Reducing Respondent Burden in the ACS: Results from the 2015 CAPI Burden Reduction Field Pilot

Todd Hughes, Eric Slud, Robert Ashmead, & Rachael Walsh, U.S. Census Bureau

Background

American Community Survey (ACS)

- Conducted by the U.S. Census Bureau
- Collects demographic, social, economic and housing data
- Uses four modes of data collection over three months for each sample panel of households

Online	Mail	Telephone	Personal Visit
Start of Month 1	Two weeks later	Month 2	Month 3

Research Objectives

- To address respondent concerns about the burden associated with the number and type of contact attempts that are made during the Computer Assisted Personal Interviewing (CAPI) operation of the ACS
- To conduct a field pilot that would assist in preparing for later nationwide implementation

Research Questions

- In treatments that stopped work on cases that exceeded the cumulative burden score threshold:
  - How many cases were pulled, and what was the impact on response rates?
  - Were there differences in perceived contact burden?
  - Were there differences in Field Representative (FR) behavior?
- Were there differences observed between FRs who could see the cumulative burden score, versus those who could not see the score?

Key Limitations

- Challenges with the accuracy and timeliness of paradata
- Modest number of randomized treatment assignments in the pilot
- Other influences on FR behavior, such as differences in supervisory interventions
- Some cases that exceeded the threshold were pulled as late returns for self-response, and the final outcome for the case only reflects the late return
- Compliance was lower than desired with the required twice-daily transmission, which impacted the timely updating of burden scores

Methodology

What is a Cumulative Burden Score?

- A tally of each contact attempt (in any mode) as a separate increment of contact burden
- We assigned a score based on our assessment of the relative perceived burden of various contact attempts
- Threshold set at the 95th percentile based on historical data
- Once the cumulative burden score exceeded the threshold, the case would be pulled from the active workload so no further attempts could be made

Score Comparisons	
Higher Score ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ Lower Score	
Personal Visit ━━━━━━━━━━━━━━Telephone━━━━━━━━━━━━━━━━Mail	
"Strong" Reluctance ━━━━━━ "Soft" Reluctance ━━━━━━━━━━No Reluctance	
Made Contact ━━━━━━━━━━━━━━━━━━━━━━━━━━━Did Not Make Contact	
Left Message/Materials ━━━━━━━━━━━━━━Did Not Leave Message/Materials	

Treatment Groups

	Control	Treatment 1	Treatment 2	Treatment 3
Cases pulled when exceed threshold?	No	No	Yes	Yes
Transmit once or twice daily?	Once	Twice	Twice	Twice
Burden score displayed for each case?	No	No	Yes	No

Comparison Groups | Experimental Groups

Permutation Based Inference

- Under the null hypothesis treatment assignments are simply labels assigned at random to observations
- Calculating the value of test statistic under each possible configuration of treatment assignments yields the exact distribution of the test statistic under the null hypothesis
- Comparing the observed test statistic to the permutation distribution gives a p-value and a test-based CI (see Zieffler, Harring, and Long [2011])
- Permutations of treatment groups were created by reassigning treatment groups to FSA
- 100,000 Monte Carlo samples of those permutations yields a close approximation to the exact distribution of the test statistic, and a test-based confidence interval

Daily Processing

Interviewer **transmits** in the morning before starting work to receive updated scores, and cases that exceed the threshold are automatically removed

Interviewer makes contact attempts on active cases, and records paradata for each attempt

Interviewer **transmits** paradata at the end of the work day

Overnight, a centralized system updates the Cumulative Burden Score and prepares files with stop work instructions and updated scores for interviewers to pick up the next morning

Sample Design

- Formed ten matched SSFA pairs within Regional Offices, one selected to be in the pilot, the other is a control
- Two other SSFAs were included with certainty, given expected high impact of pilot procedures in these areas
- Within selected SSFAs, form FSAs into groups of three, by measure of interviewing difficulty, and assign randomly to Treatments 1, 2, or 3
- All Field Representatives (FRs) within an FSA follow the same treatment

Regional Office

Survey Statistician Field Area (SSFA)			Survey Statistician Field Area (SSFA)		
Field Supervisor Area (FSA)	Field Supervisor Area (FSA)	Field Supervisor Area (FSA)	Field Supervisor Area (FSA)	Field Supervisor Area (FSA)	Field Supervisor Area (FSA)
FR FR	FR FR	FR FR	FR FR	FR FR	FR FR

	Control	Treatment 1	Treatment 2	Treatment 3
FSAs	444	46	46	46
FRs	2,299	236	221	227

Source: American Community Survey Paradata, August 2015.

Results

How many cases were pulled for exceeding the threshold?

	Control	Treatment 1	Treatment 2	Treatment 3
Pulled cases	3	19	187	171
Total cases	44,911	4,299	4,135	4,213
Percent of pulled cases out of total cases	< 0.1%	0.4%	4.5%	4.1%

Source: American Community Survey Paradata, August 2015.

A small number of cases assigned to Control and Treatment 1 were pulled due to the reassignment of cases across treatment groups. Most FRs in Treatments 2 and 3 did not have any of their cases pulled, while a sizeable proportion, 38 percent and 37 percent respectively, did have one or two cases pulled. At most six assigned cases were pulled from any FR.

What was the impact on FR behavior?

- The quality of the paradata used to calculate burden scores is affected by FR compliance with procedures that require them to record information about each contact attempt
- Given that FRs may be motivated to be less compliant with recording CHI entries if cases are removed from their workload when they exceed the cumulative burden score threshold, it was necessary to assess FR paradata-reporting behavior during the pilot
- Indicators of FR paradata-reporting behavior were found to change little across control and treatment groups during the pilot study, including the proportions of entries for:
  - not attempting contact
  - observing the household from the FR's vehicle
  - personal visit versus telephone attempts
  - attempts made before noon, early afternoon, late afternoon, and post-6 p.m.
  - weekday versus weekend attempts
  - attempts in which low-, medium-, or high-burden "strategies" were reported

Another way to assess FR reporting accuracy is to compare the days with reported interviewing payroll hours with whether there were corresponding paradata entries that day documenting their contact attempts and results.

Type of Day	Control	Treatment 1	Treatment 2	Treatment 3
Payroll and paradata sent	83.5%	84.1%	81.6%	84.4%
Payroll sent only	16.5%	15.9%	18.4%	15.6%
Total FR-days	25,493	2,656	2,451	2,666

Source: American Community Survey Paradata, August 2015.

The percent of payroll-only FR-workdays with interviewing hours was roughly the same among the controls and Treatments 1 and 3. The apparently higher percentage of payroll-only days in Treatment 2 was not significantly larger than the percentage for Treatment 3 [p-value = 0.187, 90 percent CI (-0.7%, 6.5%)].

What was the impact on response rates?

	Control	Treatment 1	Treatment 2	Treatment 3	Treatments 2-3
August (Test Month)	93.4%	93.1%	92.1%	91.5%	91.8%
July	93.1%	93.5%	93.8%	93.3%	93.6%
Estimated Change* (August –July)	†	-0.5%	-1.7%	-1.7%	-1.7%

*This is not necessarily equal to the August – July above. Differences are calculated at the FSA-level and averaged by the total cases over the two months. †Not calculated because not all FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

We estimate that the response rate was a borderline-significant 1.3 percentage points lower [90 percent CI (0.0%, 2.6%), p-value =0.104] for Treatments 2 and 3 combined versus Treatment 1. Comparing this estimate (1.3 percent) to the percentage of cases pulled (4.3 percent), one can argue that approximately two-thirds of pulled cases would not have resulted in complete interviews if the burden stopping rules were not in place.

What was the impact on perceived contact burden?

Source: American Community Survey Paradata, August 2015.

- The plot on the right hand side is the right tail of the plot on the left
- There are similar distributions of cases with burden scores less than 40 for all treatment groups, but Treatments 2 and 3 had only 0.3 percent of cases with burden scores over 60, while control and Treatment 1 had over 2.0 percent of cases with burden scores over 60
- Treatments 2 and 3 had less than 0.1 percent of cases each with burden scores over 80, while Control had over 0.5 percent of cases and Treatment 1 had 0.4 percent of cases with burden scores over 80

Conclusions

- Implementation of the cumulative burden score stopping rule was effective at reducing some metrics of the perceived contact burden of ACS CAPI operations, while also having a small negative impact on response rates
- The Census Bureau will continue to prepare for a nation-wide implementation of the cumulative burden score and an associated stopping rule in June 2016
- In consideration of the results of this research and the feedback received during debriefing sessions conducted with many of the field staff involved in the pilot, we do not see significant benefits for showing the cumulative burden score to the FR versus not showing the score
- The full report is available online: www.census.gov/library/working-papers/2016/acs/2016_Hughes_01.html