BIGSURV CONFERENCE| OCTOBER 2018

# IS MORE DATA "BETTER DATA"?

## ASSESSING THE QUALITY OF COMMERCIAL DATA APPENDED TO AN ADDRESS-BASED SAMPLING SURVEY FRAME

MAKING RESEARCH RELEVANT

Rebecca Medway| Nicole Guarino| Carol Wan | Danielle Battle | Michael Jackson



AIR®

# Motivation

- Interest in appending new data sources to the U.S. National Household Education Survey (NHES) address-based sampling frame
- Overarching goal: assess utility for targeted and adaptive designs, sampling, weighting, etc.
- Motivated by research into utility of data already available on NHES frame
  - Ability to predict response outcomes and key estimates for NHES somewhat limited[1]
  - New data source offers many more variables on a wider variety of topics
- Today we will talk about step 1: assessing general quality and cost of the data

1. Jackson & Medway (2017); Jackson & McPhee (2017); Jackson, Steinley, & McPhee (2017)

# Research Questions

1. **Cost:** What are the costs associated with using the new data?

2. What is the **quality** of the new data?
   - Breadth
   - Coverage
   - Accuracy

3. **In comparison to**: For the above questions, how does this compare with the data already available on the NHES frame?

# Data: NHES

- Household survey that provides descriptive data on the educational activities of the U.S. population
- Sponsored by National Center for Education Statistics (NCES)
- Uses an address-based sample
- Screener phase is used to sample a child about whom an adult reports
- Paper-only since 2012, transitioning to web-push mixed mode
- Using data from two most recent administrations:
  - 2016: last official administration (n=205,000)
  - 2017: web test (n=97,500)

# Data: Commercial Data

| | Existing | New | |
|---|---|---|---|
| **Unit** | Address-level data | Person-level data | |
| **Number of variables** | About 20 | Over 200 | |
| **Type of variables** | Basic demographics (HH, HoH) | Voting-related | Consumer-related |
| **Matching procedures** | Proprietary | 5 match attempts – exact match, then 4 lesser (e.g., city differs, ZIP differs) | |
| **Timing** | Same as sample draw | 1-2 years after sample draw | |

# Cost: File Review and Preparation

- New data cheaper than existing source; however, both files quite inexpensive to purchase

- New file requires much more extensive processing
  - Many more variables
  - Person-level data → address-level data

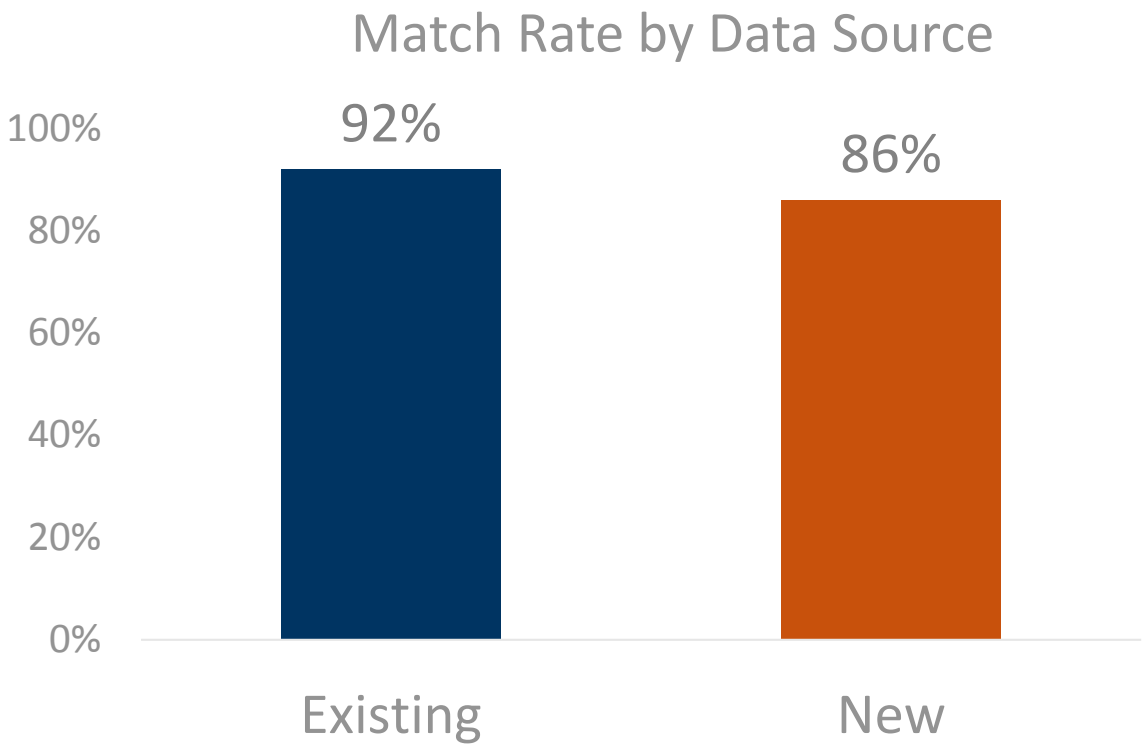
- Examples of file preparation tasks:
  - To review 800,000 person-level matches: established and applied rules for identifying and removing suspicious matches
  - To go from person-level to address-level data: established and applied aggregation rules for about 200 variables

# Quality: Match Rate

- **Match rate** = percentage of sampled addresses for which any appended data is available
  - **Existing data**: at least one variable is populated for address
  - **New data**: at least one person-level record matched to address with at least one variable populated

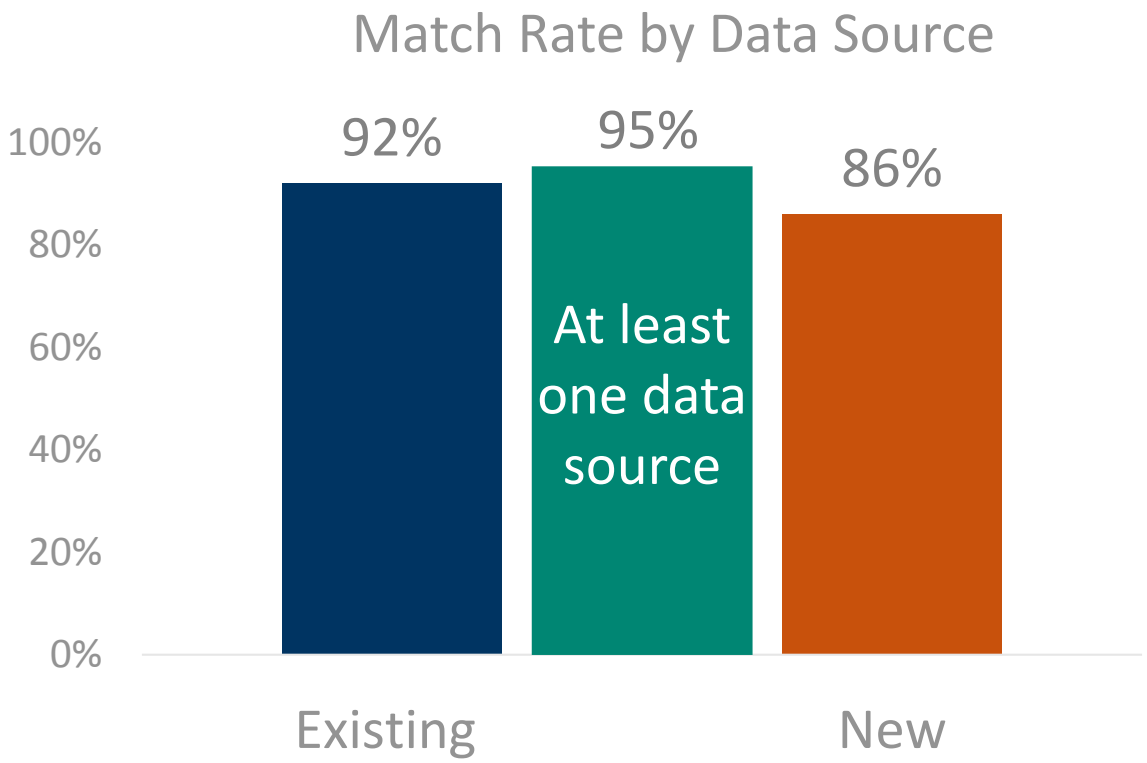
# Quality: Match Rate

- Almost all NHES addresses match to existing data source.
- Though still relatively good, new data source match rate was lower.
  - Most addresses had 1-2 person-level matches



# Quality: Match Rate

- Almost all addresses that matched to new data source:
  - Also matched to existing data.
  - Came from first, strictest match attempt (98%).



# Quality: Match Rate

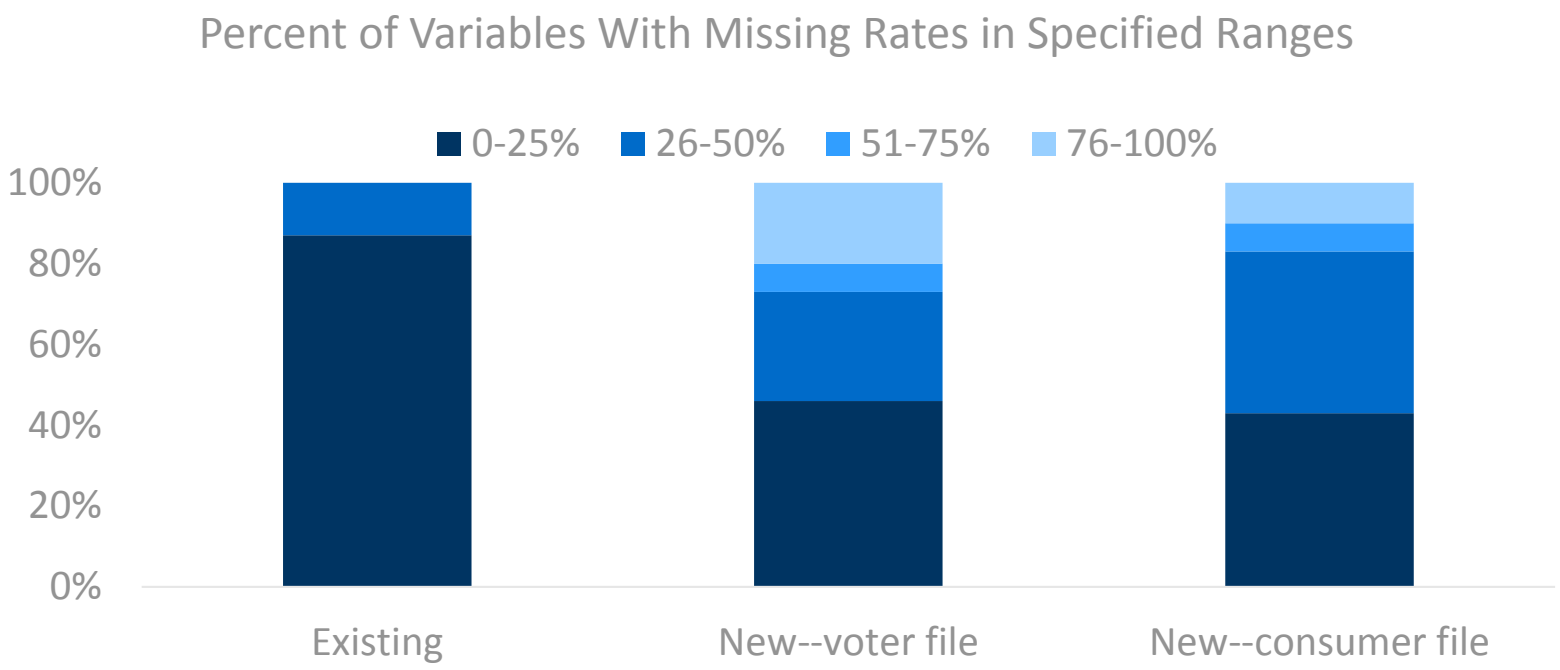
- Existing data source match rate higher than new data source for all subgroups examined (by 3 to 8 percentage points)
- Both data sources relatively less successful at matching for some types of addresses than for others:
  - High poverty areas
  - High minority areas
  - Areas with low concentrations of children
- No impact on match rate: survey year, urban/rural

# Quality: Missing Data Among Matched Cases

- New data source offers many more variables than existing one – but to what extent is data missing among matched NHES addresses?
- Limited to variables where can definitively determine “missing”
  - Existing data: 16 variables in 2016; 15 in 2017
  - New data - voter file: 45 variables
  - New data - consumer file: 30 variables

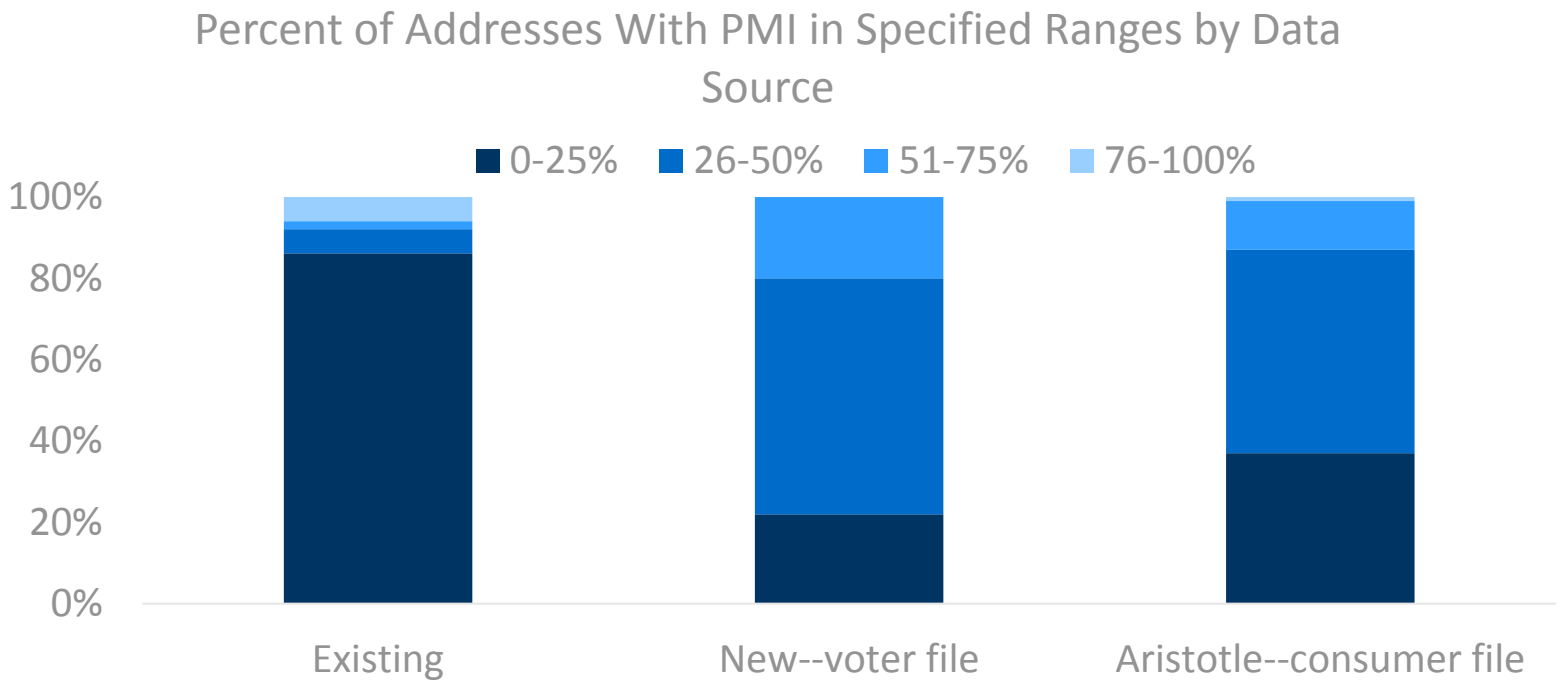
# Quality: Missing Data Among Matched Cases

- **Item missing rate**: percentage of matched addresses without info available for that variable
- Variables from new data source more likely to have extensive missing data
  - However, both data sources have a similar **number** of variables low missing rates



# Quality: Missing Data Among Matched Cases

- **Percentage missing information**: percentage of *variables* for which *matched address* is missing data
- Matches to new data source more likely to be missing data for many variables



# Quality: Agreement Between Commercial Data and NHES Responses

- Identified variables on commercial data files that were also captured on NHES
  - Calculated agreement rate and Kappa statistic for each variable
- Wide range in agreement of commercial data with NHES responses

**Agreement Rate with NHES Responses**

| | Existing | New |
|---|---|---|
| Anyone in household age 65+ | 86% | 89% |
| Anyone in household age 18-64 | **72%** | **89%** |
| Owns home | 86% | 88% |
| Hispanic household | 83% | 83% |
| Any children in household | **73%** | **67%** |
| Household income (categorical) | 45% | 46% |
| Number of people in household | 35% | 35% |

# Conclusions and Next Steps

- Findings for both data sources similar to findings from other studies.
  - Data not available for all addresses
  - High missing rates for some variables
  - Variation in quality of data across variables as compared to self-reports
    - » Lower quality: child presence indicator
- Though new data source is not perfect, it offers several potential benefits
  - Many more variables on a wider variety of topics
  - Adds data about 3% of addresses where we previously had nothing
  - “Opportunity” to evaluate quality more thoroughly
- Therefore, we will evaluate its utility for weighting, propensity modeling, targeted mailings, etc; this work is in progress

REBECCA MEDWAY

SENIOR SURVEY METHODOLOGIST

202.403.6369

RMEDWAY@AIR.ORG

MAKING
RESEARCH
RELEVANT

THANK YOU

# References

Jackson, M. and McPhee, C. (2017). *NHES:2016 Tailored Incentive Experiment Report.* Washington DC: Internal report, American Institutes for Research.

Jackson, M. and Medway, R. (2017). *NATES:2013 Nonresponse Bias Analysis Report: Evidence from a Nonresponse Follow-up Study*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jackson, M., Steinley, K., and McPhee, C. (2017). *What will work for whom? Identifying subgroups for which a higher monetary incentive will be effective.* Paper presented at the 2017 American Association for Public Opinion Research Conference.