How Much Is Enough? Moving Toward Smart Stopping Rules for Data Collection

Federal Committee on Statistical Methodology
Research Conference

November 4, 2013

Wan-Ying Chang
and
Lynn Milan

NCSES, National Science Foundation

Disclaimer

This paper reports the general results of research undertaken by staff at the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation (NSF). The views expressed are attributable to the authors and do not necessarily reflect those of the survey sponsors: the NSF and the National Institutes of Health (NIH).

Background

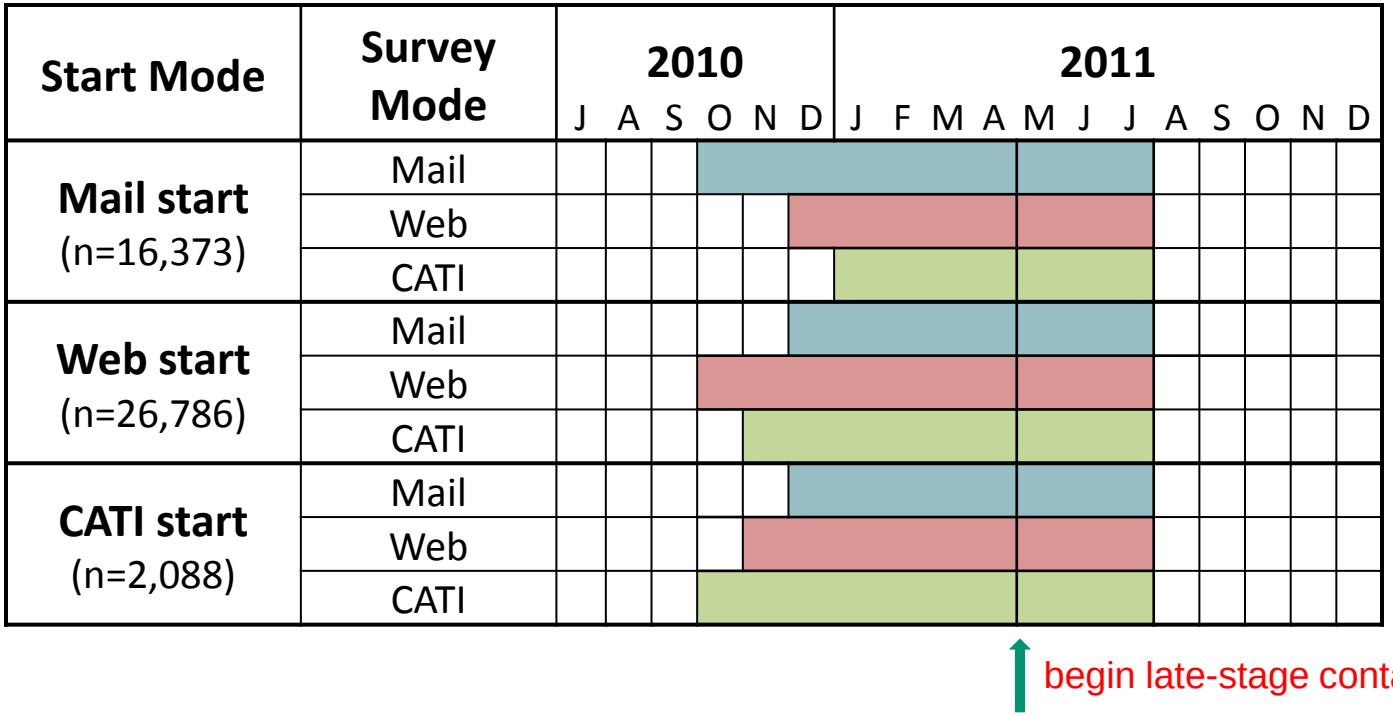
- To obtain the highest possible response rate, surveys often extend their data collection periods
- Increasing timeliness of data and reducing project costs argue for shortened data collection periods
- Survey efficiency may be improved with minimal loss of data quality by applying adaptive interventions and monitoring key indicators throughout data collection
- Key Question: What stopping rules can be developed to guide decisions about when data collection should be concluded?

Survey of Doctorate Recipients (SDR)

- Sponsors: NSF and NIH
- Design: Biennial longitudinal survey (n≈45,000)
- Target population: U.S.-granted doctorate recipients in science, engineering, and health fields who are under age 76
- Question topics: Demographics, education, career history, and employment outcomes
- Sampling frame: Survey of Earned Doctorates

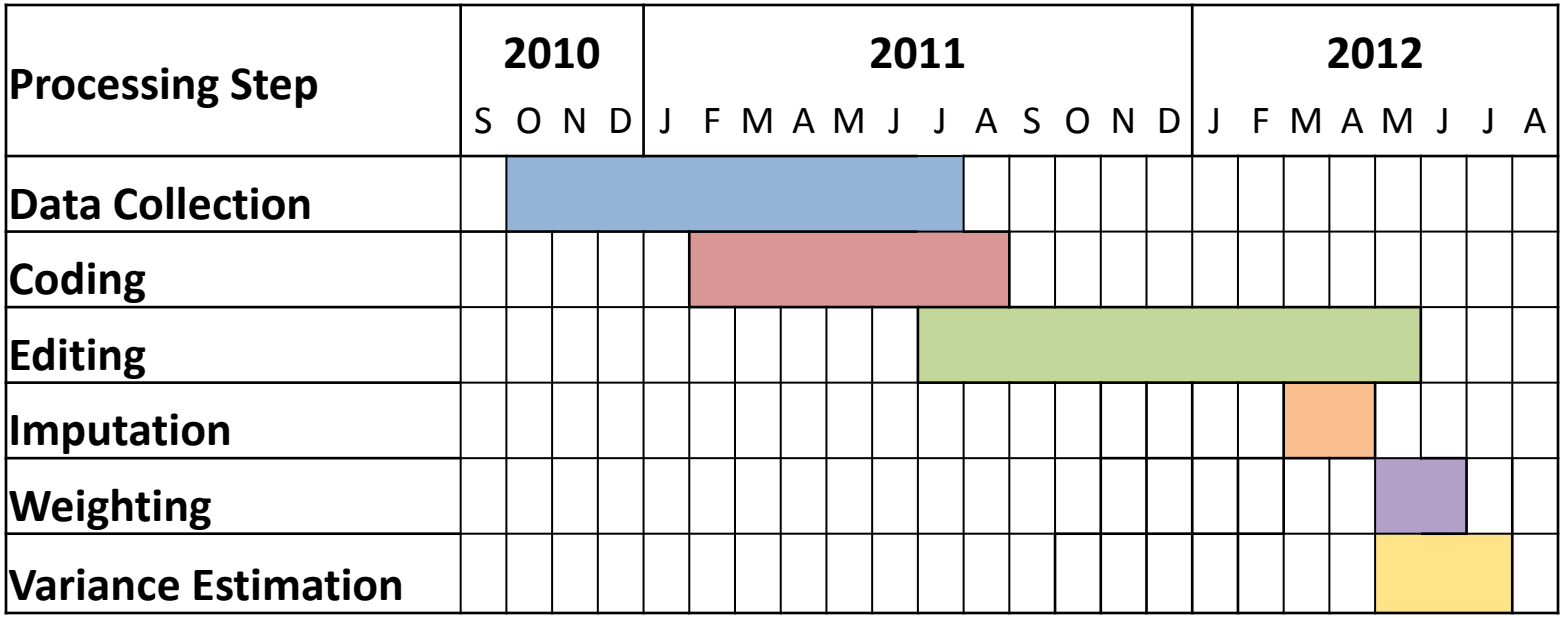
2010 SDR Data Collection Protocol

- Multi-mode: mail, web, telephone
- Start mode based on reported preference or mode used in prior cycle
- Eventually all nonrespondents became eligible for the late-stage protocol, which included a monetary incentive



2010 SDR Data Processing Timeline

- From the start of data collection to the end of data processing took 22 months



Study Overview

- Retrospective study of 2010 SDR responses and paradata
- Steps included
  1. Monitor sample representativeness
  2. Track reporting quality
  3. Study potential bias due to nonresponse
  4. Summarize stability and precision of the survey outcome estimates
  5. Analyze cost and efficiency of collection effort
  6. Search for stopping rules

1. Sample representativeness

- **Full-sample R indicator**

  summarizes the variability in the full sample in terms of the probability of responding,

  $$\hat{R}_\rho = 1 - 2\hat{S}_\rho,$$

  where

  $$\hat{S}_\rho{}^2 = (N-1)^{-1}\Sigma_S w_i\left(\hat{\rho}_i - \bar{\hat{\rho}}_U\right)^2$$

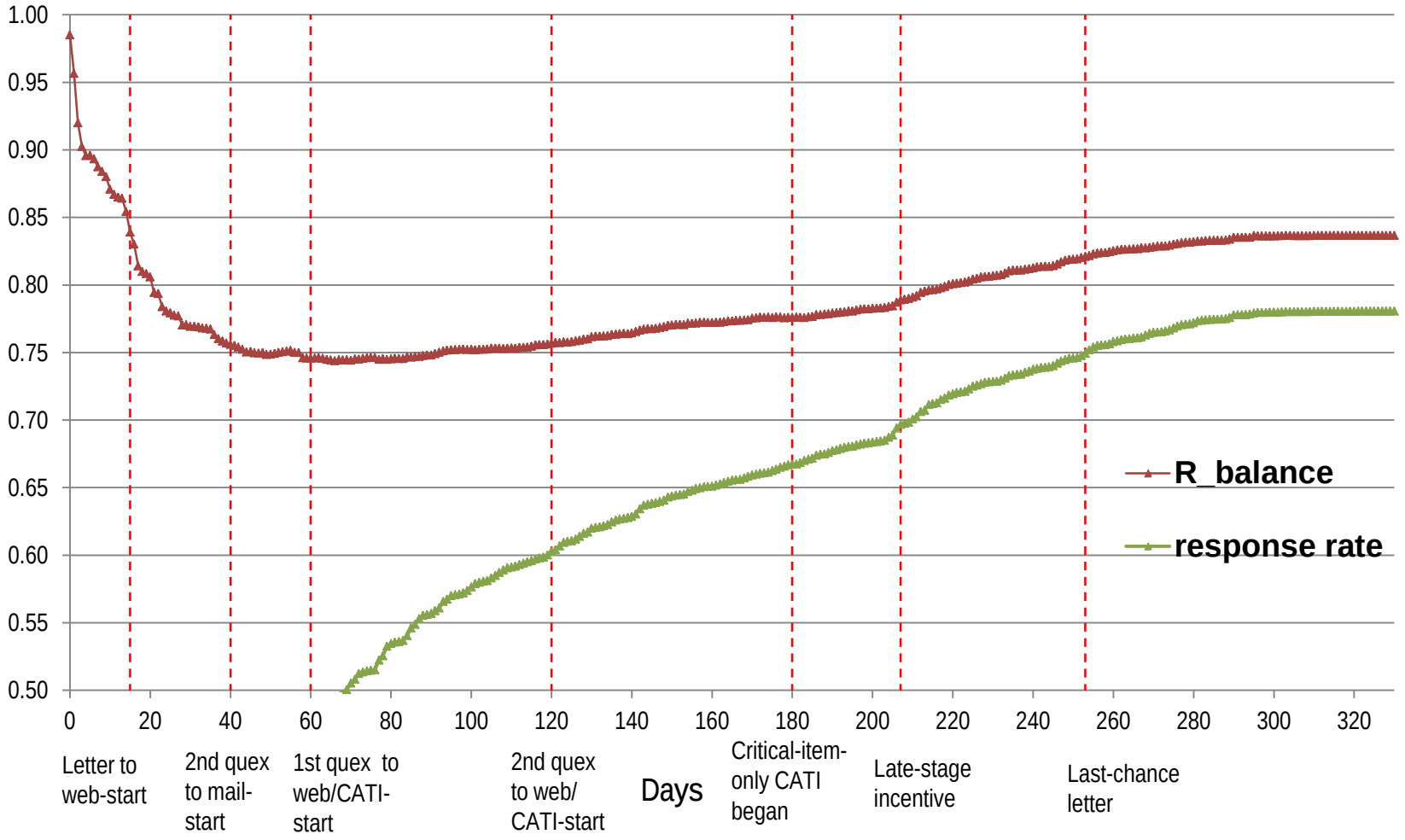
- **Partial R indicator**

  Used to identify over- or under-represented subgroups,

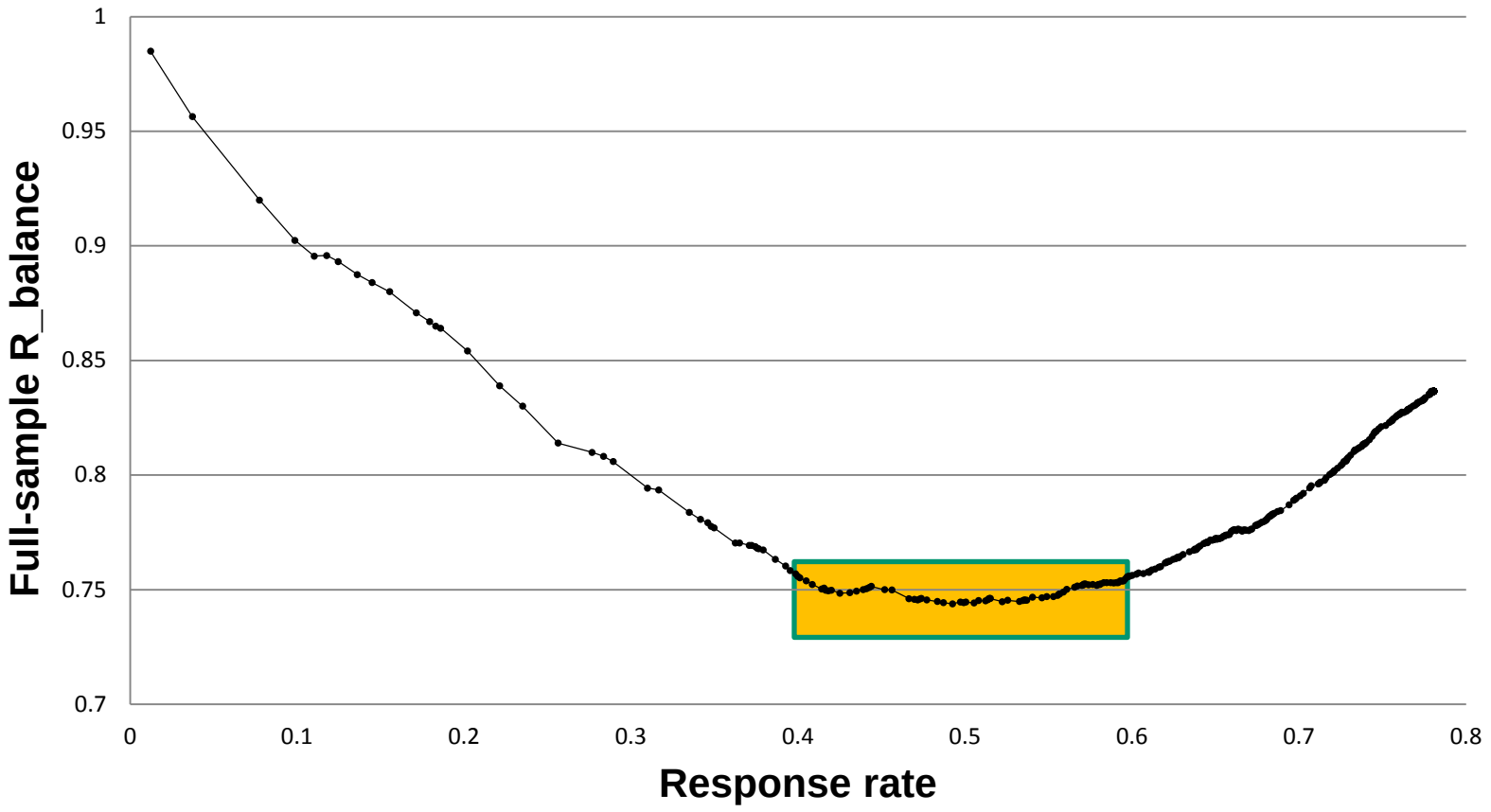
  unconditional partial R of a categorical variable Z with K levels is

  $$S_b(\rho_x|Z) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{K} N_k\,(\bar{\rho}_{X,k} - \bar{\rho}_x)^2}$$

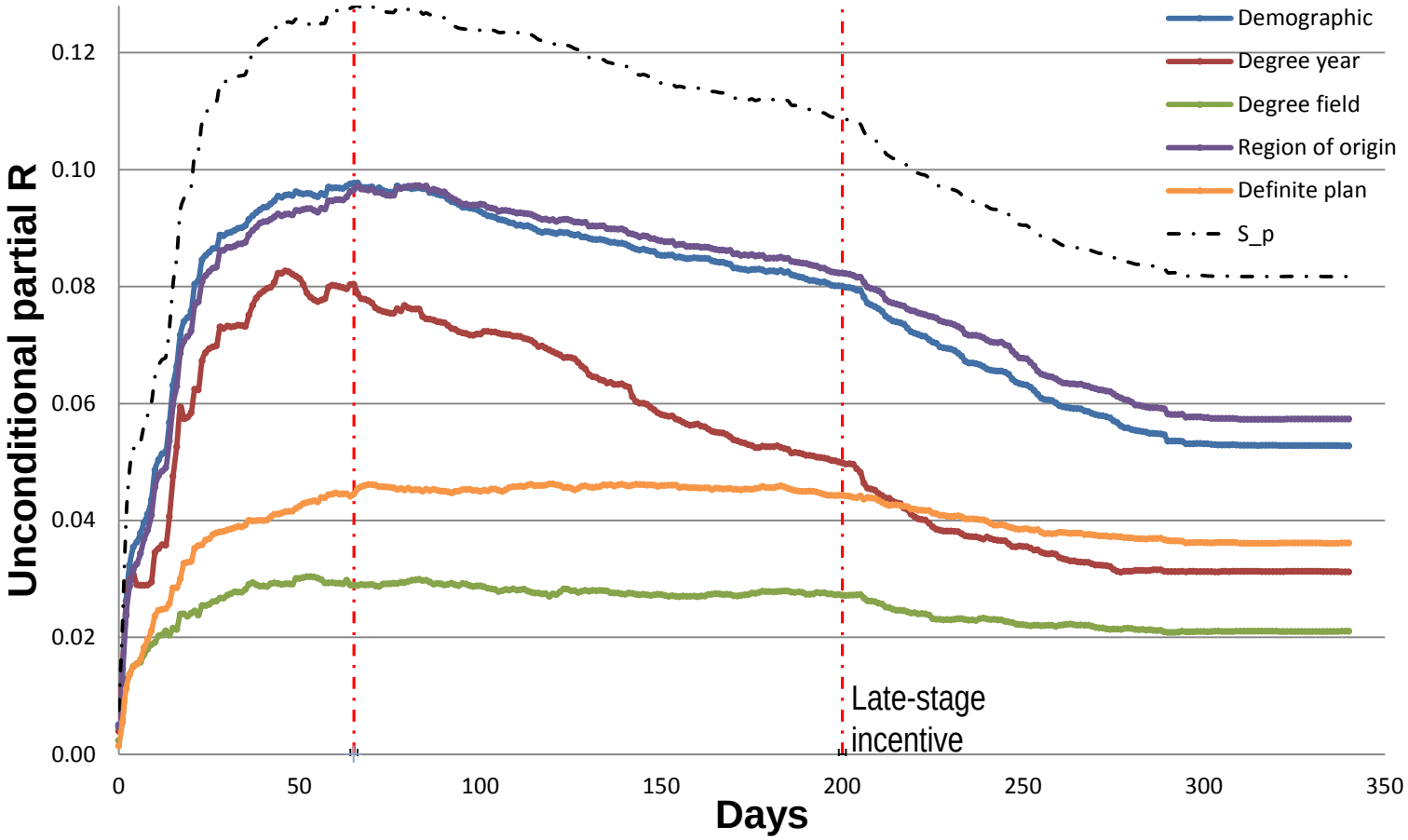
Tracking Full-Sample R Indicator by Major Contact Milestones



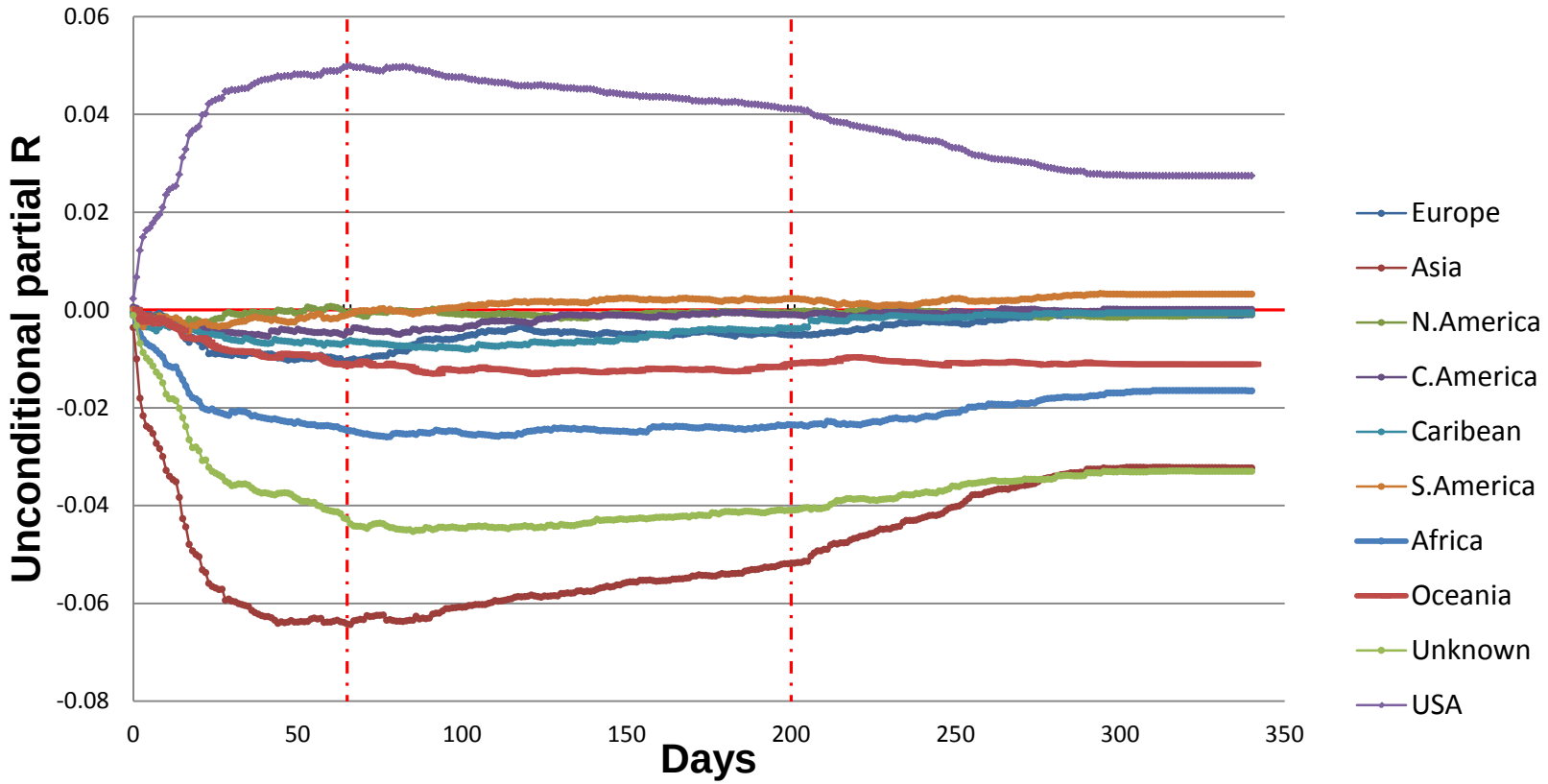
- **Little change in R indicator when response rate climbed from 40% to 60% (day 40 to day 120)**



- **Subgroups by demographic or region of origin have higher level of differential response pattern**

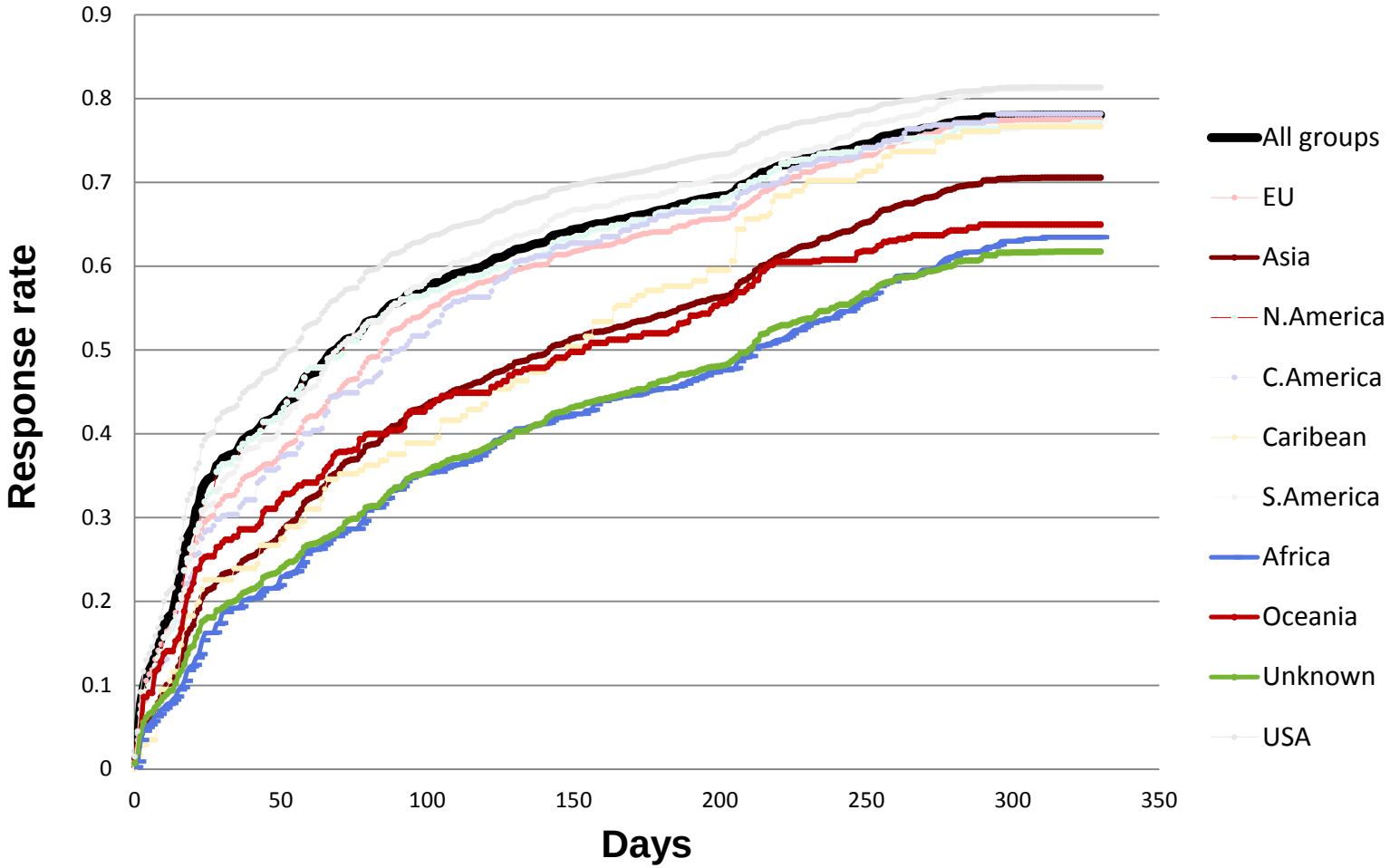


Unconditional Partial R at Category Level for Region of Origin



Unconditional partial R for level *k* of variable *Z* is $\sqrt{\frac{N_k}{N}}\left(\bar{\rho}_{X,k} - \bar{\rho}_X\right)$

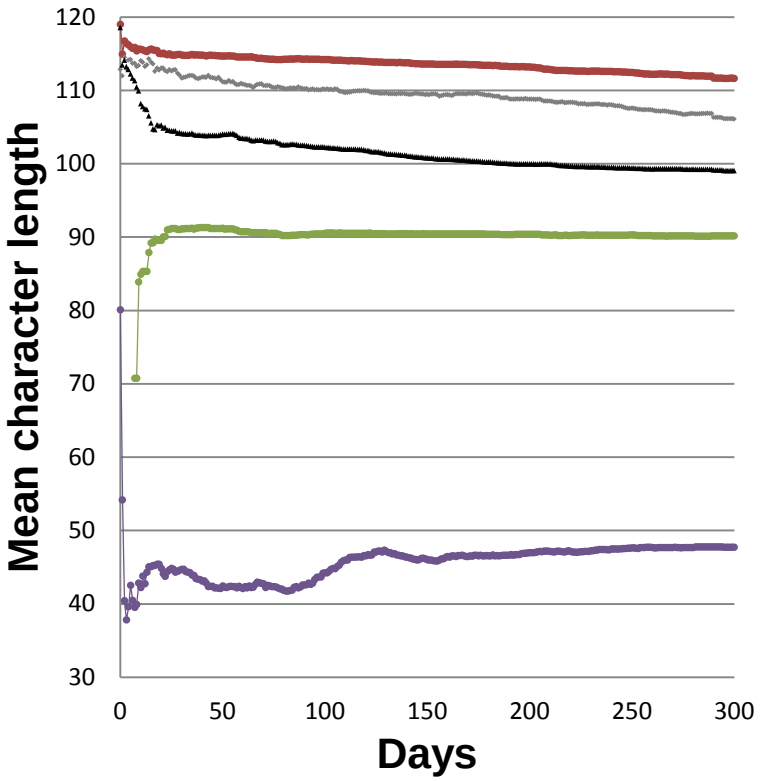
How Does Unconditional Partial R Compare to Response Rate at Category Level?

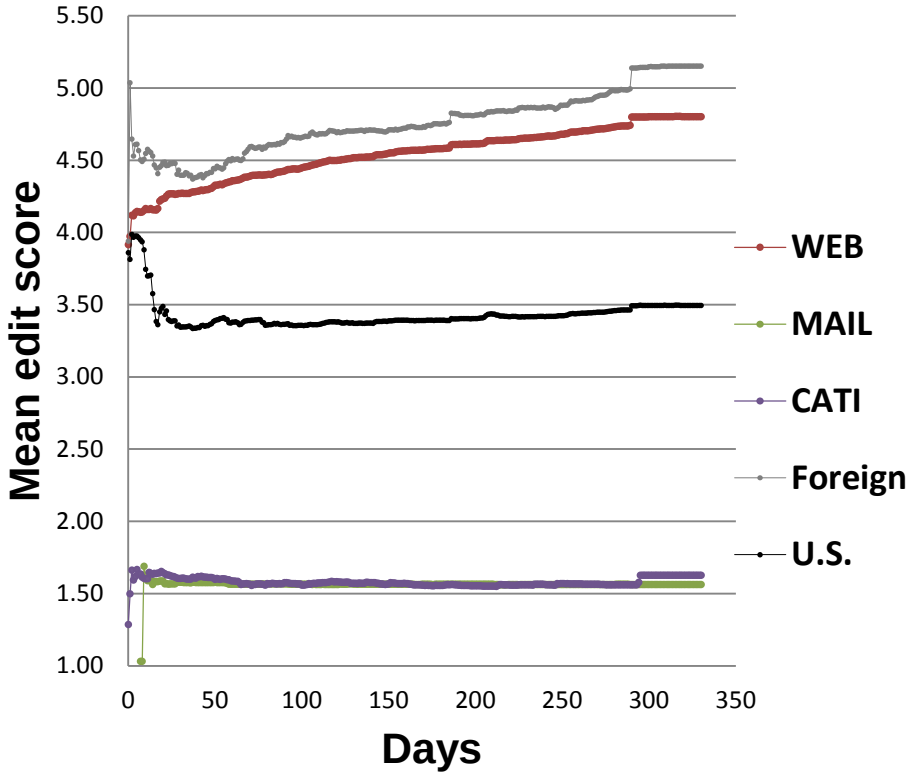


2. Reporting quality

- Measures of reporting quality were tracked
  - Character length of occupation titles and descriptions
  - Percent of edits made to survey variables
  - Percent of survey items imputed
- Longer verbatim length and lower edit and imputation scores imply higher data quality
- Mean scores of the measures were compared by mode and respondents' residency location

- **Verbatim length differences were due to mode effects**
- **Web and foreign responses showed gradual deterioration in data quality**

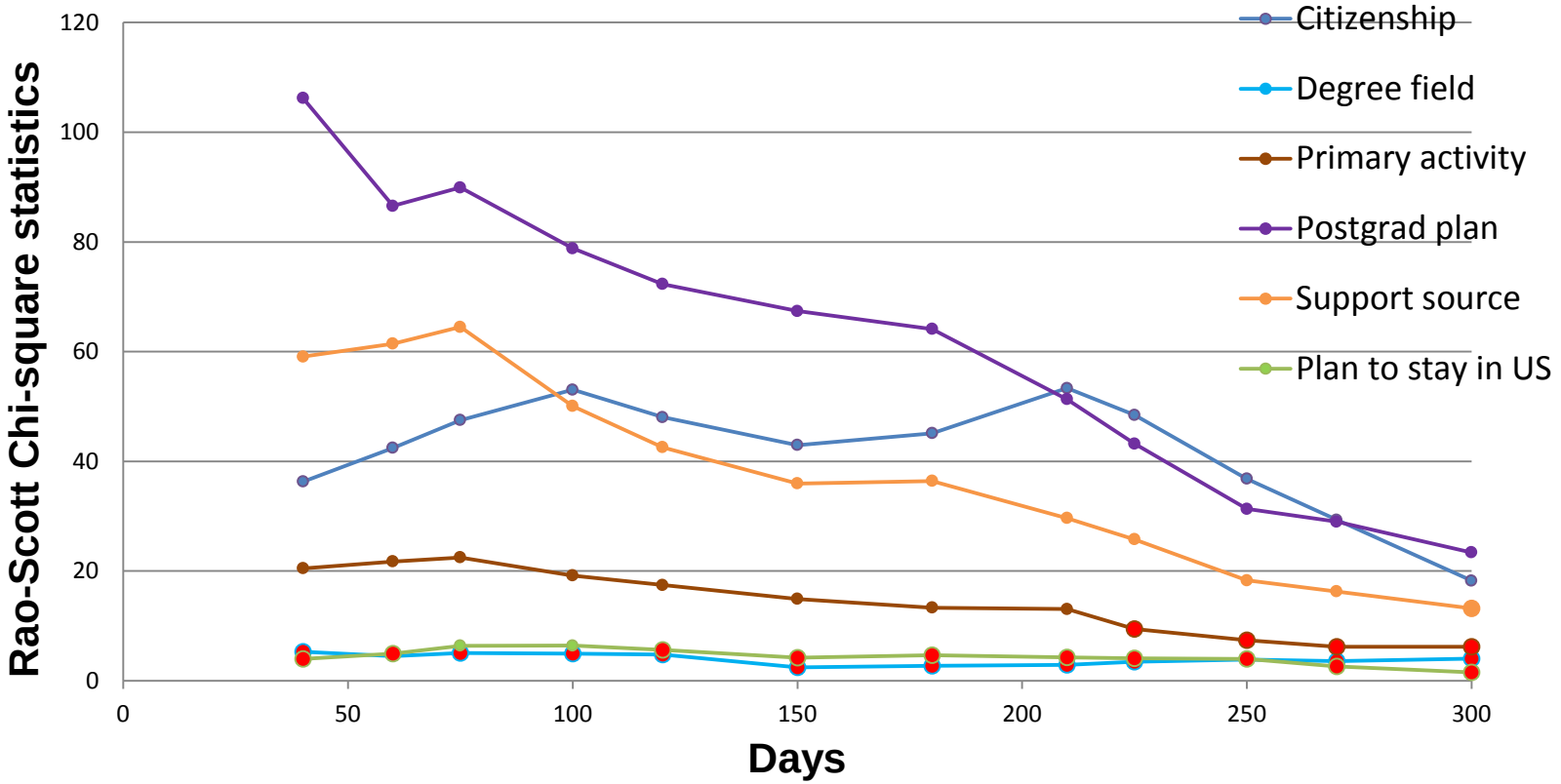




3. Potential bias due to nonresponse

- Base weights of response set were adjusted for nonresponse using propensity models
- Estimates of selected frame variables were monitored over time
- Estimates were compared to full-sample estimates using Chi-square distance and Rao-Scott Chi-Square Statistics

- **Estimates of variables included in the propensity models tracked the full-sample estimates closely from the start**
- **Differences from full-sample estimates decreased over time and continued decreasing during the late-stage data collection**



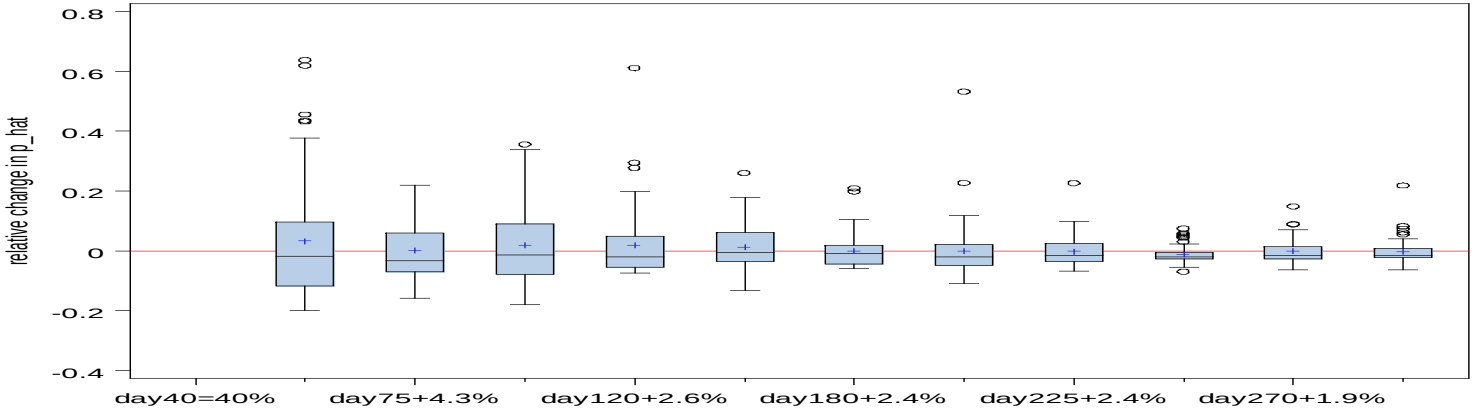
Note: Estimates marked in red are NOT statistically different from the full-sample estimates

4. Stability and precision of survey outcomes

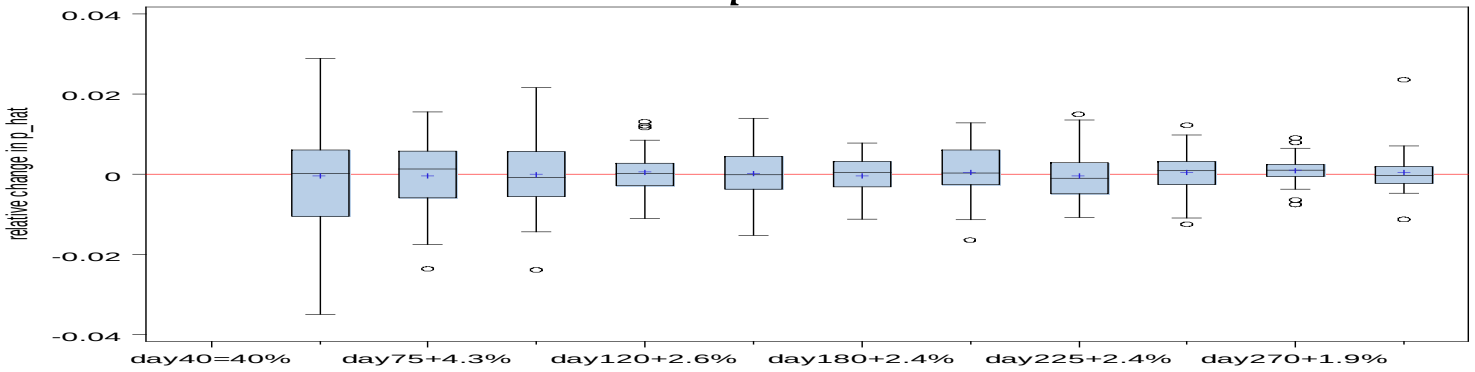
- Estimates of survey outcomes were calculated using the nonresponse adjusted weights, and standard errors were calculated using replicate weights
- 176 domains defined by employment outcomes, race/ethnicity, sex, region of origin, and degree year were selected to represent a wide range of domain sizes (ranging from 0.02% to 43% of the full-time employed population)
- Estimates of proportion for these 176 domains were calculated for 12 time points corresponding to major contacts during data collection
- The 176 domains were classified into 4 groups by domain size for comparing relative change and CV of the estimates

- **Relative change of estimated proportions decreased over time**
- **Small domains experienced significantly more fluctuations**

$\hat{p} \in (0.02\%, 0.25\%)$

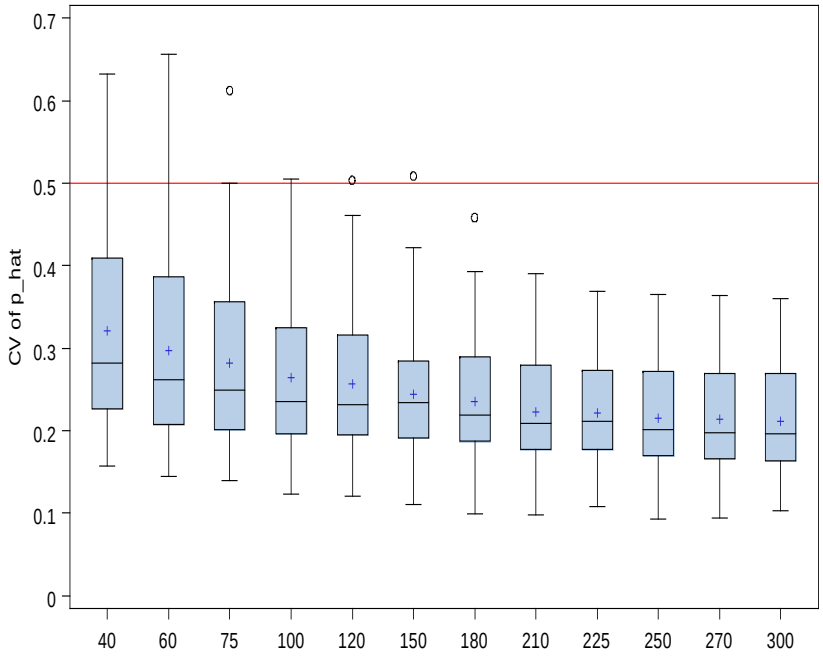


$\hat{p} > 5\%$

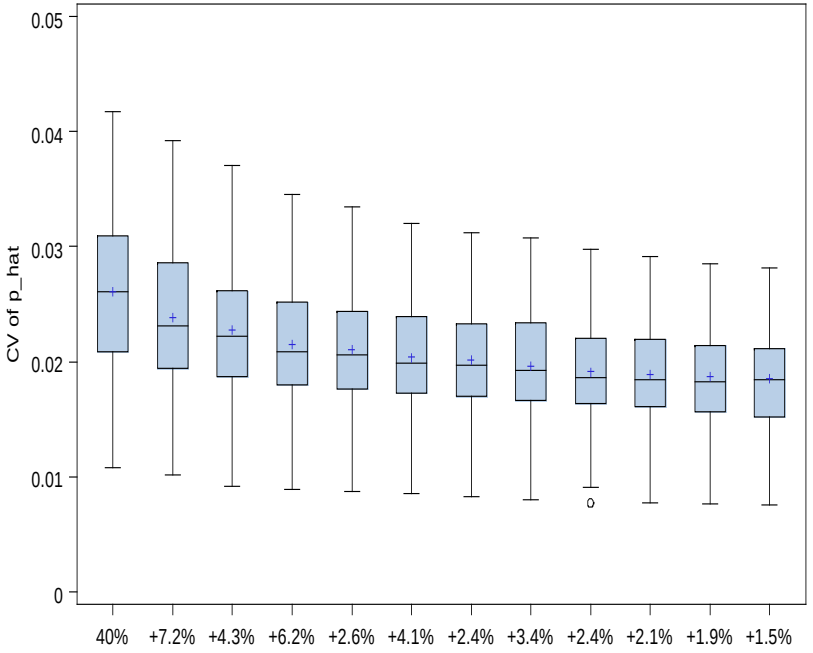


- **CV decreased over time**
- **After day 250 little change was observed**

$\hat{p} \in (0.02\%, 0.25\%)$



$\hat{p} > 5\%$

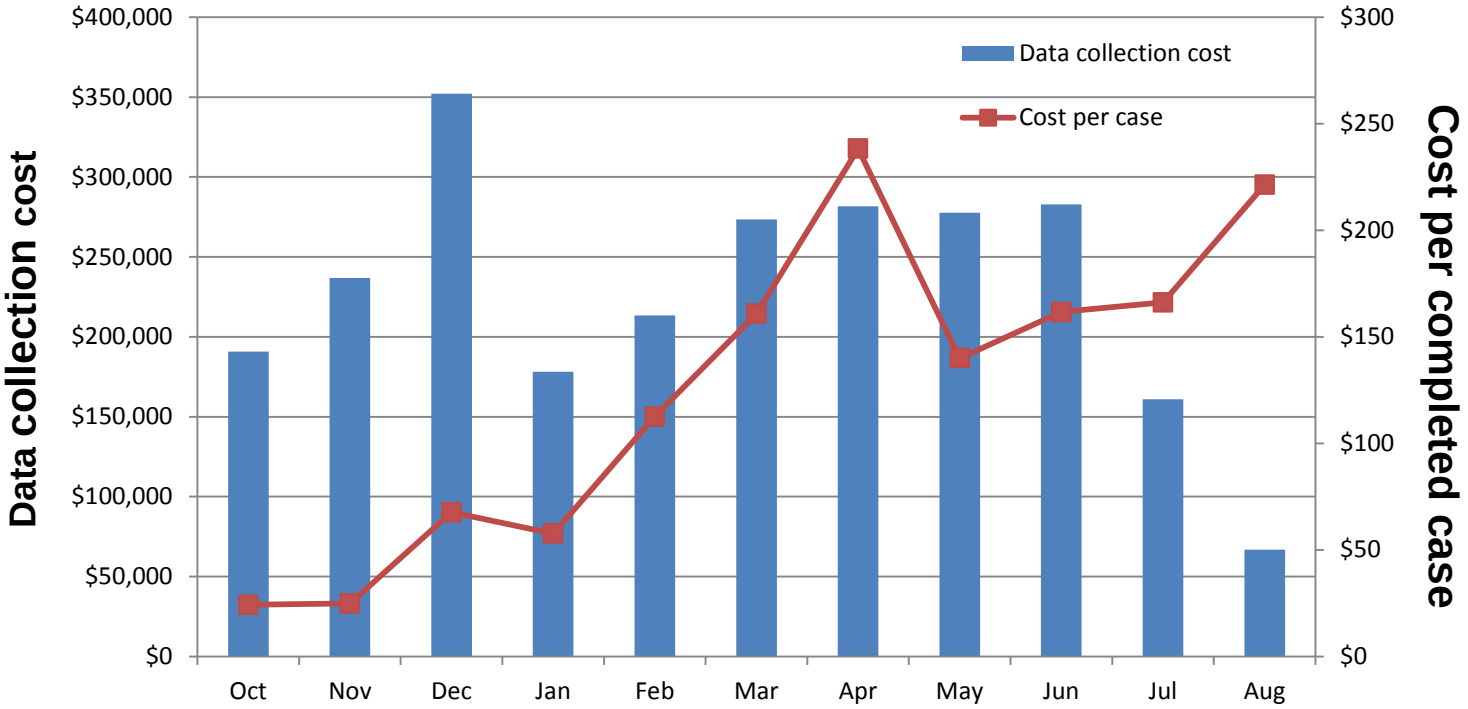


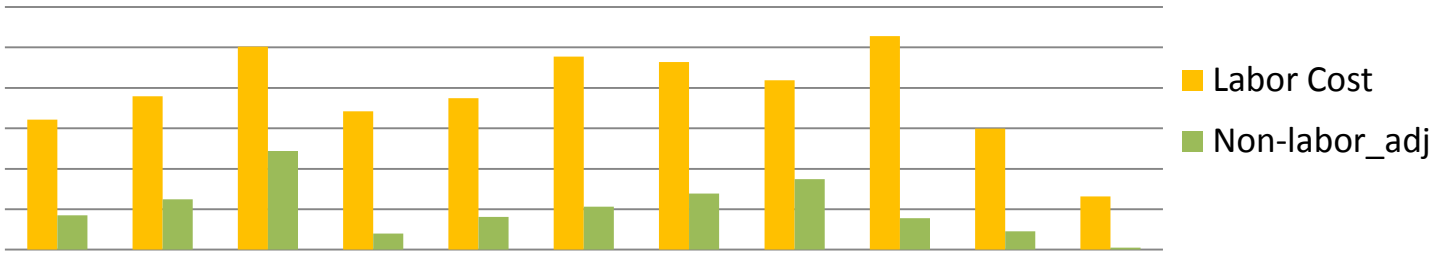
Days

5. Cost and efficiency of collection effort

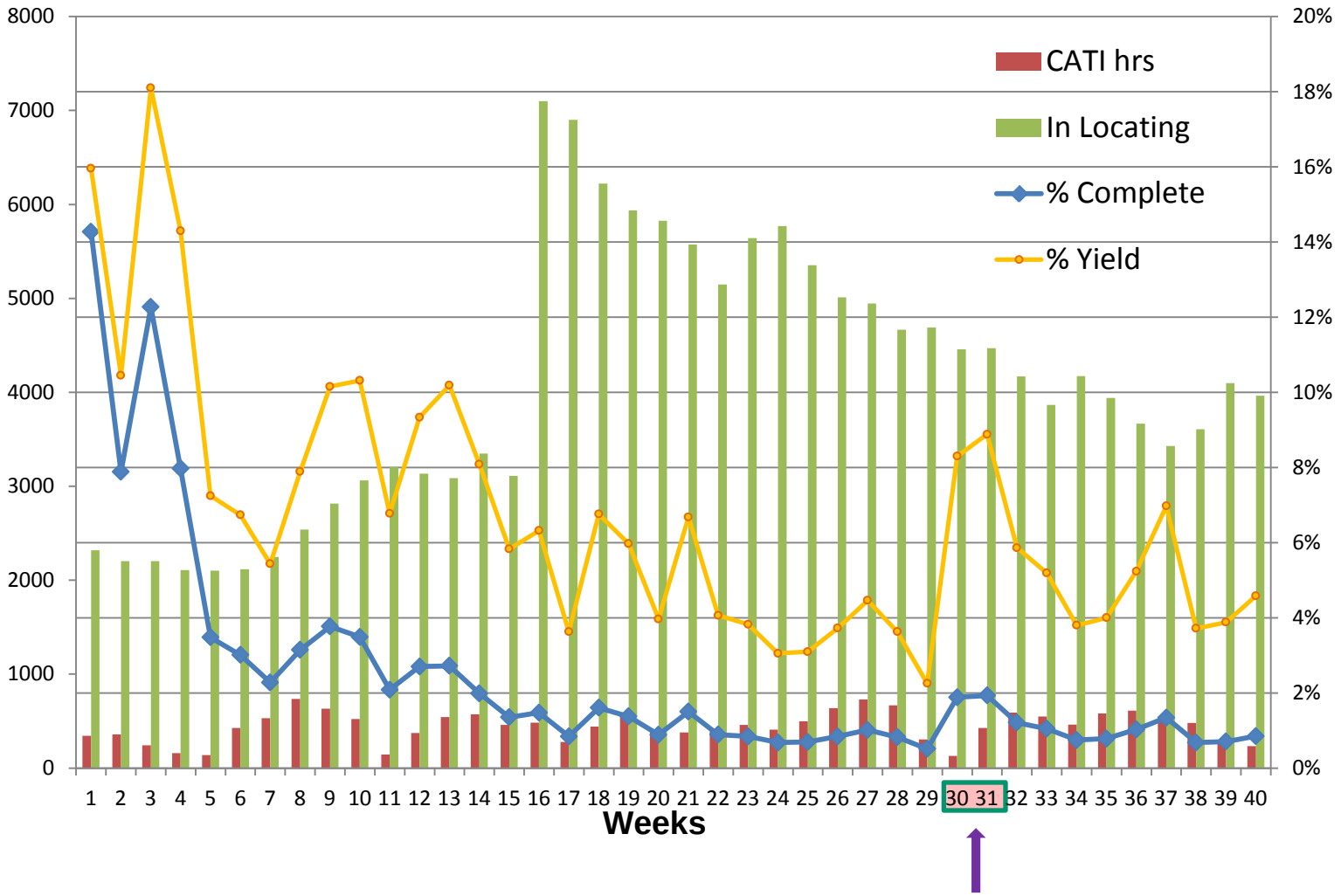
- It was difficult to calculate the data collection cost precisely as a function of time
- The monthly invoice for data collection, mailing type and quantity, CATI hours, and number of cases in locating were used to approximate cost and intensity of effort over time
- The subset of open cases changes over time; at the late stage the open cases consist mostly of hard-to-reach or reluctant cases

- **Late-stage incentive was effective in cost per complete**





CATI and Locating Effort vs. Yield Rate



6. Stopping rule or better orchestrated data collection?

- When data collection interventions and available sample cases both change over time, efficiency is not gained by simply abbreviating the data collection period

- Late-stage incentive was crucial to adding the reluctant group, improving the sample representativeness, reducing nonresponse bias, and improving stability of small domain estimates

- Establishing flow processing will enable real-time monitoring of key data quality measures and evaluating efficiency and timing of various data collection interventions

Future Research

- 2013 SDR data collection followed a compressed schedule while keeping all contact strategies, the efficiency of the new timeline needs to be studied and compared to previous survey cycles
- Continue experimenting and searching for optimal timing and duration of contact strategies
- Improve paradata collection and develop flow processing for SDR

Please direct questions and comments to…

Wan-Ying Chang, Mathematical Statistician
WChang @ nsf.gov

Lynn Milan, Program Officer
LMilan @ nsf.gov

Thank you!