

AAPOR | MAY 2018

**DO YOU “VOTE” FOR WEB OR PAPER?
LEVERAGING VOTING DATA TO PREDICT
RESPONSE BY WEB IN A MIXED-MODE (WEB-
PUSH) SURVEY**

Cameron McPhee
Mahi Megra | Michael Jackson | Rebecca Medway

MAKING
RESEARCH
RELEVANT

Background

- Mixed mode survey designs that include web surveys are becoming increasingly ubiquitous
 - Help control cost
 - Potentially increase representativeness
- Mixed mode designs still tend to achieve lower final response rates compared to paper only designs (Dillman et. al., 2014; Messer & Dillman, 2011)
- Experimentation with ASD and RSD suggests different modes can be leveraged for different respondents to improve response rates and representativeness.

Background

- Previous research suggests address-level auxiliary data can be moderately effective at modeling mode preference (McPhee & Jackson, 2017)
 - Too much error in predictions to be effective out of sample
 - Likely not strong enough correlation between previously available auxiliary data and response mode preference
- Voting history data shown to be predictive of survey participation in general (Tourangeau et al., 2010)
- Known correlation between likelihood to vote and participation in online panels (e.g., see Baker et al., 2013)

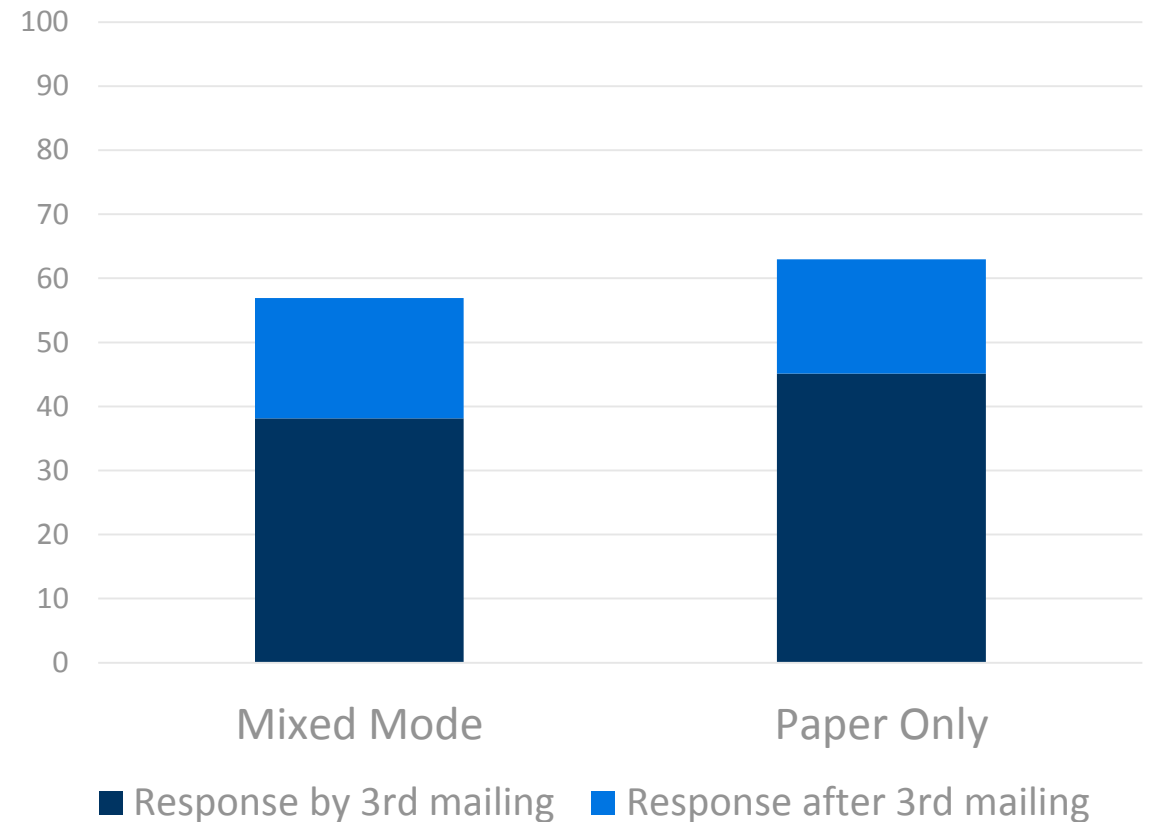
Research Questions

- 1. Are individuals who would respond under a paper only treatment, but not a mixed mode treatment identifiable through a predictive model in advance of data collection?**
- 2. Can this model be effectively used to tailor survey administration mode in order to maximize response rates?**
- 3. Do voting history data and other extensive commercial data improve the utility of the model?**

The National Household Education Survey (NHES)

- Large-scale (n ≈ 200,000) household survey sponsored by National Center for Education Statistics
- Typically administered as a two-phase mailed paper survey (using ABS sampling frame)
- 2016 administration included a mixed-mode web “push” experiment (n ≈ 35,000)
- Focus here primarily on screener (phase 1) response

NHES:2016 Screener Response Rates by Treatment



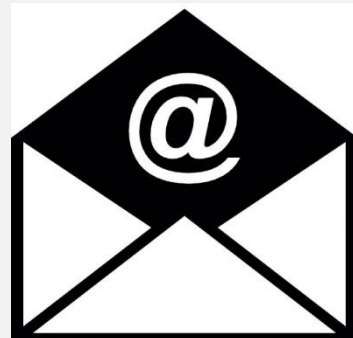
Web Push vs. Paper only Mailing Protocol

Mixed-Mode

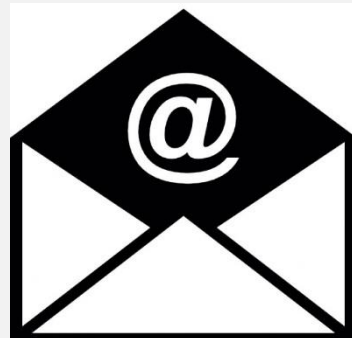
Advance Letter



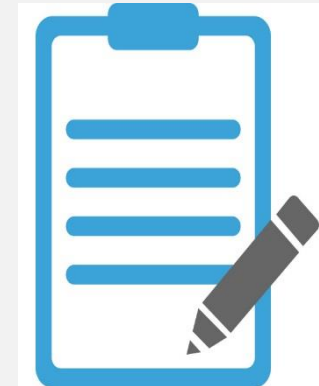
1st Mailing



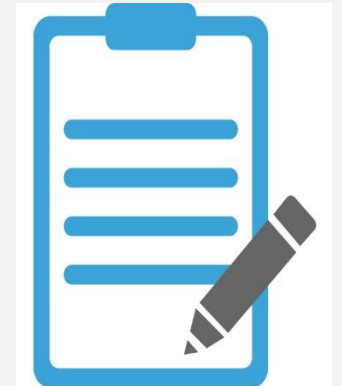
2nd Mailing



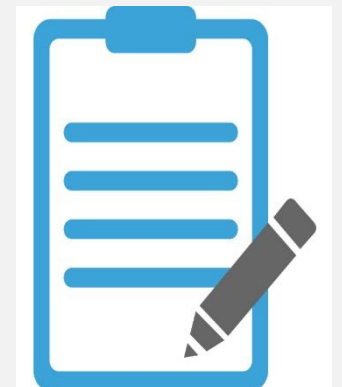
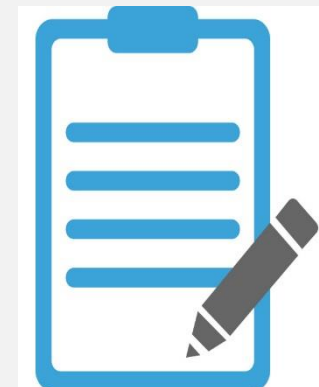
3rd Mailing



4th Mailing



Paper-only



Methodology - Variables

- Original Frame variables
 - USPS address characteristic variables
 - Appended address-level demographic variables
 - ACS/Decennial Census Block-group variables from Census Planning Database (PDB)
- Aristotle Voting and Commercial data (250+ variables)
 - Person-level voting history variables
 - Person-level commercial data

Methodology – Data Reduction

- Used conditional random forests (*cforest* in R Party Package) to select variables most predictive of each outcome under each treatment (mixed mode and paper only)
 - Separate models for Frame-only and Frame + Aristotle data
- Selected 25 variables identified as most important from the forest to include in final logistic regression models
 - included union of mixed mode and paper variables

Methodology – Logistic Regression Models

Steps for each model

1. Run binomial logistic regression using most important predictors from forests
 - All variables interacted with treatment (Mixed-mode vs. Paper)
2. Eliminate non-significant main-effects (and related interactions)
3. Calculate paper response propensity and web response propensity for each case (by using treatment indicator)
4. Calculate each case's **“Sensitivity score” = Paper response propensity – Web response propensity**
5. Create four sensitivity quartiles (least sensitive to most sensitive)
6. Calculate each sensitivity quartile's **“Paper treatment effect” = Paper response rate minus mixed-mode response rate**

Methodology – Logistic Regression Models cont.

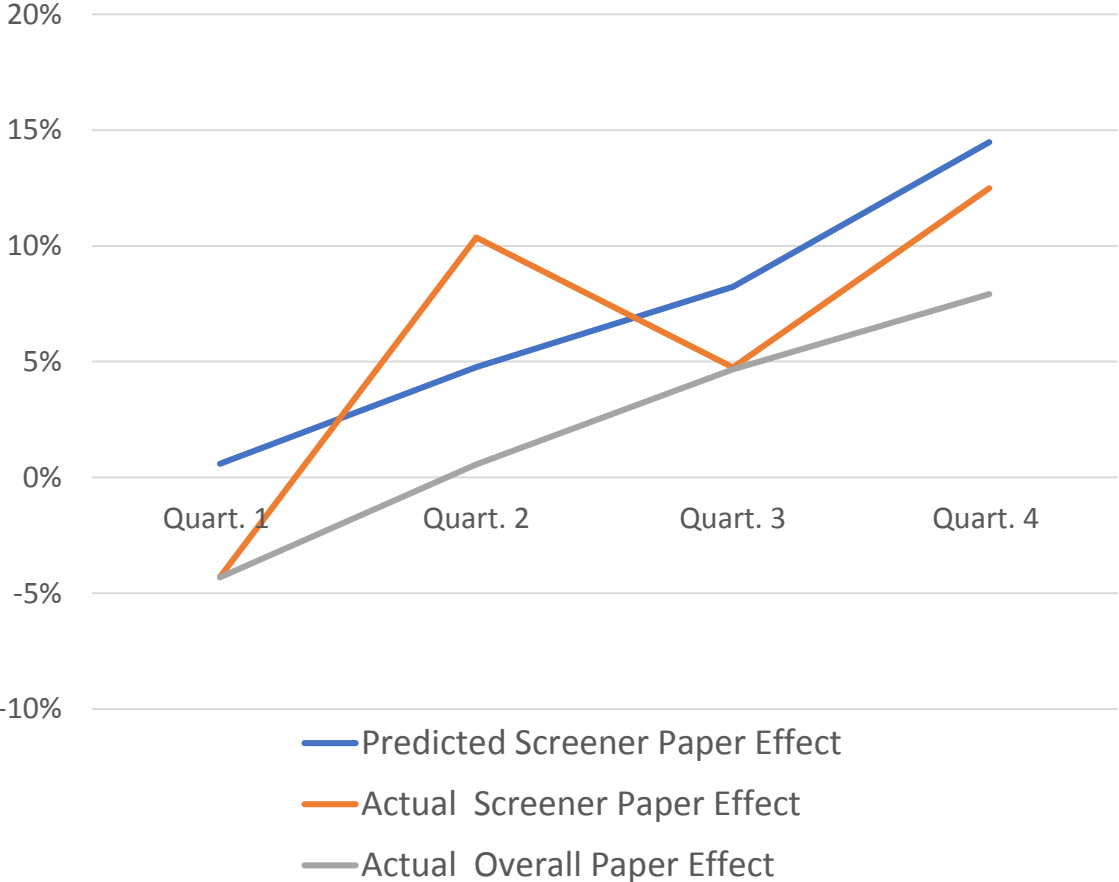
Four separate models

Independent Variables	Dependent Variables	
	Response after 2 nd mailing	Final screener response
Frame Variables Only	Mode sensitivity	Protocol sensitivity
Frame + Aristotle Variables	Mode sensitivity	Protocol sensitivity

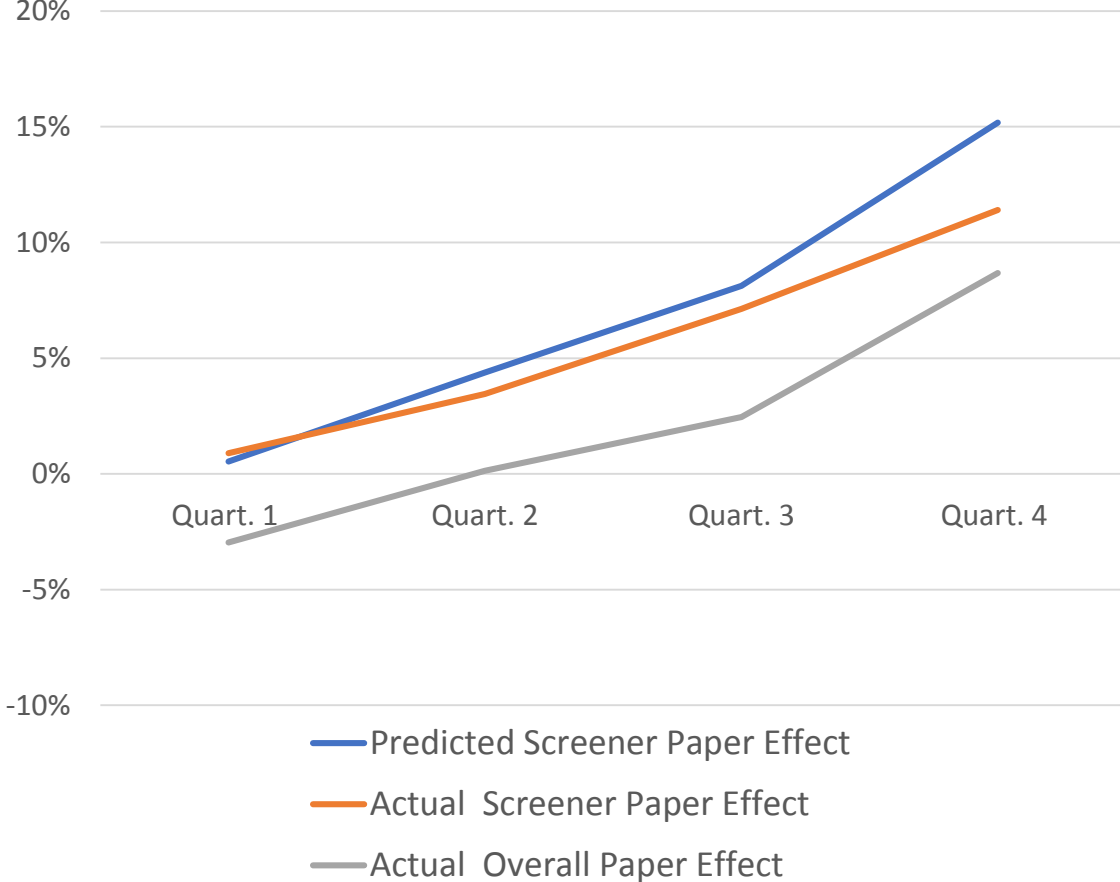
- All results computed on training data and validated on test sample not used in predictive model

Results: Mode Sensitivity (Early Response)

Treatment Effects by Sensitivity Quartile
Frame Data Only

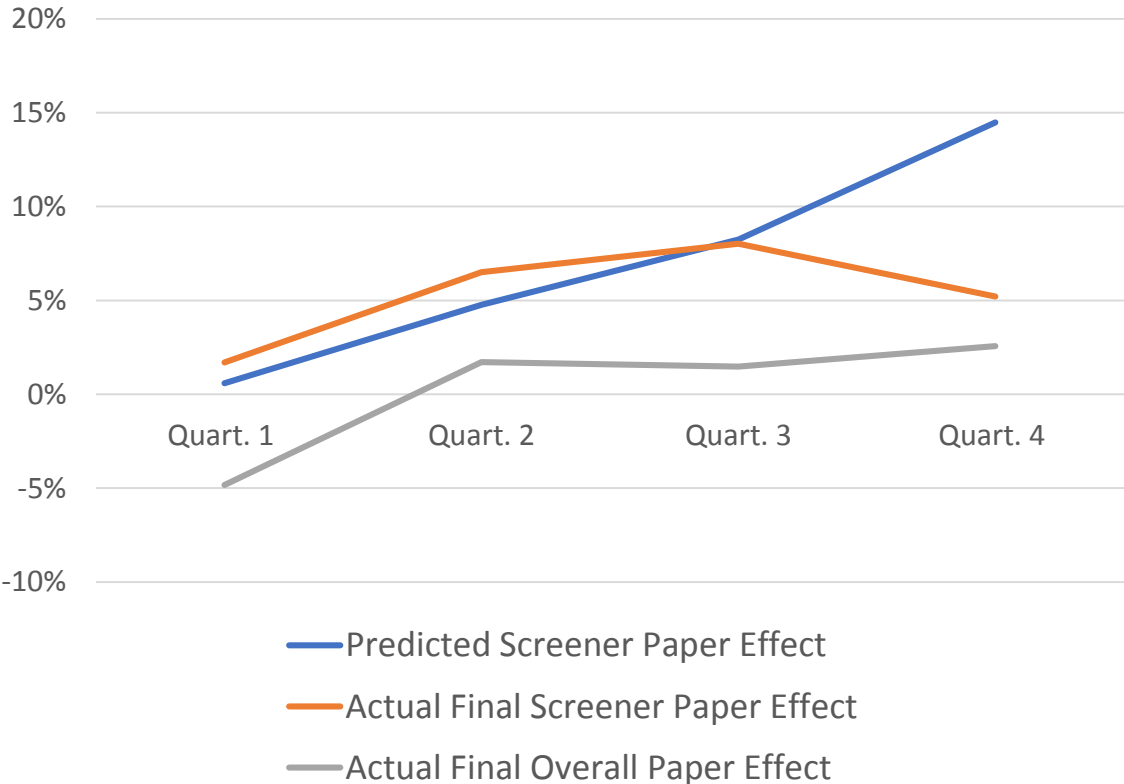


Treatment Effects by Sensitivity Quartile
Frame + Aristotle Data

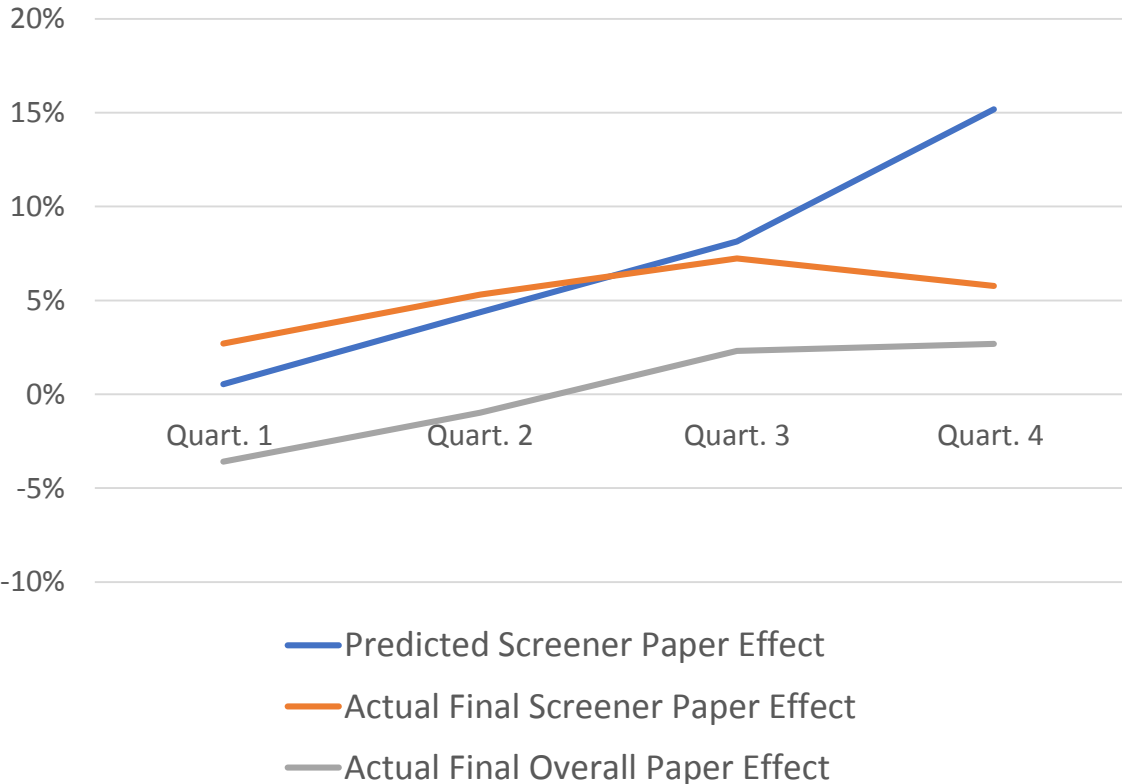


Results: Mode Sensitivity – Final Response Rates

Treatment Effects on Final Response Rates by Sensitivity Quartile
Frame Data Only

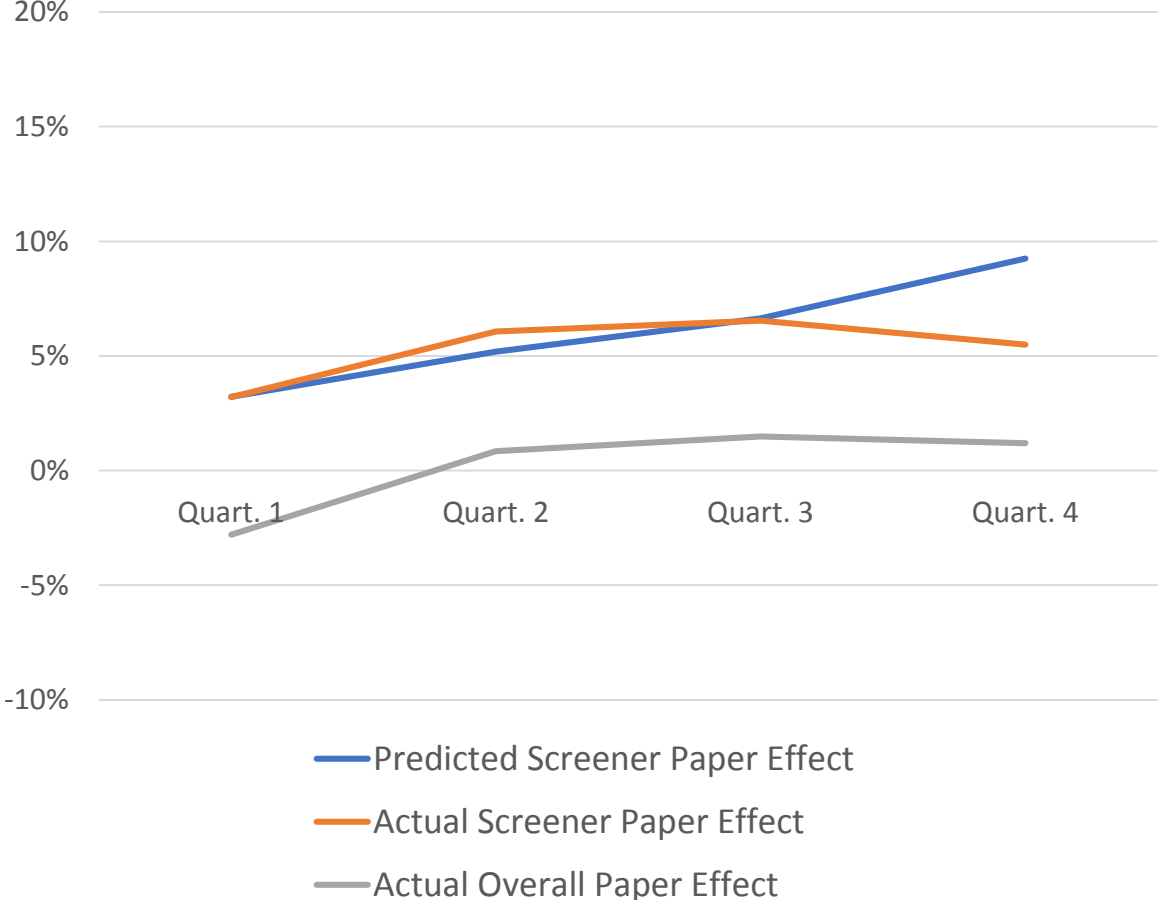


Treatment Effects on Final Response Rates by Sensitivity Quartile
Frame + Aristotle Data

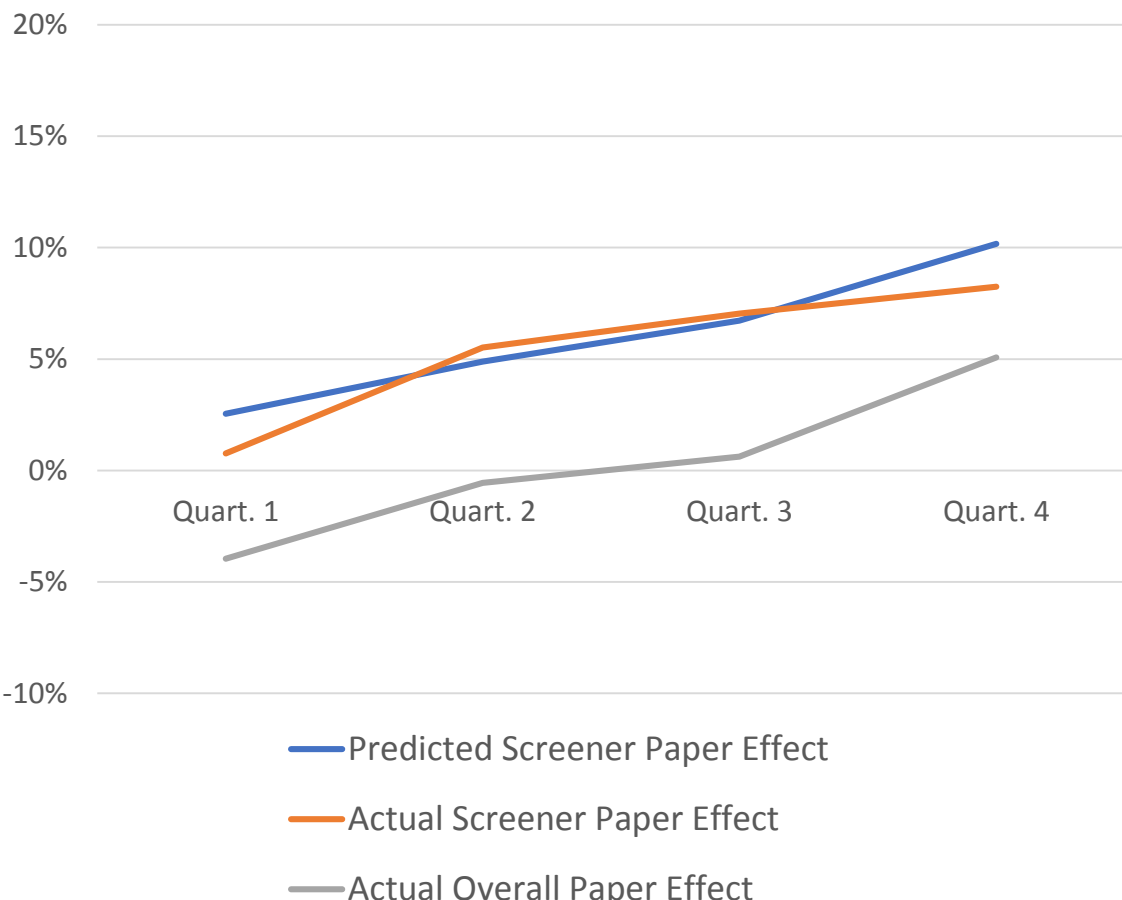


Results: Protocol Sensitivity (Final Response)

Treatment Effects by Sensitivity Quartile
Frame Data Only



Treatment Effects by Sensitivity Quartile
Frame + Aristotle Data



Summary

- Model seems to do a better job predicting mode sensitivity than protocol sensitivity
- Aristotle data does seem to add utility to the model
 - Is the added value worth the added effort?
 - Perhaps there is a better variable selection strategy
- Protocol Model could potentially be used to identify a group that prefers paper and maintains response rate improvements through topical (second phase) collection
- Perhaps response rate differences between mixed-mode and paper only protocols is too small to detect through this type of modeling
- Does this paper preference group even exist?

Next Steps/Considerations

- Improve variable aggregation and explore alternative variable selection approaches
- Explore alternative modeling approaches (e.g., Model-based recursive partitioning [MBRP])
- Rerun validation using five-fold cross validation to ensure robustness
- Apply modeling results to field a mode-based adaptive design experiment in NHES:2019
 - Original idea: Use paper-only for subset in which it achieves the largest gain in final RR (e.g. quartile 4 based on protocol sensitivity score)
 - Results suggest that we may not be able to accurately model protocol sensitivity
 - But, modeling mode sensitivity (effect on response as of wave 2) is more promising
 - Can we adjust our planned design to reflect this?

CAMERON MCPHEE

PRINCIPAL RESEARCHER

202.403.6557

CMCPHEE@AIR.ORG

MAKING
RESEARCH
RELEVANT

THANK YOU

AMERICAN INSTITUTES FOR RESEARCH | AIR.ORG

Copyright © 20XX American Institutes for Research. All rights reserved.



Results: Model Fit

	Mode Sensitivity (Early Response)		Protocol Sensitivity (Final Response)	
	Frame Only	Frame + Aristotle	Frame Only	Frame + Aristotle
McFadden's R-squared	0.070	0.093	0.072	0.096
Accuracy Rate	64%	66%	66%	67%
AUC	0.68	0.69	0.68	0.71

- Aristotle Variables add predictive value in both models
- Model can likely be improved with additional examination and data manipulation

Results: Variable Selection

Demographic/Address characteristic variables

- Head of household age
- Head of household age
- Number of persons 65 or older
- Inferred ethnicity (household mode)
- Ethnicity of head of household²
- Home owned or rented
- Home owned or rented
- Length of residence at address (household mean)²
- Predicted household net worth (household mode)
- Estimated household income (household mode)¹
- Education level of head of household¹
- Sampling stratum (race/ethnicity stratum)
- Percent of block group without high school degree
- Percent of block group speaking non-English language²
- Percent of block group non-Hispanic black living alone²

Other

- Census low response score
- Household responded to something by mail in last 24 months (household mode)²
- Date phone number was validated (household mode)¹

Voting Variables

- Number of times voted (household mean)
- Voted in 2014 general election (household mode)
- Voted in 2016 general election (household mode)
- Voted in 2008 general election (household mode)¹
- Voted in 2012 general election (household mode)²
- Voter propensity for general election (household mean)
- Voted in 2016 presidential primary election (household mode)

Lifestyle Variables

- Aerobic exercise indicator (household mode)
- National "do not call" flag (any household member)
- Interest in animal rights or pets (household mode)¹
- Investor flag (household mode)
- "Broader living" flag (household mode)¹
- Someone in household is a golfer (household mode)²
- Charitable donor (health) flag (household mode)²
- Someone in household interested in décor (household mode)²
- Sports and leisure purchase flag (household mode)²
- Someone in household dieting (household mode)¹
- Interest in taxes or tax reform (household mode)¹
- "High tech living" flag (household mode)¹

RED = Aristotle BLUE = Frame (including PDB)
¹Wave 2 model only
²Overall model only