How measuring demographic variables could be the next data and replication crisis

Ben Leff, Co-Founder, Verasight

Demographic Variables in Survey/Public Opinion Data

- Accurate demographics (e.g., age, gender, race, income) are key for survey analyses, survey targeting, and weighting

- We will show 3 examples of common measurement approaches that lead to misleading conclusions
  - Modeled demographics from voter/consumer files
  - ChatGPT/AI
  - Survey routers/exchanges

- Strategies for accurately measuring/validating demographics

Example 1 – Modeled Demographics from voter file/consumer databases

Example One: Using a prominent voter/consumer file to model demographics

Why do researchers use modeled demographics?

- Create sampling frames to target specific audiences
- Reduce number of questions asked in survey

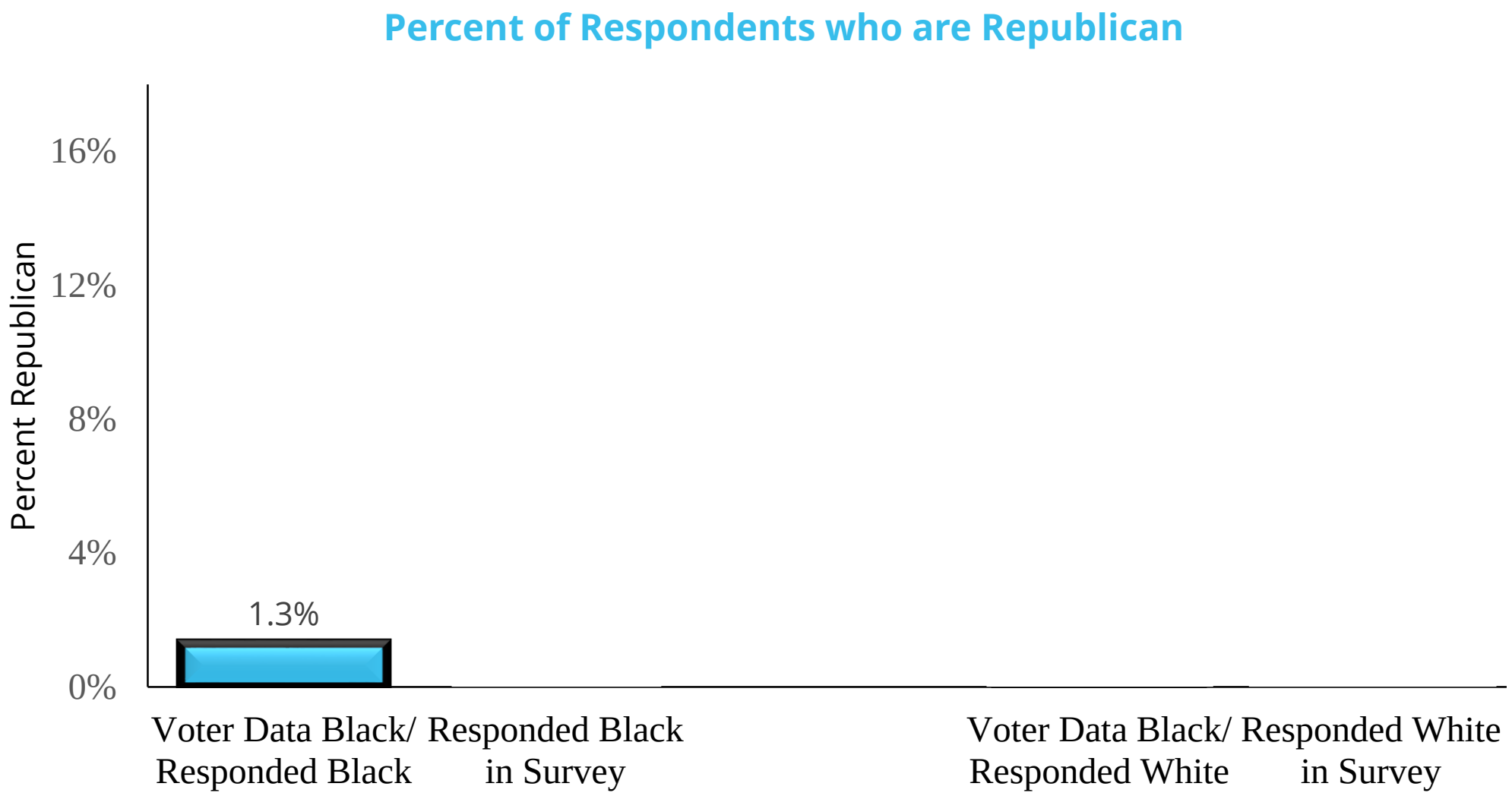
How did we test accuracy of modeled demographics?

We conducted a large survey of residents in a midwestern city and county and compared survey responses to modeled demographics

How did modeled demographics perform?

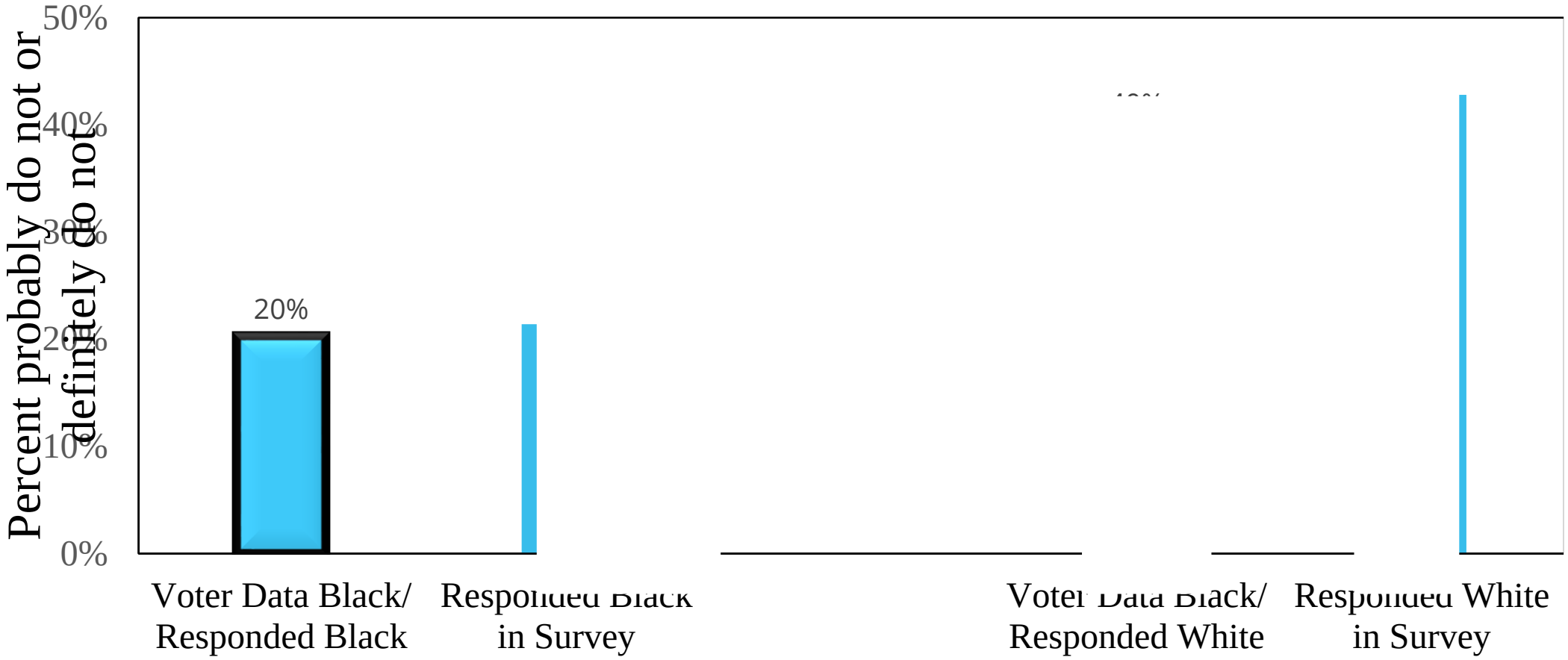
- 33% of respondents' self-reported race did **not** match the racial category modeled
- The largest mismatch was respondents coded as African American in the data self-reporting as white

Modeled Demographics vs. Self-Reported Demographics



Modeled Demographics vs. Self-Reported Demographics

Do Police Use More Force Against Black respondents?



Why this matters?

- When demographic profile data are included, researchers rarely ask the same demographics

  Analyses will use problematic data

- These demographic data are used to generate sampling frames to target specific respondents

  Campaigns, companies, and researchers are sampling the wrong respondents

Example 2 – ChatGPT

Can Open AI better predict race than modeled demographics?

Research Plan

- We tasked OpenAI's "Data Analyst" GPT with predicting racial identification
- Provided: First Name, Last Name, House address, and Demographics (age, gender, education level, income level, etc.)

The Result

Classification error was greater than 33%

Example 3 – Survey Marketplaces and Routers

Are you the parent or guardian of any children under the age of 18 living in your household?

Yes

No

2% Complete

NEXT

Please list the age and gender for each child living in your household under the age of 18.

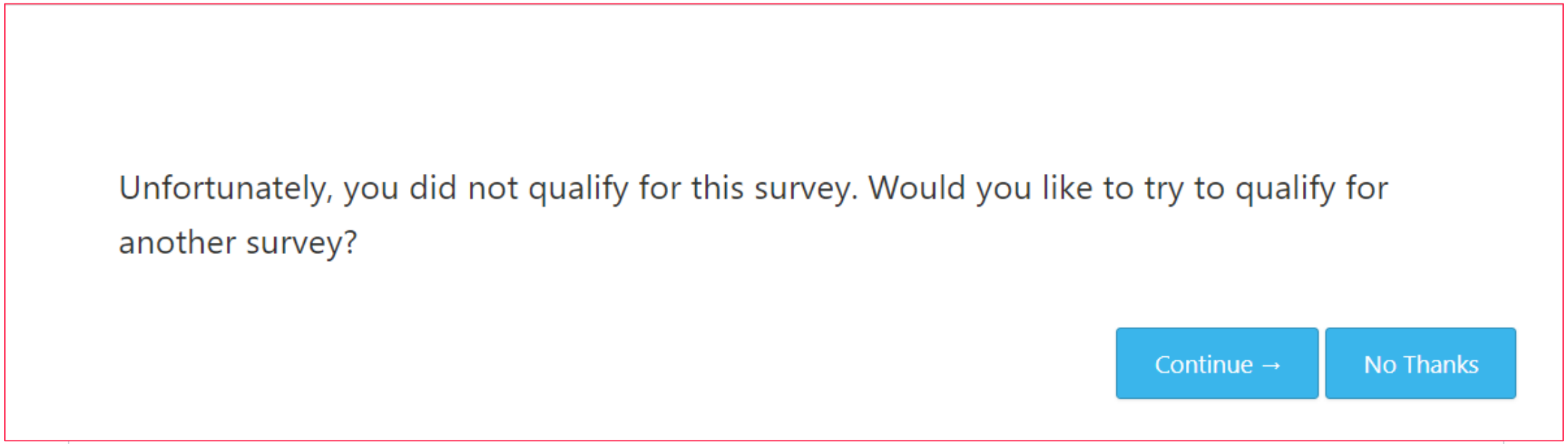
Unfortunately, you did not qualify for this survey. Would you like to try to qualify for another survey?

Continue → No Thanks

2% Complete

NEXT

Are there any children in your household for whom you have sole or joint care responsibility?
Please select all that apply



- Yes aged between16-17
- No

Unfortunately, you did not qualify for this survey. Would you like to try to qualify for another survey?

Continue → No Thanks

To begin with, please tell us a little bit about yourself.

Unfortunately, you did not qualify for this survey. Would you like to try to qualify for another survey?

Continue → No Thanks

- ☐ Boy age 3
- ☐ Girl age 3
- ☐ Boy age 4
- ☐ Girl age 4
- ☐ Girl age 8
- ☐ Boy age 9
- ☐ Girl age 9
- ☐ Boy age 10
- 13
- ☐ Male teen age 14
- ☐ Female teen age 14
- 16
- ☐ Male teen age 17
- ☐ Female teen age 17

These screeners create incentives for inaccurate information

- **Inattentive, careless responses** can produce incorrect demographic information
- Even more problematic, **respondents may misrepresent demographic information** to qualify for surveys

Conclusions

Demographic variables are central to survey research, and not nearly as easy to measure as it might seem

- Know where profile variables come from and how they are validated
- Ask respondents directly for demographic information
- How you ask matters!
  - Routers and Screeners, which are ubiquitous in survey research, create perverse incentives

Passive Verification

Ask Respondents Directly at **Time 1** → Re-ask respondents in 1 – 2 months → Flag respondents with inconsistent answers

Advantages of this approach

- A sensible answer in one survey might not be sensible if taken in consideration with respondents' previous answers
- Bots/professional survey takers unlikely to remember the demographics that they previously gave
- Disingenuous survey takers who intentionally change demographic answers to try to qualify will be flagged

The Verasight Difference

100% Verified Panel (dual authentication using verified US cell phone numbers)

Ask respondents directly… then ask again (ongoing passive authentication)

96% report the same ZIP code

97.3% report the same education level

98.4% report the same gender identification

Free Survey Question and Giveaway Entry

Free Nationally Representative Survey Questions

- Researchers receive toplines, crosstabs, survey weights
- Sample size of at least 1000

Giveaways for entering question: