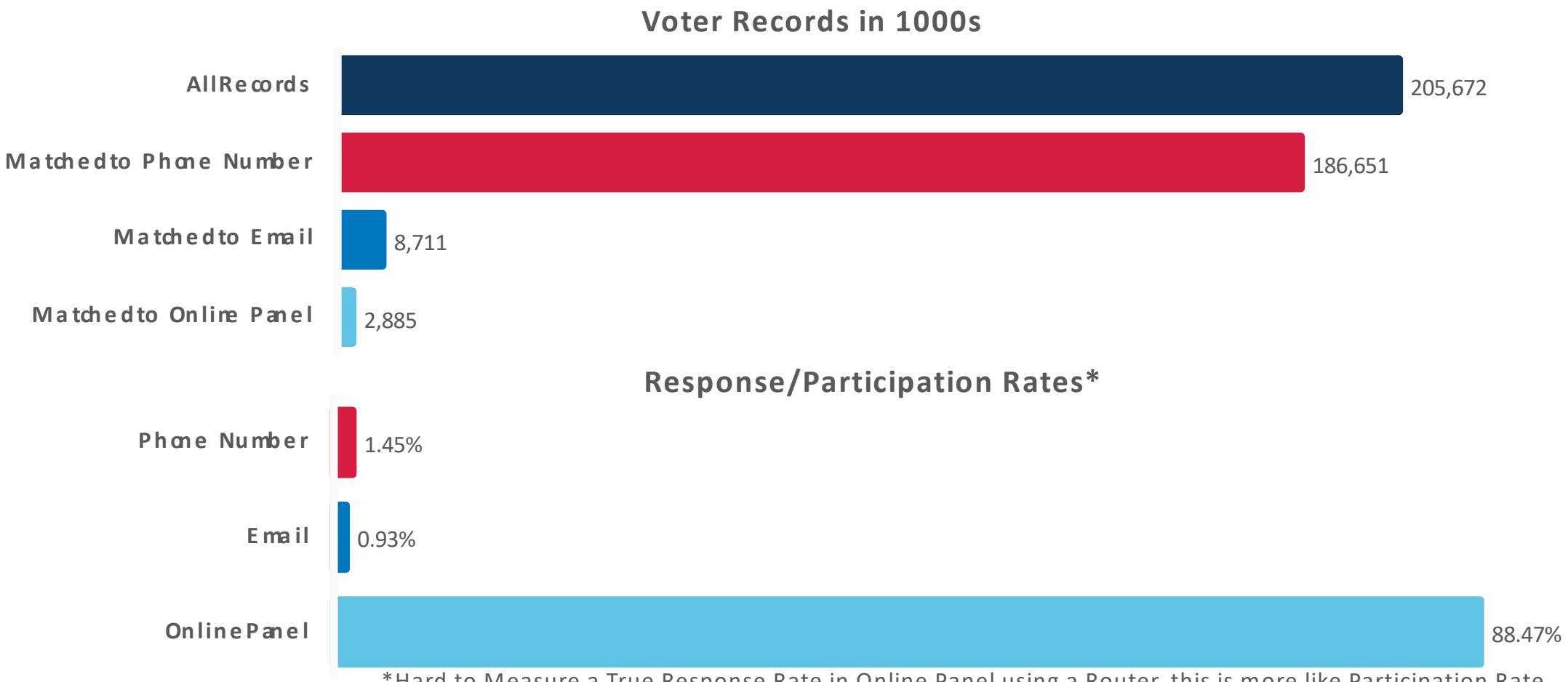
Household Matches

Last Name, Address (Zip+4)

**Found important differences in frames**

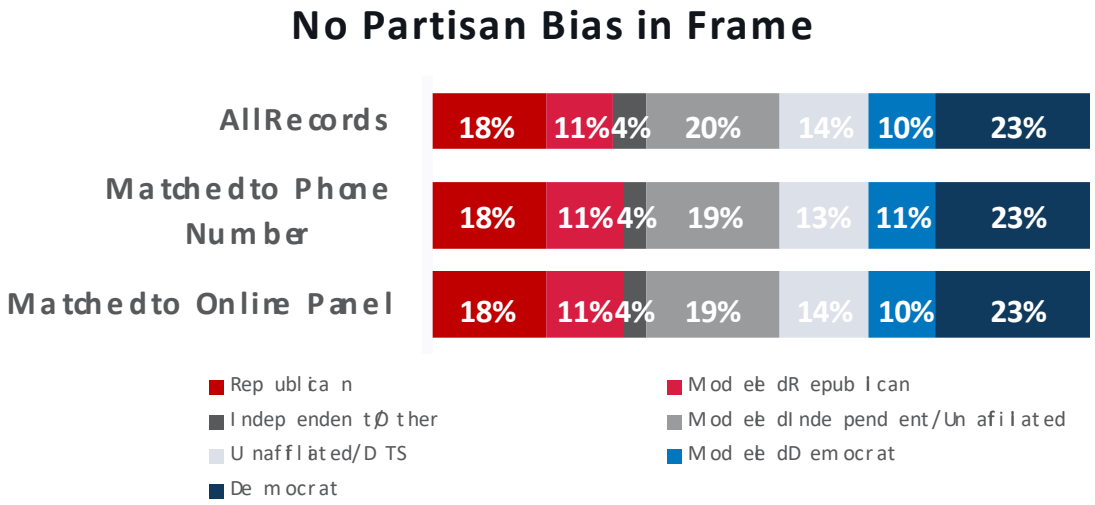
**Found important differences in questionnaire modes**

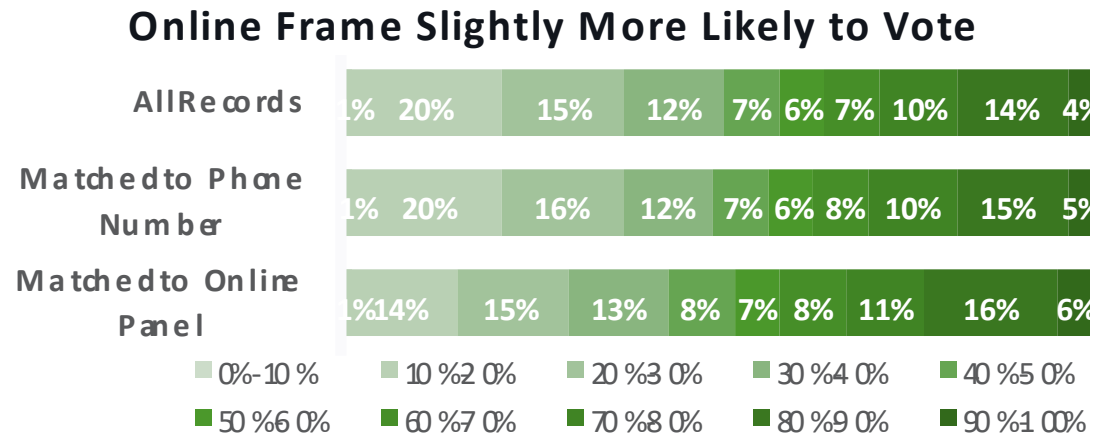
# Examining Bias in Voter File Between Collection Modes: Online Panel Low Coverage vs. Phone Low Response

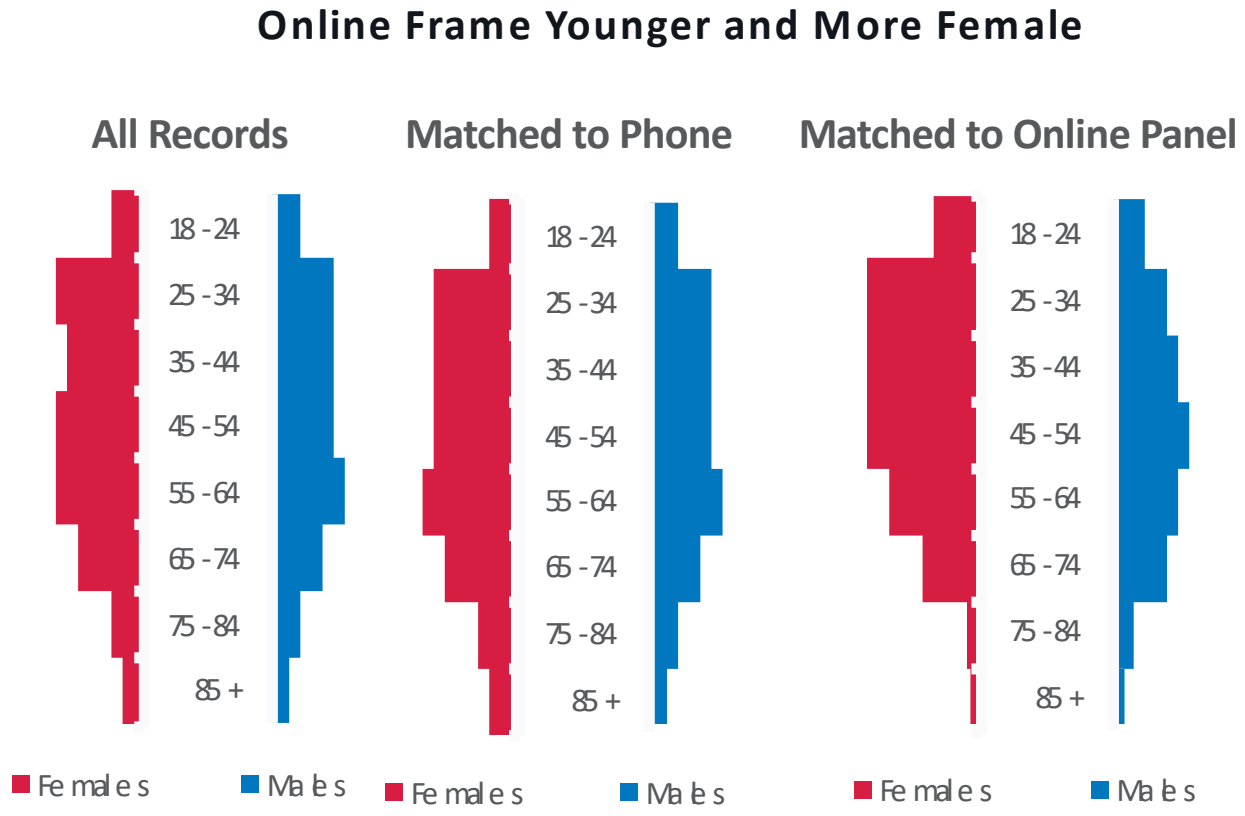



*Hard to Measure a True Response Rate in Online Panel using a Router, this is more like Participation Rate. Directly contacting via email or phone is much easier.

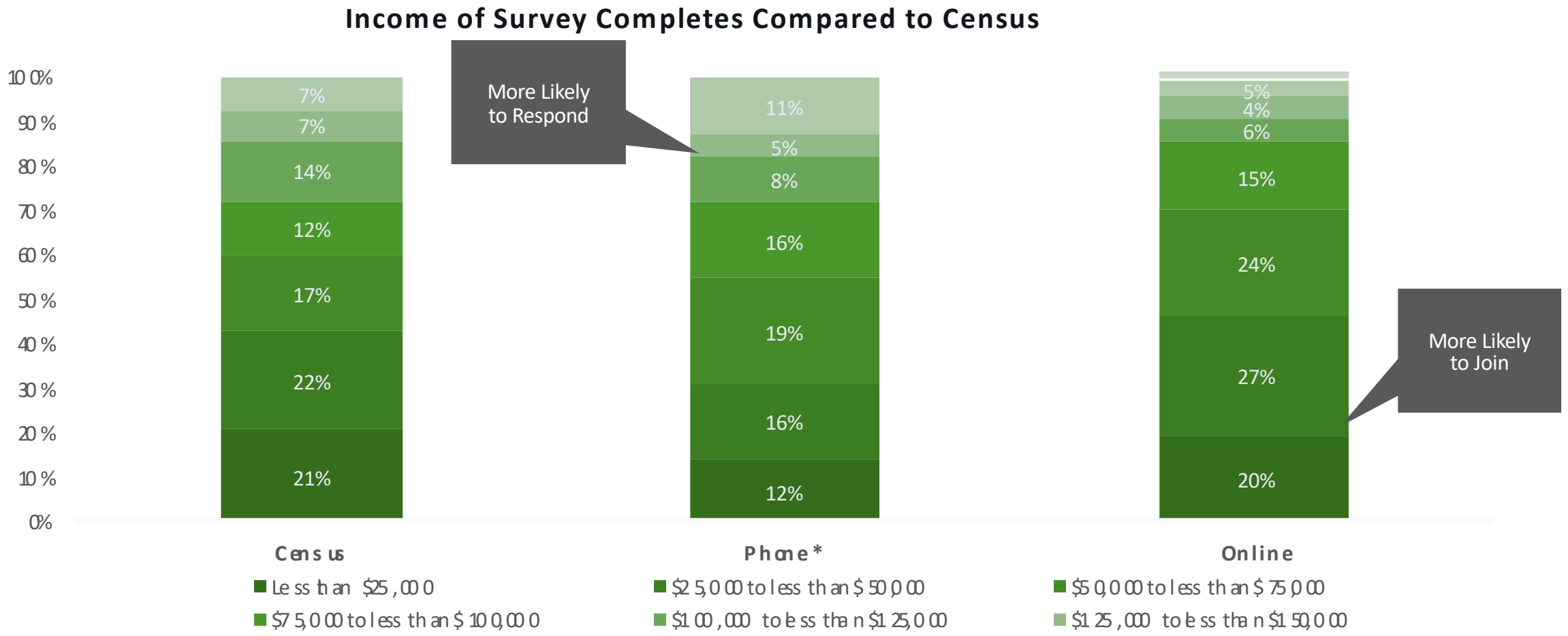
# Examining Bias in Voter File Between Collection Modes: Need to Quota on Age and Gender



# Examining Bias in Voter File Between Collection Modes: Mixture of Modes Gets Better Income Representivity



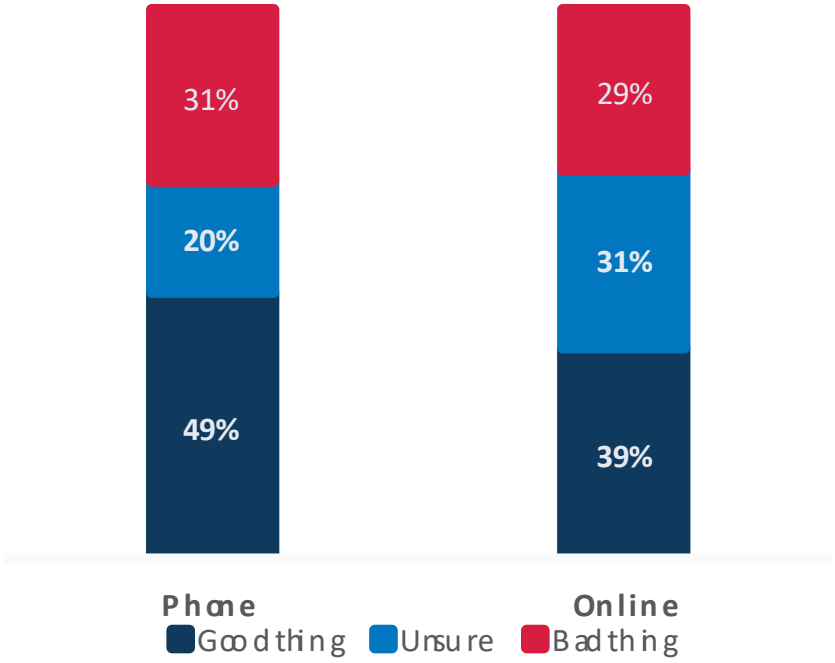



* 14% Item Non-Response

# Examining Bias in Voter File Between Collection Modes: Can be Important Mode Effects in the Questions

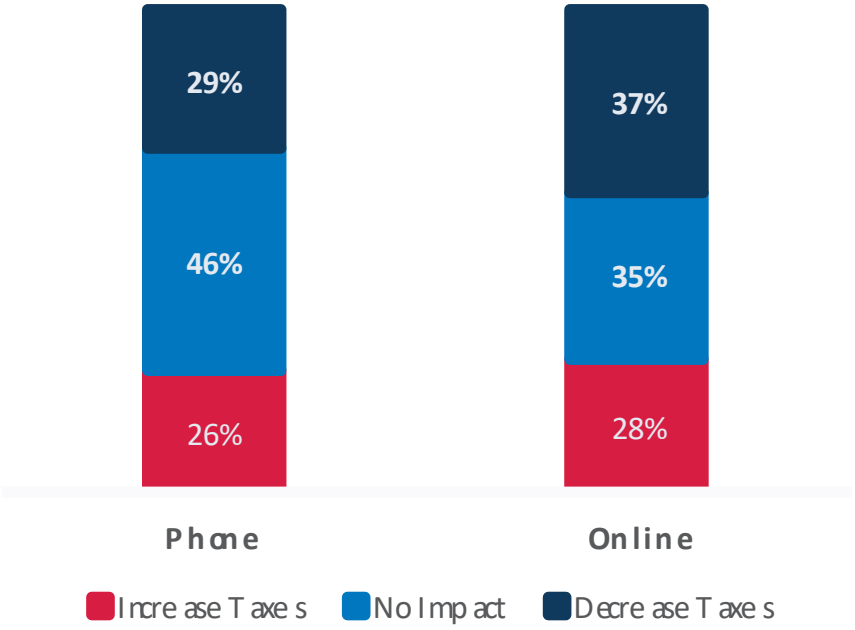
## No Stated Middle Ground

Would you say that the new federal tax law is, in general, a good thing or a bad thing for the country?*




## Explicitly Stated Middle Ground

Do you believe that the new federal tax law will increase your taxes, reduce your taxes, or will it not have much of an impact either way?**




* Answer options good/bad rotated in question/answer options. However "Unsure" not read in question, only seen in answer options.

** Answer options increase/reduce rotated in question/answer options. However, no impact always read.

# Examining the Accuracy of Likely Voter Models: Overall Research

**Problem:**

Elections can be great influenced by who turns out to vote. Lots of different models out there:

- Probabilistic Models or Cutoff Models
- Based on Stated Intention or Past Behavior

They can all lead to different results

Who's Really Ahead?

Alternative Likely Voter Methods In Same Poll Produce Very Different Results

| | All RV | Certain/ Probably vote | Certain to vote | Weighted by Self-Reported Voting | Voted 2014 (or age 18-20 and certain to vote ) |
|---|---|---|---|---|---|
| **Standard demographic weights** | | | | | |
| Doug Jones | 53% | 54% | 53% | 50% | 47% |
| Roy Moore | 45% | 45% | 47% | 47% | 50% |
| No answer | 3% | 1% | 1% | 3% | 3% |
| Margin | +8%D | +9%D | +6%D | +3%D | +3%R |
| *unweighted n=* | *1,559* | *1,386* | *1,193* | *1,505* | *1,028* |
| **Standard + 2016 vote weights** | | | | | |
| Doug Jones | 49% | 49% | 49% | 44% | 43% |
| Roy Moore | 47% | 49% | 49% | 52% | 53% |
| No answer | 4% | 2% | 1% | 3% | 4% |
| Margin | +2%D | 0% | 0% | +8%R | +10%R |
| *Unweighted n=* | *1,559* | *1,386* | *1,193* | *1,505* | *1,028* |

SurveyMonkey

Survey of registered voters in Alabama interviewed November 30 to December 6, 2017

**Which likely voter model is more accurate?**

**Research:**

Used Aristotle Voter File in Virginia and Alabama

Dual Frame RDD/Voter File Survey

- Three surveys (combined N=1,971)
- Tried to match RDD sample back to voter file on first/last name, zip, age.

**Compared 9 Methods:**

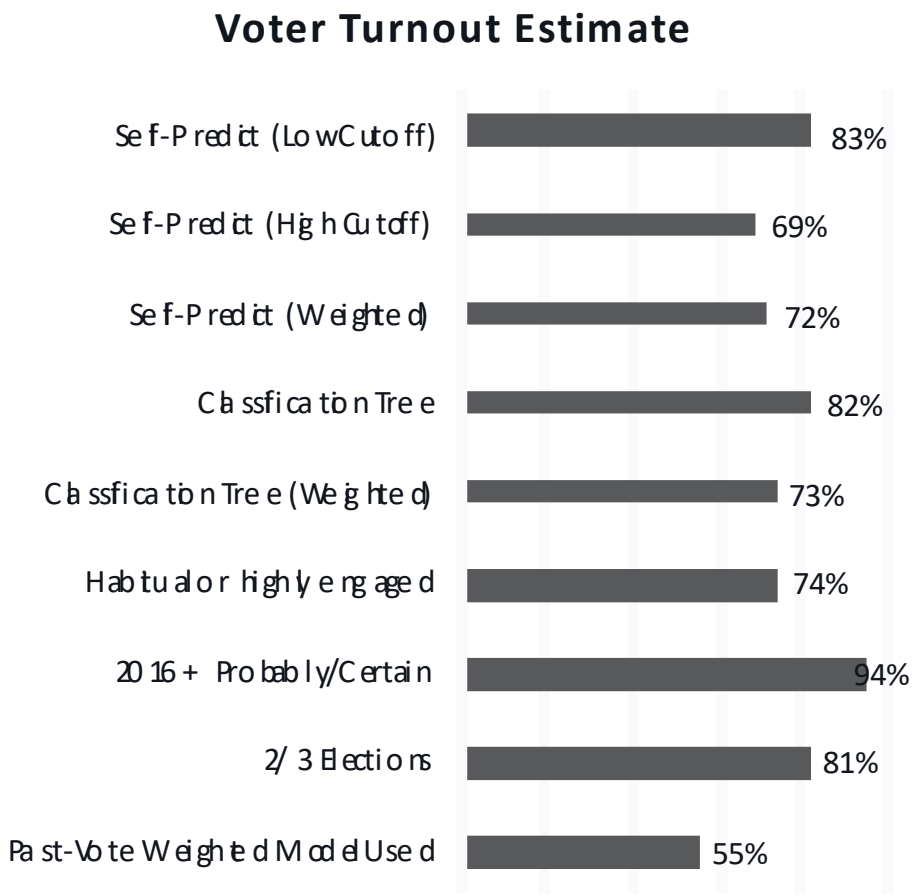
Cutoff Models:

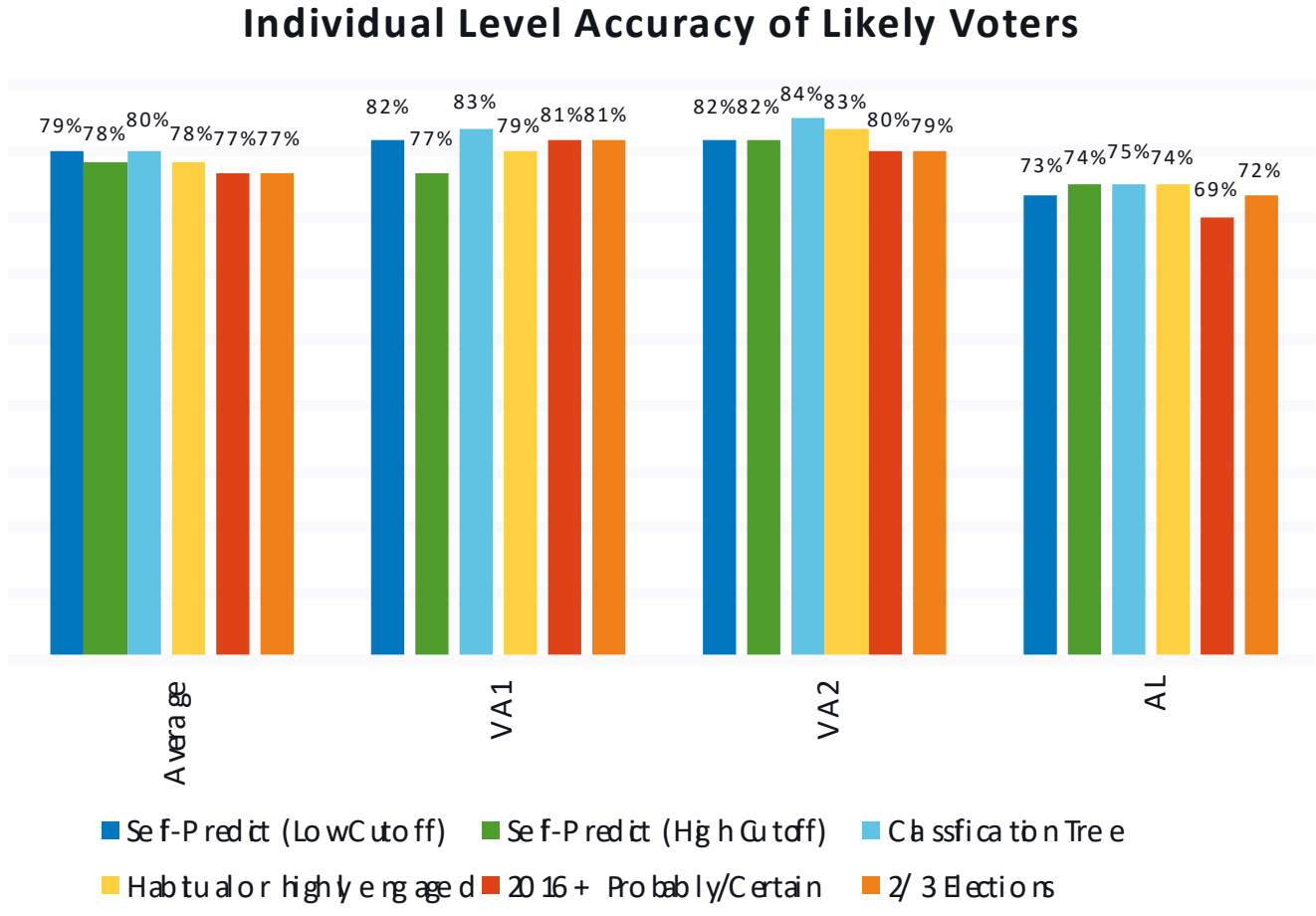
- Self-Predict (Low-Cutoff)
- Self-Predict (High-Cutoff)
- Classification Tree Cutoff
- Habitual or highly engaged
- 2016+Probably/Certain
- 2/3 Elections

Cutoff Models:

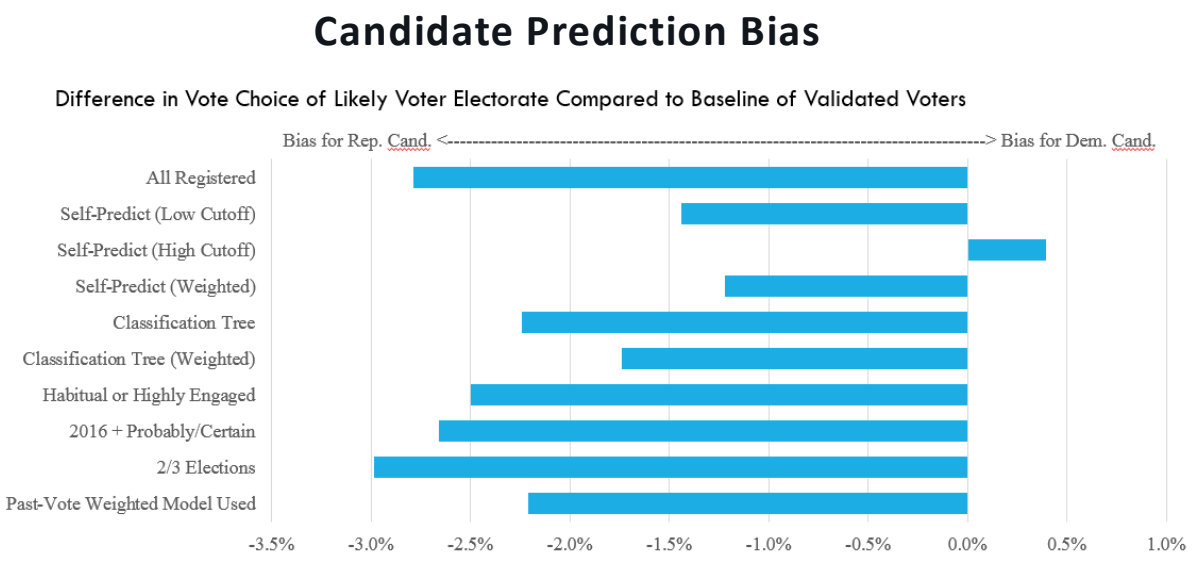
- Self-Predict (Weighted)
- Classification Tree (Weighted)
- Past-Vote Weighted Model

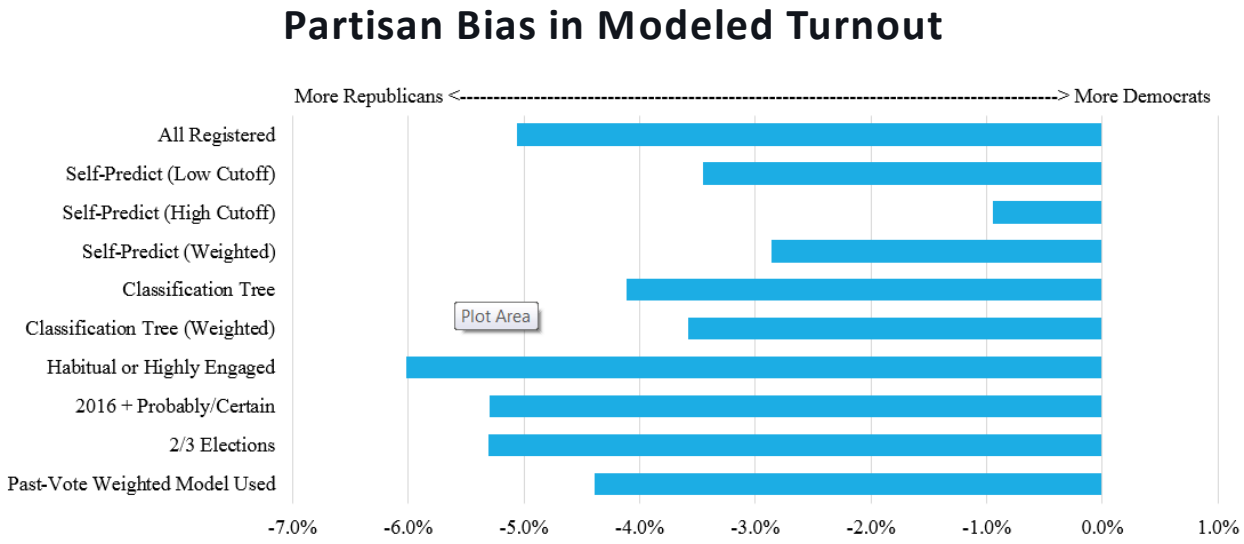
# Examining the Accuracy of Likely Voter Models: All Models Had Similar Individual Accuracy



# Examining the Accuracy of Likely Voter Models: Self-Stated Models Had Better Aggregate Estimates



# Relational validity: Overall Research

**Problem:**

A lot of new sampling techniques that don't rely on probability rules (non-probability samples).

We want to find a way to measure the ability to capture true bi-variate relationships (especially non-linear ones).

Earth Mover Distance (EMD) offers solutions:

- Allows for non-linear relationships
- Measures the amount of 'work' to bring back to known relationship
- Identifies the most efficient 'flow' to remove bias

**Can we use Earth Mover Distance as a proxy for bias measurement in non-linear bi-variate relationships?**

**Research:**

1) Simulated skewed distributions between age and party support.

25 samples of 1000 voters for each source

2) Applied technique to relationship between first election they voted in and partisanship

- Surveys from 2004-2017
- White registered voters
- 42 surveys
- 17 sample vendors
- 846,024 respondents

**Able to rank vendors better on ability to capture the bi-variate non-linear relationship between first election voted and partisanship.**

# Relational validity:
# Simulations show skewed sampling -> higher EMD

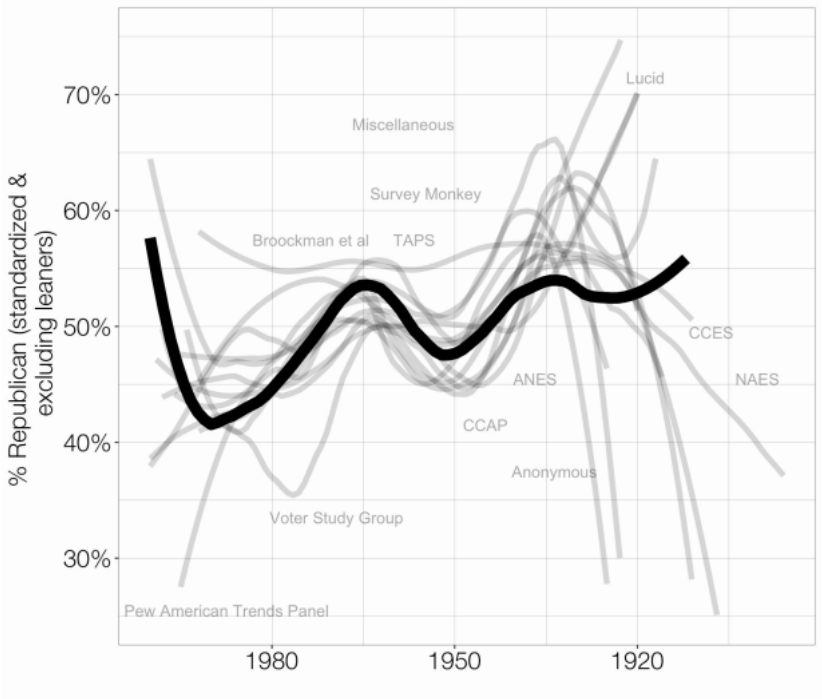
- Voting propensity tied to Age
  - Increases the EMD score between age and partisanship when **either** variable moves

**EMD Lower When No Sampling Skew**

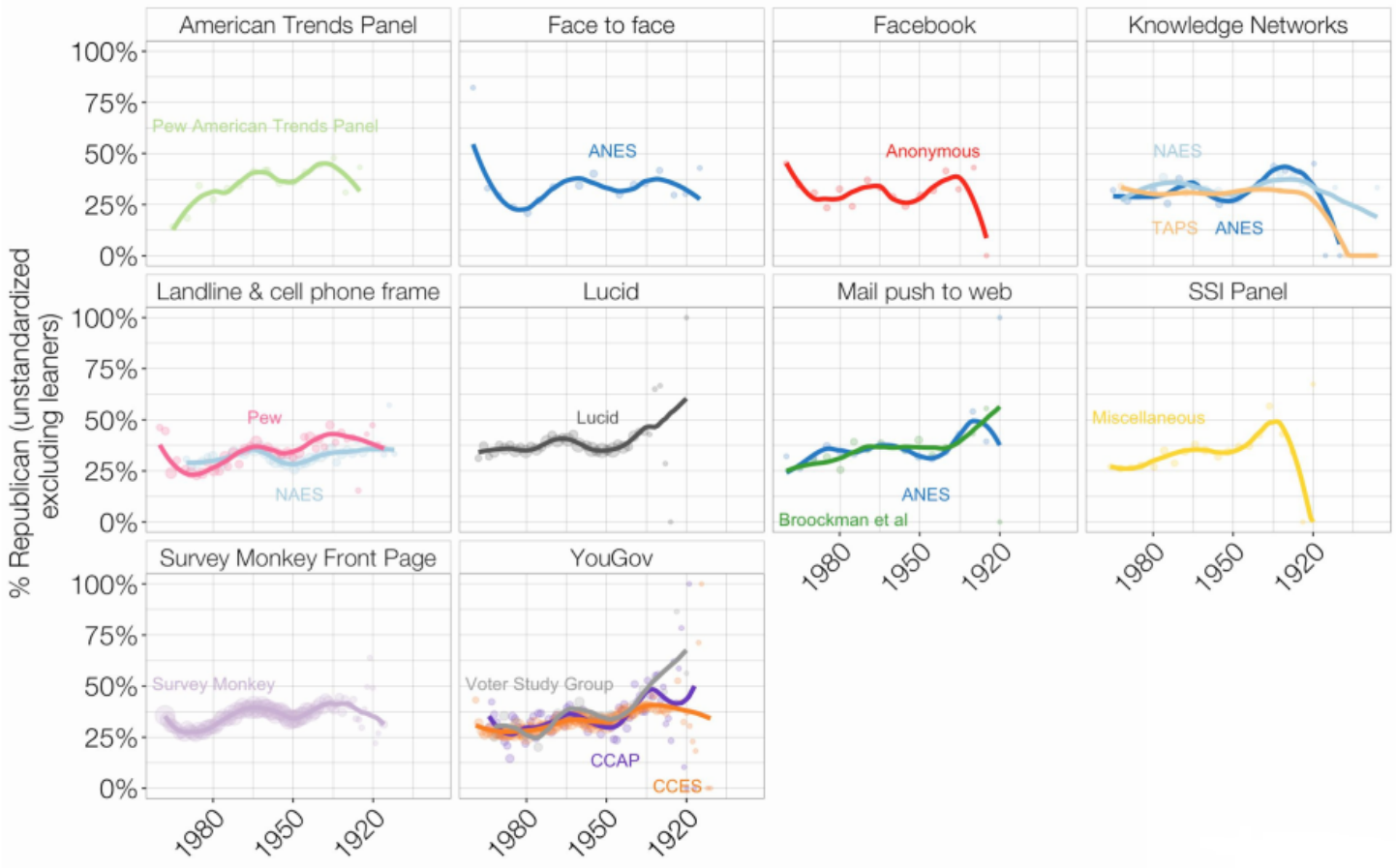
| | Over Sample Younger | No Age Skew in Sampling | Over Sample Older |
|---|---|---|---|
| Over Sample Active Voters | 6.49 | 6.20 | 6.21 |
| No Activity Skew in Sampling | 8.00 | 1.68 | 7.95 |
| Under Sample Active Voters | 7.76 | 7.73 | 7.07 |

# Relational validity:
# Clear Differences by Vendor in Capturing Benchmark

**Aggregate Probability Sampling Benchmark**




**Relationship by Vendor**



# Relational validity:
# Final Vendor and Technique Rankings

| Project | EMD (centered) | Diff. from Top Rank | EMD rank (centered) |
|---|---|---|---|
| Survey Monkey - Survey Monkey Front Page | 0.664 | 0.000 | 1 |
| Anonymous - Facebook | 1.268 | 0.604 | 2 |
| CCES - YouGov | 1.491 | 0.827 | 3 |
| NAES - Knowledge Networks | 1.593 | 0.929 | 4 |
| Pewl - American Trends Panel | 1.752 | 1.088 | 5 |
| ANES - Mail push to web | 2.151 | 1.487 | 6 |
| ANES - Knowledge Networks | 2.459 | 1.795 | 7 |
| Miscellaneous - SSI Panel | 2.659 | 1.995 | 8 |
| CCAP - YouGov | 2.736 | 2.072 | 9 |
| Broockman et al - Mail push to web | 2.765 | 2.101 | 10 |
| Lucid - Lucid | 3.269 | 2.605 | 11 |
| TAPS - Knowledge Networks | 4.788 | 4.124 | 12 |
| Voter Study Group - YouGov | 4.892 | 4.228 | 13 |

# Verifying Voter Registration Records: Overall Research

**Problem:**

Many people take voter file data as truth, but there is error in administration records that is normally ignored.

Here are some of the sources of potential error:

- Concept validity
- Measurement error
- Processing error
- Temporal error

**How much error is expected in voter file records?**

**Research:**

Used Florida state record file with L2 appends

Sampled 60,000 numbers from FL February, 2017 Voter file

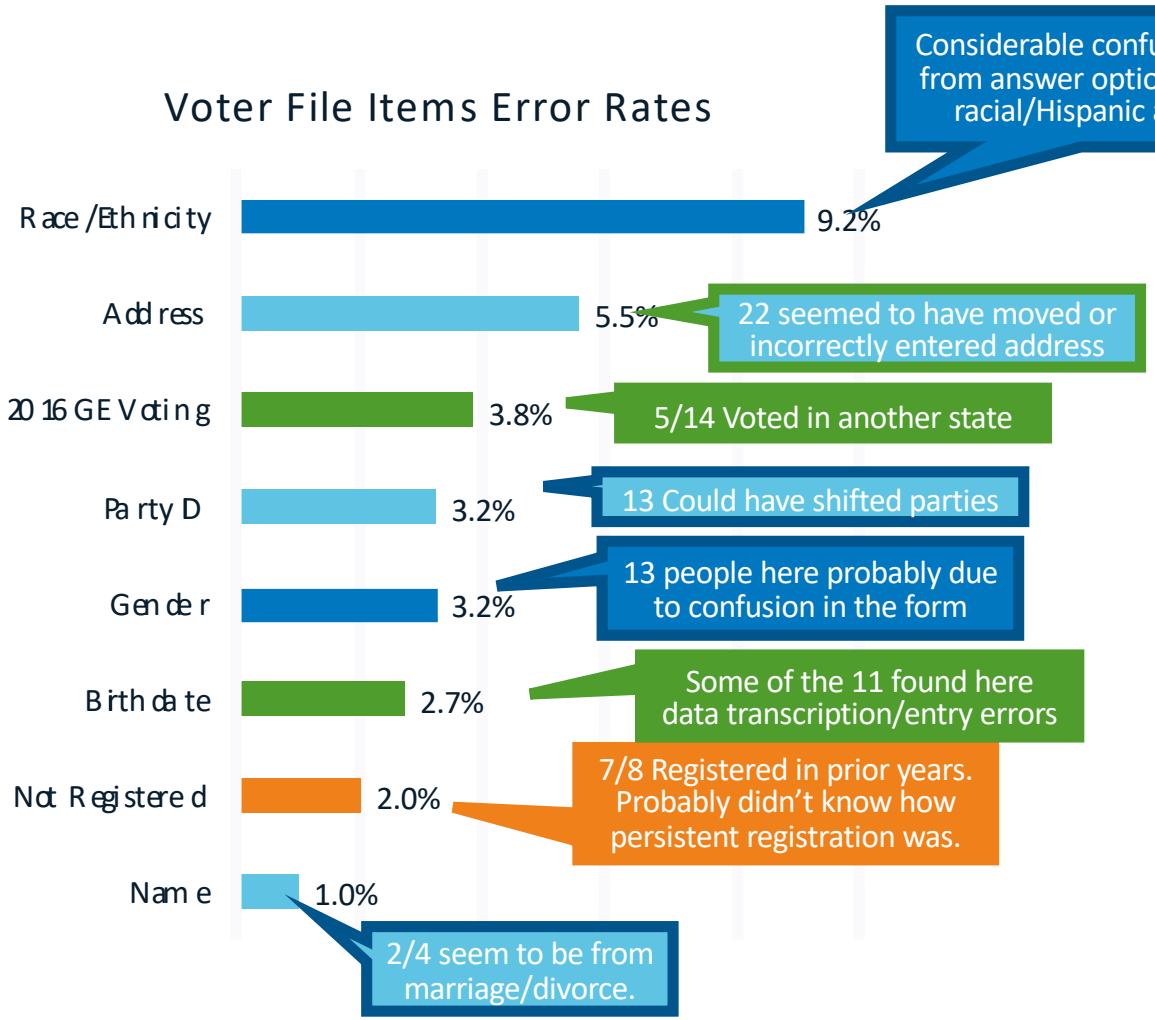
- L2 appended phone numbers (59.8%)

Conducted Telephone interviews (N=401)

- 6.5% Response Rate
- Asked Voter Registration question
- If different than what was on the Voter File
  - Got correct information
  - Asked reason for the discrepancy

**Substantial error rate (17%) in at least one field on file.**

# Verifying Voter Registration Records: Small, but Significant Errors in Voter File




There in non-negligible error in voter records.

Many voter file vendors will clean these up.

- Aristotle
- Catalist
- Data Trust
- i360
- L2
- Target Smart

Varies quite a bit from state to state.

Voter Registration forms need to be updated for easier use (treat like a survey).

# Online Panels Matching to Voter Files: Overall Research

**Problem:**

Many surveys moving to online including political for cost savings and increased capabilities.

However, online panels tend to have low coverage of the overall voter file. This can lead not only to potential coverage bias but low feasibility. To improve feasibility some have allowed household matches rather than just individual matches to be used in the online research.

**How does the full voter file compare to either individual or household matches in an online panel sample?**

**Research:**

Used National Aristotle file

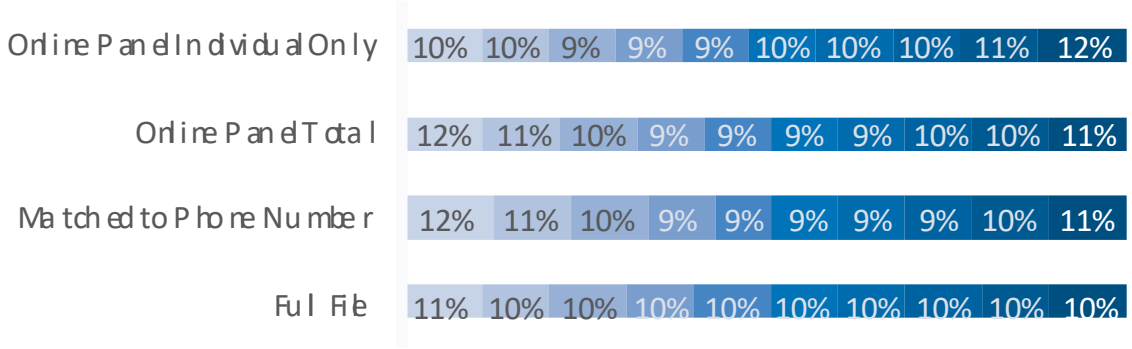
Compared the following sample frames

- Full Voter File
- Voter File with Phone Number
- Online Panel Individually Matched
  - Full Name, Address (Zip+4), Year of Birth* (I)
  - Full Name, Address (Zip+4), Year of Birth (S)
  - Full Name, Zip Code, Year of Birth (M)
  - Last Name, Address (Zip+4)
- Online Panel Household Matched
  - Last Name, Address (Zip+4)

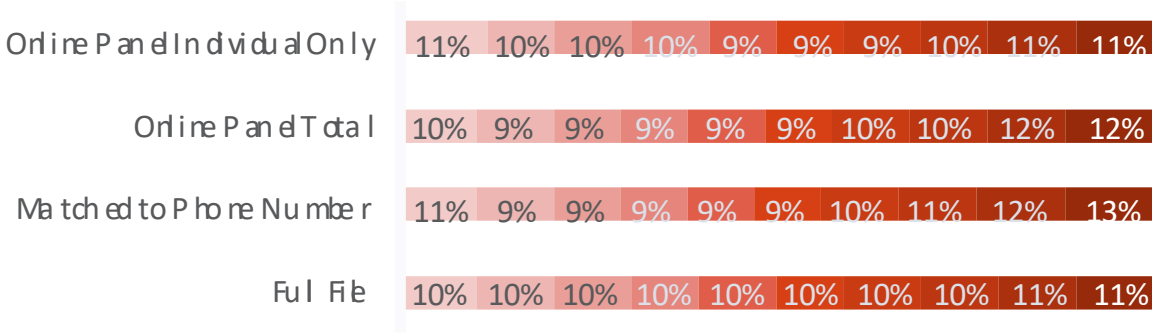
**Don't need quotas on partisanship, but maybe on Vote Propensity.**

# Online Panels Matching to Voter Files: Political Ideology not Skewed in Online Panel

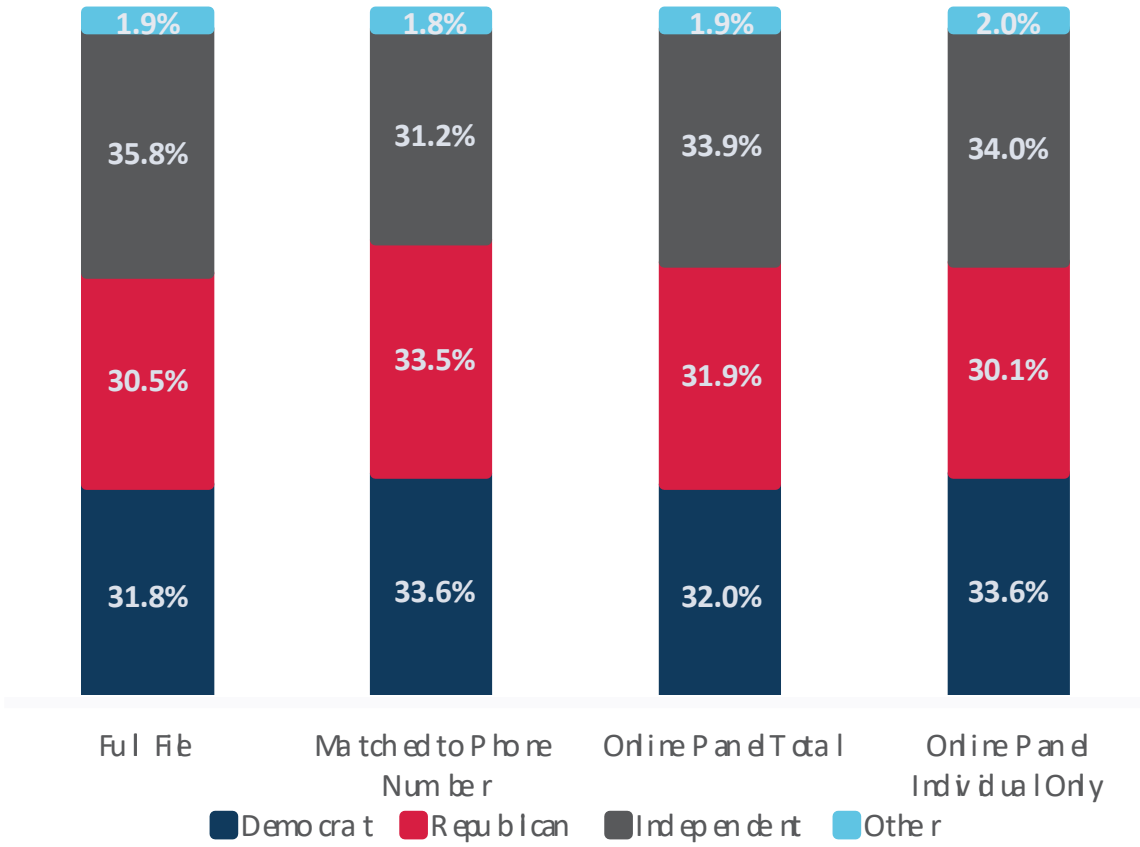
## Democrat Affinity Score Comparison



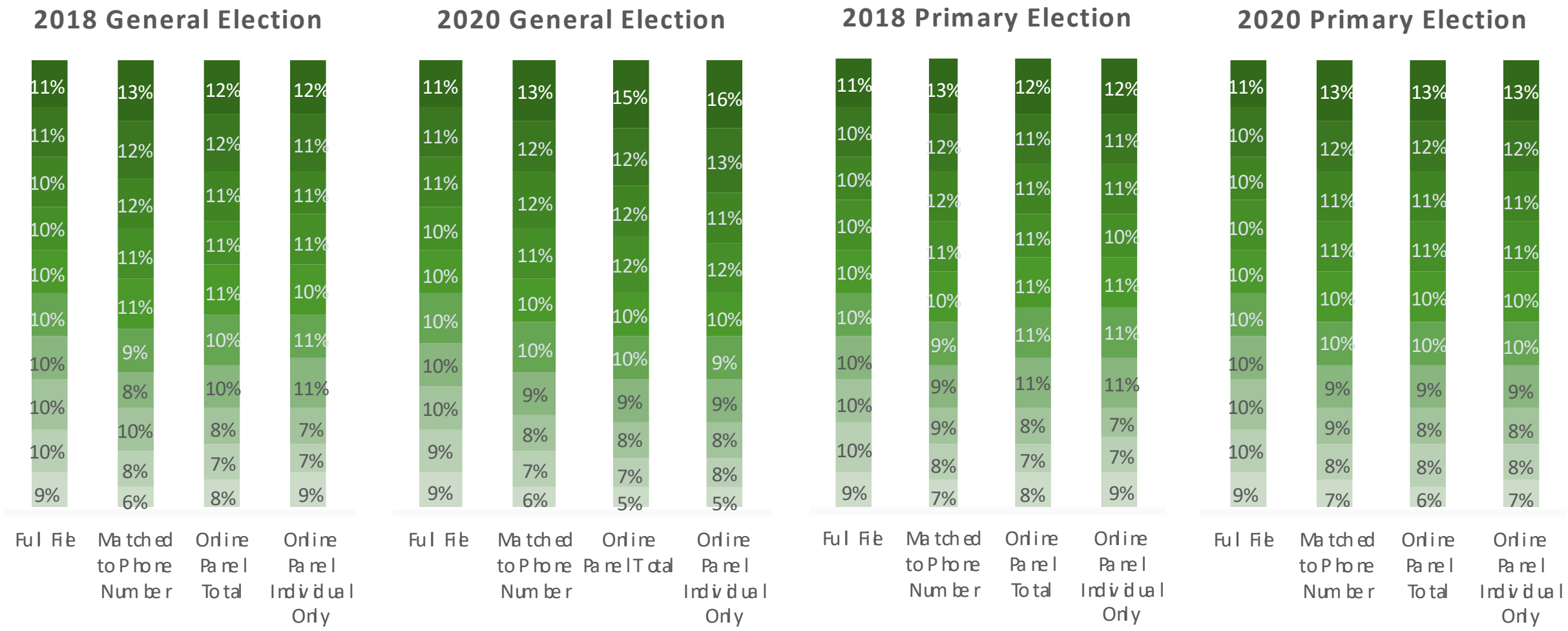

## Republican Affinity Score Comparison




## Party Code Distribution in Matched Frame



# Online Panels Matching to Voter Files: Online Panels a Little More Engaged



# Final Thoughts/Considerations

Remember to respect privacy when using Voter Files

Data will get more and more integrated (Ad Tracking, Geofencing, ect.)

Passive/Big Data from Voter Files likely to supplement, not replace surveys techniques

Continue to watch for skews in coverage and response rate for potential bias

Never going to be perfect, will continue to improve