UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

March 15, 2016

2016 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS16-RER-07

MEMORANDUM FOR ACS Research and Evaluation Advisory Group

From: Deborah Stempowski
Chief, American Community Survey Office

Prepared by: Todd Hughes
American Community Survey Office

Subject: Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation

Attached is the final American Community Survey Research and Evaluation report, "Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation". We conducted this pilot 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 American Community Survey (ACS). The pilot for the cumulative burden score stopping rule was conducted in August 2015 in roughly one quarter of the field geographies in which ACS interviewing was conducted. The results from the pilot will assist in preparing for deployment of the cumulative burden score stopping rule for the nationwide CAPI operations in the late Spring of 2016.

If you have any questions about this report, please contact Todd Hughes at 301-763-6686.

Attachment

cc:
ACS Research and Evaluation Work Group
ACS CAPI Burden Reduction Research Group

Joe Schaefer
Thomas Mathew

American Community Survey Research and Evaluation Program

March 15, 2016

Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation

FINAL REPORT

Todd Hughes, Eric Slud, Robert Ashmead, Rachael Walsh

American Community Survey Office, Center for Statistical Research and Methodology, Office of Survey and Census Analytics

(This page left intentionally blank)

TABLE OF CONTENTS

TABLES

FIGURES

Executive Summary

The multi-mode data collection strategy of the American Community Survey (ACS) can be perceived as overly intrusive by some respondents, and the Census Bureau is conducting research to address respondent concerns about the burden associated with the type and number of contact attempts that are made. In 2014 and 2015, the Census Bureau performed research to prepare for reducing respondent burden in the ACS Computer-Assisted Personal Interviewing (CAPI) operation. The Census Bureau documented the level of effort, respondent burden, costs and quality of the CAPI methods based on an analysis of 2011 and 2012 ACS CAPI paradata (Zelenak 2014 and Griffin and Nelson 2014). Additionally, an interdivisional group developed potential rules for determining when to stop contacting a sample address (i.e. stopping rules) and, therefore, reduce the number of contact attempts made in the CAPI operation (Griffin 2014). The team then modeled the effects of the proposed stopping rules on respondent burden, cost, and quality (Griffin, Slud, and Erdman 2015).

Based on this work, in August 2015, the Census Bureau conducted a field pilot in the CAPI operation of the ACS to evaluate a stopping rule based on a "cumulative burden score." This rule assigned a value to each contact attempt (in any mode) as a separate increment of burden, in order to estimate the perceived contact burden for the respondent. Once the cumulative burden score exceeded a threshold of 40, the case was pulled from the active CAPI workload, so no further contact attempts could be made. We assigned burden score increment values and the burden score threshold based on our assessment of the relative burden of the various contact attempts.

Historically Field Representatives (FRs) relied on their judgment and feedback from their supervisors to determine when to stop attempting to contact a respondent. A CAPI stopping rule based on the cumulative burden score would reflect a significant cultural and procedural shift for the field interviewing staff. The Census Bureau determined that a pilot involving a subset of the ACS CAPI sample would help identify training and tools needed for FRs and their managers to prepare successfully for full implementation. A pilot would also allow the Census Bureau to confirm that the IT systems' changes necessary to implement the burden score and stopping rule were functioning as intended. It would also help to determine whether the impacts on interviewing hours per case, response rates, and perceived contact burden were in line with the estimates produced in the 2015 research that modeled the impacts using historical data (Griffin, Slud, and Erdman 2015).

To calculate and update burden scores correctly, the August pilot required that FRs transmit ACS data from their laptops twice daily on days worked – once before attempting contacts and once after. The transmission prior to contact attempts updated the cases' burden scores with information from the day before and removed any cases with burden scores above the burden threshold from FRs' laptops. The transmission after contact attempts were used by the Census

Bureau's Unified Tracking System (UTS) to calculate and update the burden score for each case. Historically, FRs have been required to transmit once per day, at the end of their workday, therefore, transmitting twice per day reflected a change in procedures for FRs. Since these additional transmissions would enable cases to be pulled from the FR's workload at the start of their workday, it was unclear whether the FR's concern with losing cases might motivate the FR not to comply with the new transmission policy. Therefore, the Census Bureau developed experimental treatments that allowed for an evaluation during the pilot of whether FRs were more or less likely to comply with the new transmission policy depending on whether they knew it could lead to cases being removed from their workload. Additionally, the research team was interested in determining whether showing the current cumulative burden score to field staff would influence their behavior, so the pilot was designed with different treatments that would assist in making a recommendation on whether or not to show the score to field staff. The Census Bureau used the results from the pilot to prepare for deployment of the cumulative burden score stopping rule nationwide for the CAPI operations in the late Spring of 2016.

The pilot for the cumulative burden score stopping rule was conducted in roughly one-quarter of the field geographies in which ACS interviewing was conducted. The remaining three-quarters of field geographies followed the standard field protocol, forming a control group. Within the Survey Statistician Field Areas (SSFAs) selected for inclusion in the pilot, individual Field Supervisor (FS) areas and all the Field Representatives (FRs) included in that FS area (FSA) were assigned randomly to one of the three experimental treatments:

Treatment 1 (two transmissions required, burden score not displayed, cases not removed): FRs and FSs in this treatment were asked to follow the modified laptop data transmission procedures to transmit twice per day, but cases were not removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.

Treatment 2 (burden score displayed, cases removed): FRs and FSs in this treatment were asked to follow the modified transmission procedures to transmit twice per day, cases were removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was displayed on FR laptops and in the Regional Office control systems.

Treatment 3 (burden score not displayed, cases removed): FRs and FSs in this treatment were asked to follow the modified transmission procedures to transmit twice per day, cases were removed for exceeding the cumulative burden threshold, but the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.

All but two of the SSFAs selected for inclusion in the pilot were matched to control SSFAs, and these matched non-included SSFAs can together be viewed as a concurrent matched control group. Despite the original intentions with this matched design, all statistical comparisons in this report concern only contrasts among Treatments 1 to 3 within included SSFAs. In addition, many descriptive comparisons are made between Treatments 2–3 and the full set of concurrent non-matched controls from all non-included SSFAs, defined as:

> **Control (single transmission required, burden score not displayed, cases not removed):** FRs and FSs in this group were required to transmit only once per day on days worked. Cases were not removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.

In the results, we make statistical comparisons only among Treatments 1, 2 and 3. This is because in the design, pilot FSAs were randomized between those treatments. We do not statistically test comparisons between the control group and the other treatments because the assignment of SSFAs to the control group was not done entirely at random. Other limitations outlined in section four of this report (including challenges with the accuracy and timeliness of paradata, the modest number of randomized FSA treatment assignments in the pilot, and other influences on FR behavior) make this report necessarily exploratory rather than a source of definitive answers to all questions of interest. During the August 2015 pilot, the following key results were observed:

Workload:

The percent of cases pulled (out of total cases in the workload) from Treatments 2 and 3 were 4.5 percent (out of 4,135 total cases) and 4.1 percent (out of 4,213 total cases) respectively. The specific percent of cases in the pilot pulled from Treatments 2 and 3 in each SSFA ranged from 2.1 percent to 10.6 percent. The majority of FRs in treatment groups 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. In the pilot, at most six assigned cases were pulled from any single FR. Cases located in areas associated with response rate Performance Clusters 2 and 3 (i.e. areas with historically lower response rates) were pulled at higher rates than those associated with response rate Performance Cluster 1 (i.e. area with historically higher response rates).

Perceived Contact Burden:

We observed a 6.0 percent decrease in average reported contact attempts per case and a 6.2 percent decrease in the reported contacts per case in Treatments 2 and 3 combined versus Treatment 1. We also observed a 19.4 percent reduction in reported sample-person contacts with a firm refusal for Treatments 2 and 3 over Treatment 1. However, we note that the number of such contacts is overall quite small.

Implementing a stopping rule based on the cumulative burden score reduced the number of cases with high burden scores. There were similar distributions of cases with burden scores less than the threshold of 40 for all treatment groups, but only 0.3 percent of cases in Treatments 2 and 3 had 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 in the Control group more than 0.5 percent of cases had burden scores over 80. (Note: A case in Treatments 2 and 3 can have a score that exceeds the threshold due to additional attempts made on the case prior to transmitting.)

Measures of perceived contact burden were based on paradata reported by FRs, therefore we cannot be certain in all instances whether these measures reflect actual changes in attempts made or, instead, reflect changes in reporting by some FRs. However, many indicators of FR reporting behavior did not demonstrate evidence of changes in their reporting behavior (see "Field Operations" below).

Interviewing Hours and Miles:

While FRs made fewer reported contact attempts per case in Treatments 2 and 3 during the August pilot, we did not find evidence that interviewing hours or miles per case decreased. This finding may be explained in part by the hours per case ceiling in place for ACS data collection operations. There are a number of possible explanations for this, such as if FRs still have hours remaining within the allocation provided for their workload, then they may make more attempts on remaining cases after other cases are removed from their workload for exceeding the cumulative burden score threshold. These hours per case allocations were not modified during the pilot test to account for cases pulled from FR workloads. Previously released research modeling the effects of the proposed stopping rules on respondent burden, cost, and quality using 2012 paradata estimated a national reduction of 4.4 percent of interviewing hours based on the implementation of a cumulative burden score stopping rule (Griffin, Slud, and Erdman 2015). The findings by Griffin et al. may inform future decisions about appropriate interviewing hours per case allocations during production implementation of these procedures.

Response Rates:

Lower response rates were observed in Treatments 2 and 3 versus Treatment 1, due to cases being removed for exceeding the cumulative burden score threshold. We estimate that the response rate difference was borderline-significant at 1.3 percentage points lower (two-tailed p-value =0.104) for Treatments 2 and 3 versus Treatment 1. (Here and throughout, all p-values are for two-tailed hypothesis tests.) Comparing this estimate (1.3 percent) to the percentage of cases pulled (4.3 percent), one can argue that most pulled cases would not have resulted in completed interviews even if the burden stopping rules were not in place. The largest decrease in response

rate for Treatments 2 and 3 compared with Treatments 1 are for units in response rate Performance Clusters 2 and 3 with higher initial burden scores.

Field Operations:

FR compliance with the twice-a-day transmission guidelines was uneven across SSFAs. In the three treatment groups, compliance was on average 20.9 percent for the start of the work-day transmissions and 83.7 percent for the end of the work-day transmissions. FRs in Treatment 2 were more compliant (20.2 percent overall) than in Treatments 1 and 3 where the respective overall compliance rates were 16.4 and 15.4 percent. Compliance rates for Treatments 2 and 3 were significantly different (p-value = 0.035). Transmission compliance was lower than needed to ensure burden scores were updated accurately each day and cases were pulled in a timely manner, and therefore some objectives of this pilot were not fully realized. (As an indication of this, out of 748 cases in Treatments 2 or 3 attaining cumulative burden of 41 or more, 84 had an additional burden increment on a day subsequent to reaching 41, and of these only 37 were eventually pulled.)

Errors were identified, and later corrected, during the pilot in the reports that managers had available to monitor transmission compliance. These errors reduced managers' ability to intervene when FRs were not following procedures, which potentially contributed to low compliance rates. Correctness of these reports must be confirmed prior to implementation in production of case removal based on the burden score, and managers must give transmission compliance significant attention, to ensure that burden scores are accurately updated each day and cases pulled in a timely manner.

The cumulative burden score calculation relies heavily on the paradata FRs record in the Contact History Instrument (CHI); the quality of these CHI entries are 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 CHI reporting behavior during the pilot. Indicators of FR CHI-reporting behavior were found to change little across control and treatment groups during the pilot study. These included the proportions of CHI entries corresponding to:

- 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;
- FR-days in which only payroll and no CHI entries were reported;

- FR-days in which both payroll and CHI were reported with some interviewing hours and miles recorded in which personal-visit attempts were made.

FRs may have changed their behavior in some instances. There were small reductions for Treatments 2 and 3 compared to Treatment 1 and Control in the proportion of cases with at least one firm refusal (which is defined to mean that the FR indicated the respondent was hostile, not interested, hung-up or slammed the door, or intended to quit the survey), or with two or more interim outcomes reflecting either a firm refusal or other respondent refusal.

Conclusions:

Based on the results observed during the pilot, 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. These results were roughly in line with the estimates on the 2012 ACS data from Griffin et al. (2015), though the design of the pilot differed somewhat from that earlier research. We recommend that the Census Bureau continue to prepare for a nation-wide implementation of the cumulative burden score and an associated stopping rule in the spring of 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.

1. Background

The American Community Survey (ACS) utilizes a multi-mode data collection approach that may attempt to contact nonrespondents by mail, telephone calls in the computer-assisted telephone interviewing (CATI) operation, and multiple personal visits or telephone calls during the computer-assisted personal interviewing (CAPI) operation. This data collection strategy can be perceived as overly intrusive to some respondents, and stakeholders have advocated on behalf of constituents with concerns about the burden associated with the number and type of contact attempts that have been made by the Census Bureau (Poe 2011).

Traditionally, organizations collecting survey data from households associate respondent burden with the frequency of a survey, and with the time and number of respondents required to complete it (Sears 2011, McCarthy 2011). A broader definition of respondent burden includes the negative feelings experienced by participants of surveys, including frustration, anger, annoyance, or boredom (Frankel and Sharp 1981). Expanding on this concept of burden to include the negative perceptions due to repeated contact attempts, the Census Bureau has undertaken research to reduce perceived contact burden through reducing the number and type of attempts.

To respond to concerns with perceived contact burden, in 2013, the Census Bureau analyzed the number of calls made to sample addresses by telephone call centers during the ACS CATI operation. This analysis estimated the potential effect on response rates and costs of changes to ACS CATI call parameters (Griffin and Hughes, 2013). CATI call parameter changes went into effect starting with the March 2013 Mail panel (April CATI data collection), and the changes were successful in reducing respondent burden by reducing total calls by about 20 to 25 percent.

In 2014 and 2015, the Census Bureau performed research to prepare for reducing respondent burden in the ACS CAPI operation. The Census Bureau documented the level of effort, respondent burden, costs and quality of the CAPI methods based on an analysis of 2011 and 2012 ACS CAPI paradata (Zelenak 2014 and Griffin and Nelson 2014). Additionally, an interdivisional group developed potential rules for determining when to stop contacting a sample unit (i.e. stopping rules) and, therefore, reduce the number of contact attempts made in the CAPI operation (Griffin 2014). The team then modeled the effects of the proposed stopping rules on respondent burden, cost, and quality (Griffin, Slud, and Erdman 2015).

Based on this work, the Census Bureau decided to pursue a field pilot employing a stopping rule based on a "cumulative burden score." This rule assigns a value to each contact attempt in any mode as a separate increment of burden, in order to estimate the perceived contact burden for the respondent. Once the cumulative burden score exceeds a threshold of 40, a level corresponding to the 95th percentile of cumulative burden among all CAPI cases in 2012 ACS data (Griffin et al. 2015), the case is pulled from the active CAPI workload, so no further contact attempts can be

made. We assigned burden score increment values and the burden score threshold were based on the research team's assessment of the relative burden of the various contact attempts.

For example, personal visits were assigned more burden points than telephone calls, contacts more points than noncontacts, indications of reluctance by the respondent more points than a lack of reluctance by the respondent, etc. Mail and CATI contact attempts establish an incoming burden score for CAPI, and the burden score increases for every CAPI contact attempt based on the type of attempt as recorded in the Contact History Instrument (CHI) and its outcome. Once the cumulative burden score exceeded the burden score threshold, the case was pulled from the active workload so no further attempts could be made.

Historically Field Representatives (FRs) relied on their judgment and feedback from their supervisors to determine when to stop making attempts to complete an interview. A CAPI stopping rule based on the cumulative burden score would reflect a significant cultural and procedural shift for the field staff. The Census Bureau determined that a pilot involving a subset of the ACS CAPI sample would help identify training and tools needed for FRs and their managers to prepare successfully for full implementation. A pilot would also allow the Census Bureau to confirm that the IT systems' changes necessary to implement the burden score and stopping rule were functioning as intended and to determine whether the impacts on cost, quality, and perceived burden were in line with the estimates produced in the 2015 research that modeled the impacts using historical data (Griffin, Slud, and Erdman 2015). The team was also interested in determining whether showing the current cumulative burden score to field staff would influence their behavior, so the pilot was designed with different experimental treatments that would assist in making a recommendation on whether or not to show the score to field staff. The Census Bureau would use the results from the pilot to prepare for deployment of the cumulative burden score stopping rule nationwide for the CAPI operations in the late Spring of 2016.

2. Methodology

2.1 Test/Pilot Overview

The pilot for the cumulative burden score stopping rule was conducted during the August 2015 ACS CAPI operations in roughly one quarter of field geographies in which ACS interviewing is conducted.

In order to implement a CAPI stopping rule to pull cases that exceeded the cumulative burden score threshold, changes were necessary to multiple data collection systems. We decided it would be feasible to employ a system that calculated the cumulative burden score once each day, after all paradata describing the contact attempts for the day had been transmitted by the FRs. The Unified Tracking System (UTS) is a reporting system that has the capability to combine data from multiple paradata and case management sources to calculate the burden score. After the

cumulative burden scores were calculated, the systems generated files that were sent to the FRs' laptops to pull any cases that had exceeded the threshold of 40 points. This process required all FRs to transmit twice each day on days that they worked—once at the start of the workday but after 8:00 AM local time (to pick up any files that would pull cases that had been determined to exceed the burden score), and once at the end of the day but before 12:00 midnight (local time) (to send in the paradata for the attempts made that day on their cases). Historically, FRs have been required to transmit once per day worked, at the end of their workday, therefore, transmitting twice per day reflected a change in procedures for FRs. Since these additional transmissions would enable cases to be pulled from the FR's workload at the start of their workday, it was unclear whether the FR's concern with losing cases might motivate the FR to not comply with the transmission policy. Therefore, the Census Bureau developed experimental treatments that would allow for an evaluation during the pilot of whether FRs were more or less likely to comply with the new transmission policy depending on whether they knew it could lead to cases being removed from their workload.

The Census Bureau employs the CHI to collect information about each CAPI contact attempt, and CHI data are the primary input in the calculation of CAPI cumulative burden values. The CHI launches automatically after the data collection questionnaire closes for each case, and the FR provides information about the specific methods used and results obtained for that contact attempt. It is possible for an FR to make an unsuccessful contact attempt without launching the data collection questionnaire, thereby not automatically launching the CHI. Field procedures require the FR to independently launch the CHI in these situations to record the information about the contact attempt, but there are limited tools for supervisors to determine whether FRs are following these procedures. Additionally, there is some ambiguity and subjectivity when filling out the CHI and classifying the methods and results of the contact attempt. As a result of these issues, the accuracy and consistency of CHI data are subject to FRs' underreporting or misreporting activities.

The pilot also tested whether or not to display the burden score to the FRs. There was some question as to whether FR or management behavior would change if the cumulative burden score was provided to the FR or their managers. Displaying the burden score on laptops might increase FR efficiency as they approached the burden score threshold. An experiment embedded within the 2010 Census that evaluated the impact of reducing the number of contact attempts allowed per housing unit showed some increases in effectiveness of the final contact attempt (Compton and Bentley, 2012). However, there was uncertainty about whether the final attempt recorded by decennial census enumerators was more effective or whether enumerators stopped recording unsuccessful contact attempts prior to their final entry. Displaying the burden score on laptops could encourage FRs to stop, underreport, or misrepresent contact attempts. Therefore, the Census Bureau decided to use a set of experimental treatments to examine this issue in the ACS

CAPI pilot. Additionally, the Census Bureau decided not to divulge to the field staff in the pilot the individual components that determined the burden score for a case.

2.2 Treatments

Field management designates geographic areas as follows: the country is divided into 6 Regional Office (RO) areas, New York (NY, number 22), Philadelphia (PHI, 23), Chicago (CHI, 25), Atlanta (ATL, 29), Denver (DEN, 31) and Los Angeles (LA, 32). Each RO is further divided into eight Survey Statistician Field Areas (SSFAs). Each SSFA is divided into 11-14 Field Supervisor (FS) areas, and multiple FRs work within each FS Area (FSA). The pilot for the cumulative burden score stopping rule was conducted in one quarter of the SSFAs. The remaining three quarters of field geographies followed the standard field protocol, forming a control group. Within the SSFAs selected for inclusion in the pilot, individual FSAs and all the FRs included in that FSA were assigned randomly to one of three experimental treatment groups:

> **Treatment 1 (two transmissions required, burden score not displayed, cases not removed):** FRs and FSs in this treatment were asked to follow the modified laptop data transmission procedures to transmit twice per day, but cases were not removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.
>
> **Treatment 2 (burden score displayed, cases removed):** FRs and FSs in this treatment were asked to follow the modified transmission procedures to transmit twice per day, cases were removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was displayed on FR laptops and in the Regional Office control systems.
>
> **Treatment 3 (burden score not displayed, cases removed):** FRs and FSs in this treatment were asked to follow the modified transmission procedures to transmit twice per day, cases were removed for exceeding the cumulative burden threshold, but the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.

In these treatment groups FRs, FSs, Survey Statisticians Field (SSFs), Survey Statisticians Office (SSOs), and other RO management received training on procedures for the pilot via memoranda. The FRs and FSs received training that described procedures only for the experimental treatment group to which they were assigned, whereas SSFs, SSOs, and RO staff received information on the procedures that would be followed in each of the treatment groups, since they supervise staff in all of the three treatments. SSOs were provided reports on transmission compliance for all FRs in the pilot and were instructed to follow up with SSFs and FSs to intervene when FRs did not comply with transmission procedures.

It is important to understand that FSAs and their associated FRs, not individual cases, were assigned to treatment groups. Therefore, when cases were reassigned to staff across FSA boundaries, those cases may have been worked under multiple experimental treatments.

All but two of the SSFAs selected for inclusion in the pilot were matched to control SSFAs, and these matched non-included SSFAs can together be viewed as a concurrent matched control group. Despite the original intentions with this matched design, all statistical comparisons in this report concern only contrasts among Treatments 1 to 3 within included SSFAs. In addition, many descriptive comparisons are made between Treatments 2–3 and the full set of concurrent non-matched controls from all non-included SSFAs, defined as:

> **Control (single transmission required, burden score not displayed, cases not removed):** FRs and FSs in this treatment were required to transmit only once per day on days worked. Cases were not removed for exceeding the cumulative burden threshold, and the cumulative burden score for each case was not displayed on FR laptops or in the Regional Office control systems.

2.3 Sample Design

The pilot study sampled 12 out of the 48 SSFAs nationally from the continental US.[1] The SSFA that covers parts of southern California and Hawaii was included, which led to some limitations for the analysis of the results from Hawaii as described in Section 3.4 of this report. Within the selected SSFAs, the aim was to randomly assign all FSAs to one of the three treatment interventions described in Section 2.2, Treatment 1 (2 transmissions only, cases not removed), Treatment 2 (2 transmissions, burden score displayed, cases removed), and Treatment 3 (2 transmissions, burden score not displayed, cases removed). By design, all FRs within each selected FSA were to be in the same treatment group.

The first stage of sample selection consisted of defining two sets of "matched SSFA pairs" within each RO, and then selecting at random one member from each matched pair for inclusion in the pilot. The intent of the matching was to find SSFAs as similar as possible with respect to ACS interviewing difficulty. This would allow comparisons of case completion and FR behavior metrics between the selected SSFAs and the matched pair controls that continued to use standard ACS practices for FR transmissions. The matching or clustering of SSFAs within ROs into similar pairs was based upon 12 criteria for each SSFA: changes in three key CAPI metrics for the effect of curtailing further follow-up based on the burden score, the proportions of ACS cases

[1] The Census Bureau desired to sample SSFAs in the continental U.S. only, as the second transmission in western time zones before midnight could occur after data processing.

in three performance clusters, and the proportions of ACS cases in six production strata. The change metrics calculated were the absolute reduction in interview completion rate, relative change in response rate, and relative difference in hours per case. Performance clusters have values of 1 to 3 to which geographic areas are assigned, indicating progressively greater difficulty of obtaining an interview. These clusters are defined from demographic, social, economic, and operational indicators known to be associated with the difficulty of obtaining interviews. Each county is assigned to one of six production performance strata, defined from the relative population density and interviewing difficulty in the PSU area. Each FR is assigned to the single production stratum for the FR's primary county, which is used to set the standards for interviewing hours per case for the FR.

Based on ACS 2012 CHI data, the changes in key CAPI metrics were calculated as part of the CAPI Phase 3 Research Report (Griffin, Slud and Erdman 2015) as if case follow-up had been terminated at the contact attempt in which the burden score first exceeded 40 for each SSFA. The change metrics calculated were the absolute reduction in interview completion rate, relative change in response rate, and relative difference in hours per case. In preparation for the August 2015 pilot, the SSFAs were clustered based on a computed "dissimilarity" score. We standardized pairwise squared distances within five categories of variables by mean pairwise distance squared within RO. The five categories were: (1) reduction in completion rate, (2) relative change in response rate, (3) relative change in hours per case, (4) vector of performance cluster proportions, and (5) vector of production strata proportions. After this standardization, the dissimilarity metric was the sum of the squared distances (1)-(5), computed pairwise between SSFAs within RO. Based on this dissimilarity metric, the SSFAs were clustered using the "aggregative clustering" function "agnes" within R (R Core Team 2015). Within each RO, two distinct pairs of nearest-neighbor SSFAs according to this clustering method were found. However, it was decided that two specific SSFAs (number 78,Washington, DC in the Philadelphia RO and number 77, Houston, TX in the Denver RO) had shown such a high impact of changes due to case termination using burden score in the earlier research (Griffin, Slud & Erdman 2015) that they should be included in the pilot study with certainty. These SSFAs were far from any others in their ROs with respect to the dissimilarity criterion described above and would have no corresponding "matched" control SSFAs. Thus, only 10 matched pairs of SSFAs were determined, two pairs in each RO other than Philadelphia and Denver and one in those ROs. The SSFAs, along with the indication of which SSFA in the pair was selected for inclusion in the pilot study are shown in Table 2.1.

Table 2.1 Selected pilot SSFAs and their matches

RO	Selected Pilot SSFA	Matched Control SSFA
ATL	2974	2972
ATL	2975	2978
CHI	2578	2574
CHI	2576	2577
DEN	3173	3172
DEN	3177*	
LA	3277	3273
LA	3275	3276
NY	2275	2276
NY	2277	2274
PHI	2374	2376
PHI	2378*	

*These SSFAs were designated as high impact and were selected to be in the pilot with certainty. They do not have a matched pair control group.
Source: American Community Survey Paradata, August 2015.

The next stage of the pilot study design was to assign individual FSAs within already selected SSFAs to treatment groups using a stratification idea with block randomization. Within each selected SSFA, all FSAs were ordered into successive tiers of three, with those FSAs ordered first which contained the lowest workload proportion in performance cluster 1, the easiest-to-complete cluster. If the number of FSAs within SSFA was not divisible by three, then the final tier contained fewer than three FSAs. Within each tier of FSAs within each SSFA, Treatments 1, 2, and 3 were sampled randomly (equiprobably) without replacement. Thus, in tiers with three FSA's, exactly one was assigned to each of the three treatment groups, while in tiers with one FSA, that FSA was randomly assigned to a treatment group chosen randomly from {1,2,3}. The assignment of FSAs to treatment groups is detailed in a memorandum (Olson 2015) from Timothy Olson, Acting Associate Director for Field Operations, to RO Directors.

2.4 Research Questions

The research questions evaluated in this report are as follows:

2.4.1 Workload

How many cases were pulled from the workload in the pilot test areas, by day of data collection and cumulatively, by treatment?

How many cases were pulled from the workload in the test SSFAs and FSAs by treatment?

What was the distribution of the total counts of cases pulled per FR at the end of the interview period by treatment?

What were the proportions of cases pulled in each of Treatments 2 and 3 by RO and by the three response rate performance clusters?

2.4.2 Perceived Contact Burden

What were the differences in the total national numbers of contact attempts, contacts, and contacts with reluctance between each of Treatments 1 to 3 and the Control group?

What were the differences between these counts for the treatment groups during the pilot and during the previous month of the same year (July 2015)?

What were the differences in distributions of cumulative burden scores for cases in the various Treatments?

2.4.3 Interviewing Hours and Miles

What were the differences in hours per case in the ACS workload, hours per completed interview, miles per case in the workload and miles per completed interview between each of the treatment groups and the control groups and between each of the treatment groups during the pilot and during the previous month?

2.4.4 Response Rates

What were the differences in CAPI response rates between each of the treatments and the control by FR performance cluster?

2.4.5 Field Operations

How well did FRs follow key test procedural changes such as transmitting at the end and start of each work day and stopping work on cases once the burden score threshold had been reached?

Was there evidence of systematic changes in how FRs followed procedures to record CHI data for each attempt because of the pilot interventions? Specifically, were there more CHI records marked as “not attempting contact” in the Treatments versus the concurrent controls?

Was there evidence that payroll and mileage data appeared to be inconsistent with CHI records? For example, what was the percentage of FR-days on which payroll hours were entered and no CHI entries are recorded, and what was the percentage of days on which payroll miles were recorded and no personal visit entries are recorded in CHI?

How many cases had CHI records (excluding inbound calls) made after a case exceeded the burden score threshold?

For all stopped cases, what was the distribution of final outcome codes for the case?

How did FR behavior in attempting to contact households change under the new stopping rule, as measured by changes in CHI data for contact attempts by time of day and contact strategies used?

How will the observed changes in total hours worked by FRs impact FR salaries, benefits eligibility, availability for other survey work, and field staffing overall?

2.5 Analysis Design

The formal design of the experiment described in Section 2.3 enables the research questions in Section 2.4 to be addressed and partially answered. Methodological aspects of the analysis strategy are next discussed in four subsections to explain how comparisons and conclusions can be reached with differing degrees of statistical validity.

The discussion of assumptions and limitations (Section 4) will revisit many of these same topics, since moderate sample sizes and data limitations make definitive answers to some of the research questions impossible based on this pilot study data.

2.5.1 Comparison Groups

Several comparisons can be made to estimate the mean effects of treatment interventions on response, burden, contacts and compliance. Specifically, we can compare Treatments 2 and 3 to the following:

1. Concurrent Treatment 1;
2. Concurrent matched controls;
3. All concurrent controls; and
4. The previous month (paired by same FSA).

The randomization of FSAs among Treatments 1, 2, and 3 allows these treatment groups to be directly comparable. The differences between treatment groups measure the extent to which the different treatment interventions lead to different results. While Treatment 1 FRs are asked to make an additional daily transmission, they are otherwise unaffected by the burden score pilot. Because of the randomization, we think that comparisons between Treatment 2 and 3 and Treatment 1 are best suited to estimate treatment effects and to test hypotheses.

A second possible comparison for the treatment interventions are the FSAs in the non-selected matched SSFAs who followed the standard field protocol. These matched non-selected FSAs are comparable to the treatment-group FSAs in sharing the same RO and being very similar with respect to characteristics used in the clustering (performance clusters, production strata, and estimated impact of case removal based on analysis of 2012 ACS data – see Section 2.3 for

details). Recall that two SSFAs chosen for the pilot did not have matched control equivalents, which complicates any comparisons between Treatments 1, 2 and 3 and controls.

The non-matched control FSAs that followed the standard field protocol are expected to have different characteristics from the selected treatment-group FSAs, so that the comparison of their metric outcomes (nonresponse, burden, contacts, and costs) with those of selected FSAs would be much less interpretable. Still, estimating outcomes for the concurrent control allows a descriptive comparison with treatments, but statistical testing for comparisons involving controls was not done given that SSFAs had not been asssigned to the control group by random selection. We are especially interested in whether the concurrent control outcomes are generally similar to those of Treatment 1.

Finally, comparisons can be made between outcomes from August 2015 and earlier ACS data collection months. The initial research plan identified data from July 2015, August 2014, and July 2014 as possible comparison months. We noticed that 2014 metrics differed substantially from 2015 among non-pilot SSF areas, and, as a result, decided to consider only July 2015 data for comparisons over time. Changes from 2014 to 2015 were likely due to changes in procedures and performance standards during the year that are not relevant to the pilot study.

We drew some comparisons with outcome metrics calculated for ACS CAPI data in July 2015 for the same FSAs within selected SSFAs that were included in the August 2015 pilot. Because FSAs had different cases and sometimes different FRs in July, the comparability of outcome data (whether at the case- or FR-level) across July and August 2015 is less than for the contrasts of Treatments 1, 2, and 3. As the results show, there do not seem to be important systematic differences across July 2015 and August 2015 ACS CAPI outcome metrics. Therefore, comparisons with the previous month for the same FSAs allow a general and useful check on changes due to the twice-daily transmission and (in Treatments 2 and 3) due to case removal.

2.5.2 Inference

It is of interest in this report to make statistically valid comparisons between the treatments. Although the study was designed as a randomized experiment, as described in Section 2.3, the methodology to draw inferences from an experiment with such a design is not standard because of the structure of the survey features.

Recent approaches to this type of analysis fell into two categories: design-based and model-based. Van den Brakel and Renssen (1998) proposed a finite-population design approach in which each set of treatment observations is viewed as a subsample of the larger sample and can be used to separately estimate the finite-population parameter. Differences of these finite-population parameter estimates are of interest, and an approximate sampling variance of the difference is derived. Alternatively, Van den Brakel and Renssen (2005) and Van den Brakel (2008) applied a measurement error model to the problem, which posits a true intrinsic value for

each individual. Their linear model assumed the individual's observed value was equal to the sum of the true value, a treatment effect, an interviewer effect, and random measurement error.

While the observations in the pilot study come from a survey in which observations are probabilistically sampled, for the purposes of this report we are not concerned that our sample is representative of the population as a whole. Our focus is on the internal validity of the inference and therefore our population of interest is only those observations in the pilot. The randomized design of the pilot study allows some conclusions regarding significance of observed differences of outcomes by treatment groups to be made without recourse to distributional assumptions or models. If instead, we intended to make inference relating to the population from which the sample was drawn, it would be necessary to incorporate the sampling probabilities, and as a result, the variances of the different estimates would be larger than what are reported in this report.

In the following sections, we describe three approaches to inference along with discussions of hypothesis test p-values and confidence intervals for the group-contrast statistics in this setting. The first is a design-based approach which parallels that of Van den Brakel and Renssen (1998), the second is a permutation test approach, and the third is a nonparametric superpopulation approach. We believe the hypothesis test p-values and test-based confidence intervals derived from the permutation test approach are based on the most general assumptions possible in this setup, and these are reported in the results section. Results for the other two approaches are discussed in Appendix 10.2.

We designed the pilot to create a paired concurrent control SSFA for all but two of the selected pilot SSFAs. This does make it possible to use a matched-pair type inference to compare Control units with Treatment 1-3 units, but we decided that this was not the best analysis strategy because a) the quality of the matched pairs was less than desired and b) we would have to exclude the two high impact test SSFAs that were not given a control SSFA match. Thus, we do not do any formal hypothesis tests to compare Treatment groups with control groups.

2.5.3 Design-based Inference

Our first approach to inference is design-based. Imagine that each unit in the pilot has a "potential" outcome (Neyman, 1923) for each of the three treatments, which under a null hypothesis is the same for all three treatments. Although we observe each unit under only one of the three treatments, we estimate the population quantity of interest for each of the three treatments and make comparisons.

The experimental treatments were assigned at the FSA level, so the design is approximately a randomized cluster survey, where each FSA represents a cluster. Viewing the three treatments as subsamples of the larger sample (under the null hypothesis), we treat the design as approximately a Simple Random Sample (SRS) of FSAs of fixed-size from the union of all FSAs within the

selected SSFAs. Note that this interpretation does not quite reflect the actual randomization design, since FSAs were first grouped into similar tiers of three within SSFAs. Under the SRS interpretation, we estimate the average outcome of interest by

$$\hat{\bar{Y}}_r = \frac{\sum_{i \epsilon S_T} M_i \bar{Y}_i}{\sum_{i \epsilon S_T} M_i}, \text{ where}$$

$\bar{Y}_i$ is the average outcome in FSA *i*, M_i is the number of cases in FSA *i*, and S_T is the set of FSAs with a certain treatment (e.g. Treatment 1). Note that $\hat{\bar{Y}}_r$ is a ratio estimator (*r* for ratio) in the survey sampling literature (Lohr, 2009). We estimate the variance of the estimator as

$$\widehat{var}(\hat{\bar{Y}}_r) = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \sum_{i \epsilon S_T} \frac{M_i^2 (\bar{Y}_i - \hat{\bar{Y}}_r)^2}{n-1},$$

where $\bar{M}$ is the average of M_i , n is the number of clusters in sample, and N is the number of clusters in population. This is a standard survey sampling formula for the variance of ratio estimator in a cluster sample of clusters with unequal size (Lohr, 2009).

While we have an expression for the variance of the estimated mean of a single treatment group, in order to draw inferences about the difference between two treatment groups, it is necessary to estimate the covariance between the two estimates. Under the null hypothesis that the average outcome is the same for each treatment group, the covariance term will be negative.[2] We have derived expressions for the covariance under the standard large-sample approximation used in producing variances for ratio estimators. This approximation leads to the expression for an estimator of the variance of the difference of two treatment group means by

$$\widehat{var}\left(\widehat{\bar{Y}_{r_T}} - \widehat{\bar{Y}_{r_C}}\right) = \frac{1}{n_T \overline{M_T}^2} \sum_{i \epsilon S_T} \frac{M_i^2 \left(\bar{Y}_i - \widehat{\bar{Y}_{r_T}}\right)^2}{n_T - 1} + \frac{1}{n_C \overline{M_C}^2} \sum_{i \epsilon S_C} \frac{M_i^2 \left(\bar{Y}_i - \widehat{\bar{Y}_{r_C}}\right)^2}{n_C - 1},$$

where the subscripts *T* and *C* generically refer to distinct treatment groups. This approximation is similar to one proposed by Van den Brakel and Renssen (1998). We consider results from the permutational approach the best reflection of the distribution of the test statistics under the actual experimental design, so the design-based results are not presented in the main body of the report. Instead, a comparison of the results from the design-based approach with that of the permutational approach is presented in Appendix 10.2.

---

[2] This is a result of the fact that we have a finite population to which we randomly assign treatments to all units. In the case of two treatments with a common mean μ we have that $\alpha \bar{Y}_T$ + (1-α)$\bar{Y}_C$ = μ , where α is the proportion of units assigned to treatment. This is true for any randomization of the unit assignments. The treatment and control group means are negatively correlated because if $\bar{Y}_T > \mu$, then necessarily $\bar{Y}_C < \mu$.

2.5.4 Permutational Inference

Under the null hypothesis that there are no differences between outcomes due to treatment, treatment assignments are simply labels assigned at random to observations. Therefore, calculating the value of a test statistic under each possible configuration of FSA treatment assignments and weighting them equally yields the exact distribution of the test statistic under the null hypothesis. This is true for any test statistic of interest. Comparing the observed test statistic to the exact distribution yields a p-value for the test of the null hypothesis along with a test-based confidence interval. This flexible method of inference is called a permutation test. See Zieffler, Harring, and Long (2011) for additional information.

The experimental design of the pilot (Section 2.3) randomized clusters of observations, FSAs, to treatment groups within blocks defined by workload proportion in performance cluster 1. In this setting, under the null hypothesis permutations of the treatment groups are created by re-assigning treatment groups to FSAs according to the design. The contrast we are interested in testing (1 versus the union of 2 and 3, or pairwise 1 versus 2, 2 versus 3, etc.) determines which FSAs' treatment assignments we permute. For example, for testing Treatment 2 versus Treatment 3, we calculate only permutations of the FSAs that received those two treatments in the pilot. Instead of calculating all allowed permutations, we utilize 100,000 Monte Carlo samples of those permutations, from which we estimate a close approximation to the exact distribution of the test statistic.

The two test statistics of interest for any two comparison groups are as follows:

1) Differences between group means during the test period,

$$\hat{\Delta} = \hat{\theta}_T - \hat{\theta}_C = \frac{\sum_{i \epsilon S_T} M_i \ \bar{Y}_i}{\sum_{i \epsilon S_T} M_i} - \frac{\sum_{i \epsilon S_C} M_i \ \bar{Y}_i}{\sum_{i \epsilon S_C} M_i},$$

where $\bar{Y}_i$ is the average outcome in FSA i, M_i is the number of cases in FSA i, and S_T and S_C denote two sets of FSAs for comparison (e.g. Treatment 2 vs. Treatment 3).

2) Difference of differences of group means,

$$\hat{\Delta} = \hat{\theta}_T - \hat{\theta}_C = \frac{\sum_{i \epsilon S_T}[(M_{1i} + M_{2i})] \ (\bar{Y}_{1i} - \bar{Y}_{2i})}{\sum_{i \epsilon S_T}[(M_{1i} + M_{2i})]} - \frac{\sum_{i \epsilon S_C}[(M_{1i} + M_{2i})] \ (\bar{Y}_{1i} - \bar{Y}_{2i})}{\sum_{i \epsilon S_C}[(M_{1i} + M_{2i})]},$$

where $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$ are the average outcome in FSA i during August (the test period) and July, M_{1i} and M_{2i} are the number of cases in FSA *i* during August and July, and S_T and S_C denote two sets of FSAs for comparison (e.g., Treatment 2 vs. Treatment 3).

When the groups of observations used for comparison are similar, the differences between group means during the test period will yield valid inferences, but when the comparison groups are not

similar the inferences may be biased. Using the difference of differences of group means test statistic helps to reduce that bias by using the previous month's FSA averages as a control. In general, we found that these two test statistics agreed. Therefore, we mainly report results using the differences between group means during the test period. Results concerning the difference of differences of group means are listed in Appendix 10.1 but referenced where necessary in the main body of the report. Note that we report corrected Monte Carlo p-values (Davison and Hinkley, 1997), which consist of the ratios (k+1)/(R+1) where R=100,000 is the number of sampled permutations and k the number of permutations resulting in test statistics as or more extreme than the observed test statistic.

2.5.5 Nonparametric Superpopulation Inference

Our third approach is a non-parametric rank-based approach. Let the unit of analysis be the FSA and the variable of interest be either (a) the FSA-level outcome or (b) the change in the FSA-level outcome from July to August. Using the outcome (a) will result in a comparison using statistic 1) in Section 2.5.4, and using outcome (b) will result in a comparison using statistic 2). Let Y_i be the variable of interest for FSA i. We assume that the FSA-level outcomes within a treatment group are independent observations from a common continuous distribution. If we further assume that observations between treatment groups are independent, then we can test the null hypothesis that the treatment group distributions are equal using the rank-sum test (Hollander and Wolfe 1999, Chapter 4). If we also assume the distributions differ by a shift parameter then we can obtain an estimate of that difference and an associated confidence interval. A comparison of the results from this approach with that of the permutational approach and the design-based approach is presented in Appendix 10.2.

2.5.6 Measurement Issues from CHI

Measurement issues concerning FR behavior and incentives had to be addressed in this pilot using ACS CHI data, which may contain random measurement error as well as measurement error dependent on treatment. Two previous research studies at the Census Bureau provide baseline levels of CHI reporting compliance or systematically explore the methodology of measurement of FR behavior using the CHI instrument on Census Bureau administered surveys (Bates et al. 2010, Virgile 2015). Bates et al. (2010) found that overall, the CHI data are of sufficient quality for analysis, but there is some evidence to suggest FRs may be underreporting contact attempts, noting the number of missing paradata records due to instrument design. Based on that research, as well as comments from Field Division staff, a redesigned instrument was implemented in January 2014 to reduce the potential for underreporting contact attempts. A comparative analysis looked at the CHI data quality under the redesign, concluding the redesign minimized the potential for under reporting in the data used in this pilot (Virgile, 2015). Additional research is needed on the behavioral aspects of FR CHI reporting and their relation to local management practices at the FS, SSF, and RO levels.

Anecdotal evidence from the Census Bureau's Field Division suggests that FR contact attempt reporting in CHI systematically omits or underweights certain categories of attempts, especially "observed household from vehicle" and other types of attempts that find no one at home. Additionally, analysis of CHI case-records by Field Division staff and other researchers indicates the occasional use of "not attempting contact" CHI entries to correct and re-characterize previous entries. As we describe under Limitations below, inaccurate CHI reporting of these and other types distorts the calculation and interpretation of the burden score and weakens the conclusions that can be drawn from the pilot study.

3. Data

3.1 Data Sources

The analysis of this pilot study used CAPI paradata from the month the burden score was piloted, in August 2015, as well as July and August 2014 and July 2015 for comparison purposes. Several sources of paradata were combined when creating the analytic dataset. CHI data provide information from the FR pertaining to each contact attempt made to the sample unit as well as additional levels of effort not classified as contact attempts (e.g. locating activities, ACS geocoding). For comparative purposes, the number of times the FR reported "not attempting contact" was also included as a metric for analysis. Additional information pertaining to each contact attempt included whether and with whom the FR reported making contact; any reported respondent reluctance, behaviors, or concerns; and FR strategies, as all of these events reported in CHI incremented the burden score.

Because the first two modes of data collection increment the burden score prior to CAPI data collection, we first had to merge the CHI data records with UTS paradata in order to determine how each contact attempt increased the burden score. The UTS paradata pertaining to mail and CATI operations were used to calculate the initial burden score of each case *prior* to the first CAPI contact attempt. Using the initial burden score as the starting point for each case, we sequentially ordered the recorded contact attempts for each case to re-construct the cumulative burden after each attempt.[3] Because UTS was not calculating the burden score in 2014, initial burden scores were only available for CAPI cases worked in July and August 2015.

The sequenced data were also used to determine the FRs' compliance with the transmissions protocol for the test. UTS provided a separate dataset for each day in the interview period.

[3] UTS did not retain the daily burden score calculations, thus requiring the reconstruction. Our total re-constructed burden score matched the total burden score from UTS in 99.7 percent and 99.5 percent of cases in August and July of 2015 respectively. The reason for the discrepancy in the small number of cases is unclear at this time.

Available only for August 2015, these paradata included minimal information, including only the burden score that was displayed for Treatment 2, and a record indicating the stop work order was sent for cases in Treatments 2 and 3. Combining the sequenced data with the UTS daily burden score calculations and stop work orders provided an indicator of FR transmission compliance. If the FR did not comply with the daily transmissions protocol on a day when the burden score exceeded the threshold, UTS would not have calculated the daily burden score and issued a stop work order to pull the case from the FR's laptop.

At the case-level, additional metrics required additional data sources. The Regional Office Survey Control System (ROSCO) provided the final disposition of each case. The performance cluster, the rating of difficulty gaining respondent cooperation in that particular area based on sociodemographic characteristics related to survey response, was also merged onto the data at the case-level. These data are available based on the geocode assigned to the case. After merging additional geocoding done by FRs during data collection, 0.3 percent of cases were unable to be geocoded and sufficiently matched to receive a cluster rating.

The Cost and Response Management Network (CARMN) provided information at the FR-level, including the FR's assigned FS and SSF area codes. Additionally, this paradata source provides the hours and miles charged daily by each FR to specific project and task codes. For the purposes of this research, interviewing hours charged to two project codes were summed to determine, first, how many hours the FRs charged for the new task of daily transmissions and, second, the total hours and miles charged for interviewing ACS cases. Because CHI and CARMN are not linked, these data require merging based on FR and the calendar date from the CHI time-stamp or the reported date of hours/miles charged in the payroll system. Of the 180,323 records for each FR for each day (in July and August 2014 and July and August 2015), 63 percent of the days have matching CHI and payroll entries, 18 percent of the days have only CHI entries, and 19 percent have only payroll entries. Among the 34,674 FR-day records associated with the August 2015 interview period, none have only CHI entries, and 80.7 percent have matching CHI and payroll entries.

With respect to transmission compliance, the transmissions time-stamps from July 31 to August 7 may not be accurate because of ROSCO errors related to the time zone, which were supposed to have been local to the FR but erroneously recorded as Eastern Time. We processed the data including the affected time-stamps but checked the effect on our tabulations of restricting attention only to transmissions beginning on August 8, 2015. As indicated in the transition between Tables 5.18 and 5.19, compliance rates appear three to four percent higher in the post-August 7 data, but the pattern of contrasts across treatments is approximately the same.

In general, our analysis excluded cases and attempts from units determined to be ineligible for ACS interviews. However, these cases and attempts were included in any analyses and results

that deal with cost, such as FR hours per case or FR miles per case. In both of those situations, the total number of cases, including ineligible non-interviews, was used as the denominator.

3.2 Pulled Cases

During the pilot, when a case was pulled from the workload due to surpassing the burden score threshold, sometimes the potential respondent either had a scheduled appointment with an FR or called the telephone assistance line. In this situation, the potential respondent was instructed to complete the interview on the telephone or themselves, by mail or online. For this reason, some cases that might have been non-interviews because of surpassing the burden score were instead converted to LMR (Late Mail Return). In this report, we use the term "pulled" cases to denote those cases that exceeded the burden score, were pulled from the workload, and remained non-interviews.

3.3 Definition of a Contact Attempt

For the purposes of this research, a contact attempt was defined as any CHI entry for which an FR reported making a personal visit or telephone contact attempt or the outcome code was a respondent refusal. These events are associated with increases in the burden score. In August 2015, 77 percent of all CHI entries met this definition of a contact attempt. We note that an outcome code of refusal can be entered even with no indication of actual telephone or personal-visit contact. The purpose of such CHI attempt entries that are refusals but indicate no contact-person is often the final wrapping up or coding of a case, meaning that they are entered by the FR at a time after an attempt. In August 2015, 1.6 percent of contact attempts had no specific indication of being either a telephone or personal visit.

3.4 Case Assignment

Census Bureau data collection procedures include case reassignment, which can be across FSAs, and were a limitation of this design. In August 2015, thirteen percent of cases had multiple FRs that recorded contact attempts in CHI, and three percent of cases at least two FRs from separate FSAs that recorded contact attempts. The treatment groups were assigned at the FSA-level, and reassigning a case across FSAs could, therefore, result in a change in the case's treatment group. The planned analyses required that each case be assigned to just one FR, and, therefore, one treatment. We assigned each case to the FR who made the majority of the contact attempts. In the case of a tie in the number of contact attempts, we assigned the case to the FR with the first contact attempt. In the case where no contact attempts were made, we assigned the case to the FR with the first CHI entry. All cases had CHI entries, regardless of whether or not contact attempts were made. Based on the assigned FR, the case's assigned FSA, SSFA, and treatment group was determined. We also considered assigning cases to the FR who contributed the

majority of burden points to the case. For the overwhelming majority (97%) of cases, both methods assigned the case to the same FR.

For tabulations related to individual contact attempts, the individual FR and associated treatment group are unambiguous, and such tabulations are made in terms of the actual FR and treatment group for that attempt. For tabulations related to FRs, e.g. in connection with FR transmission and cost-reporting compliance, there is also no need for a unique association between cases and FRs, and such tabulations are made for all FR attempts on all cases they worked. However, in tabulations classifying cases by treatment group, each case was assigned a unique FR (and therefore a unique treatment) as described in the previous paragraph.

3.5 Other Data Notes

Several data anomalies, somewhat contrary to the intended design of the pilot study, are noted here and will be revisited as part of the discussion of limitations of the analyses of this research.

Since cases were sometimes reassigned across FSA and even in a few rare instances across SSFA boundaries, the association between case and treatment group was not strictly maintained in the pilot study. That is the reason for the definition described in Section 3.3 of a single case FR assignment (i.e. the FR with the largest number of CHI attempts, with attempt defined as in Section 3.2 to include only CHI entries that were either personal or telephone attempts or had an outcome code indicating respondent refusal.)

Three FSAs (317396, 317790, and 237891) were within SSFAs selected for inclusion in the pilot study but were not part of the FSA lists used to randomize assignment of FSA to treatment group. These FSAs respectively contained two, one, and one FRs and accounted for 20, seven and six cases in August 2015. FRs in these FSAs were not subject to case removal based on burden score, so data from these FSAs were included in concurrent control tabulations but not in Treatments 1 to 3.

Two FSAs, 7388 from the Denver RO and 7586 from the New York RO, had no assigned cases in August 2015. Both FSAs were assigned to Treatment 3. In both FSAs, there was no supervisor for all or part of August, and FRs were temporarily assigned to other FSAs. FSA 7586 in the New York RO did have cases and attempts for July 2014, August 2014, and July 2015 (respectively 195, 171, and 266 attempts). Due to the vacant supervisor position, the FRs in this FSA reported to the supervisor in FSA 7584 of the New York RO, also in Treatment 3, and were coded as such in the data. Therefore, for the purposes of this study, we include these FRs in analysis as though they belonged to FSA 7584 even though they were geographically located in another area. FRs in FSA 7388 of the Denver RO were reassigned to several FSAs for at least the duration of July and August 2015, but, within the CHI data, there was a single FR (found to be the FS) labelled as belonging to FSA 7388 with no assigned cases and only a few attempts. We

do include this FR's attempts at the case-level but do not include the contributions from this FR in FS- and FR-level analyses.

We removed two contact attempts from the analysis dataset from August 2015. These two contact attempts were recorded as happening on September 8 and 9 respectively but had final outcomes listed in August.

The FR transmission procedures for this pilot were designed to for the FR to transmit paradata reflecting that day's efforts by 12:00 midnight local time. This allowed the data to be received in time for overnight processing to calculate updated burden scores. The timing of this overnight processing was designed to accept all midnight Pacific Time transmissions at 3:00 am, Eastern Time. Several FSAs (7782, 7783, 7784, and 7785) partially covering Hawaii were selected for inclusion in the pilot study. These are anomalous because their time zone makes their end-of-day and following start-of-day transmissions fail to bracket overnight UTS processing, and, therefore, the results in these FSAs do not reflect the intended timing for updating the burden scores.

3.6 Baseline Tabulations

This section consists of tables with information about the number of cases by treatment group, month, RO, SSFA, performance cluster, and initial burden score. These tables are useful for reference in future sections of the report.

Table 3.1 Cases by treatment and month

Month	Control	Treatment 1	Treatment 2	Treatment 3
August, 2015	44,911	4,299	4,135	4,213
July, 2015	45,016	4,229	4,147	4,245
August, 2014	44,424	4,458	3,954	4,258
July, 2014	44,594	4,508	3,994	4,218

Source: American Community Survey Paradata, July and August 2014, July and August 2015.

Table 3.2 Cases by treatment and regional office, August 2015

Regional Office	Control	Treatment 1	Treatment 2	Treatment 3
New York	7,236	731	407	525
Philadelphia	6,709	529	560	590
Chicago	7,088	943	1,061	841
Atlanta	7,550	847	822	703
Denver	9,914	670	702	678
Los Angeles	6,414	579	583	876
Total	44,911	4,299	4,135	4,213

Source: American Community Survey Paradata, August 2015.

Table 3.3 Cases by SSFA by treatment among those in test FSAs, August 2015

SSFA	RO	Treatment 1	Treatment 2	Treatment 3
2275	NY	286	161	165
2277	NY	445	246	360
2374	PHI	355	352	391
2378	PHI	174	208	199
2576	CHI	559	618	438
2578	CHI	384	443	403
2974	ATL	392	451	340
2975	ATL	455	371	363
3173	DEN	368	368	318
3177	DEN	302	334	360
3275	LA	298	303	542
3277	LA	281	280	334
Total		4,299	4,135	4,213

Source: American Community Survey Paradata, August 2015.

Table 3.4 Cases by performance cluster by treatment, August 2015

Performance Cluster	Control	Treatment 1	Treatment 2	Treatment 3
1	28,288	3,125	3,017	2,977
2	13,479	1,079	1,024	1,149
3	3,064	86	87	74

Source: American Community Survey Paradata, August 2015.

Table 3.5 Initial burden score distribution by treatment, August 2015

Initial burden score	Control	Treatment 1	Treatment 2	Treatment 3
0	9.81%	7.93%	6.41%	6.36%
4 to 6	50.16%	51.08%	53.74%	53.10%
12 to 20	35.55%	35.75%	35.19%	35.63%
24 to 25	4.48%	5.23%	4.67%	4.91%

*Scores of 0, 4-6, 12-20, 24, and 25 are the only possible initial burden scores

Source: American Community Survey Paradata, August 2015.

Table 3.6 Initial burden score distribution by performance cluster, August 2015

	Initial burden score			
Performance cluster	**0**	**4 to 6**	**12 to 20**	**24 to 25**
1	12.82%[4]	48.10%	34.56%	4.52%
2	2.60%	54.92%	37.74%	4.74%
3	1.39%	57.75%	36.30%	4.56%

Source: American Community Survey Paradata, August 2015.

Table 3.7 FSAs and FRs by treatment, August 2015

	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.

[4] The reason for the increased percentage of "0" initial burden scores in Performance Cluster 1 over the other clusters is that Cluster 1 cases are more likely to be rural than the other clusters. Rural cases are more likely than urban cases to lack quality address and phone information used in the mail and CATI phase of non-response follow-up. As a result, a larger percentage of Cluster 1 cases will not have any mail or CATI follow-up and thus have an initial burden score of "0".

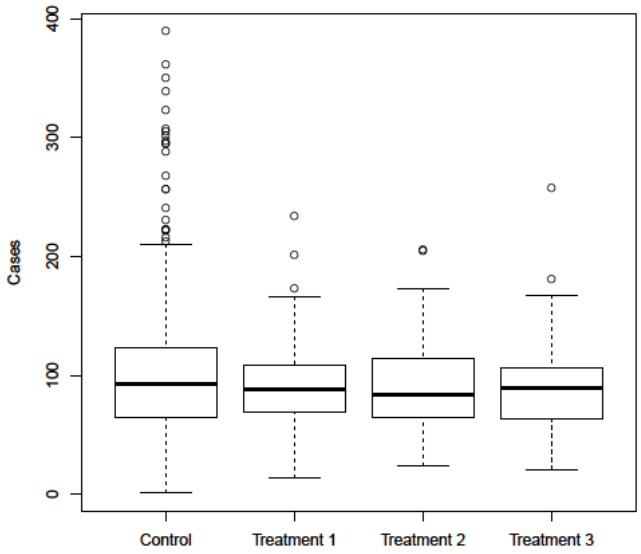


Figure 3.1 Boxplot of number of FSA-assigned cases by treatment, August 2015
Source: American Community Survey Paradata, August 2015.

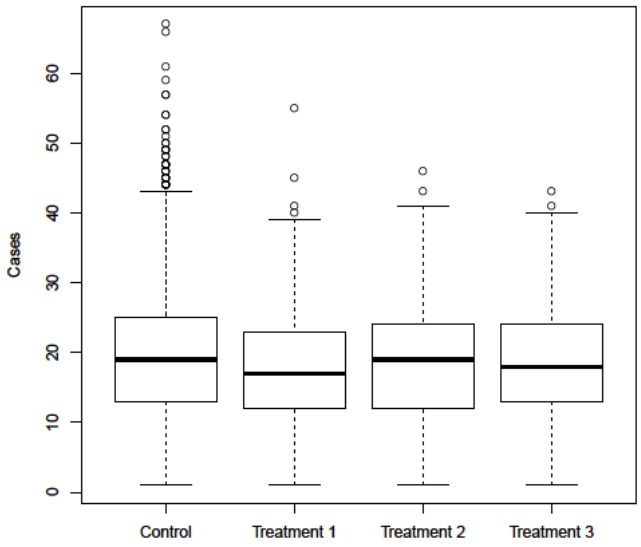


Figure 3.2 Boxplot of number of FR-assigned cases by treatment, August 2015
Source: American Community Survey Paradata, August 2015.

4. Assumptions and Limitations

The analysis and conclusions presented in this report are impacted by the following assumptions and limitations:

1. Designing an experiment under current data collection policies and procedures was difficult due to:
    a. The occasional multiple treatment assignments of cases reassigned across FSAs
    b. Insufficient baselining for statistical inferences of the pilot. There is limited detailed descriptive knowledge of current FR CHI recording behavior against which to compare the findings of the pilot. We also do not have a sense of the stability of the CHI recording behavior by RO and SSFA.
    c. The impossibility of stratifying treatment assignments on initial burden scores (derived from mail and CATI operations) which varied widely

2. CHI was used as the main data source for calculating burden scores and for many of the analyses in this report, but CHI data can be misreported by FRs, either intentionally or unintentionally. Although this report found limited evidence of CHI data inconsistencies that would indicate widespread intentional misreporting of CHI data, anonymous debriefing questionnaires completed by FRs at the conclusion of the test indicated that some FRs did misrepresent their CHI entries to prevent cases from being removed from their workload.

3. Transmission compliance was much lower during the pilot than was needed in order to ensure that burden scores were accurately updated each day and cases were pulled in a timely manner. Errors were identified during the pilot in the reports managers had available to monitor compliance, reducing their ability to intervene when FRs were not following procedures, which potentially contributed to these low rates.

4. Methodological limitations include:

    a. Lack of close comparability between the members of matched pairs of SSFAs, because the criteria used in matching are very far from being able to capture all of the important geographical and demographic differences between SSFAs.

    b. Violations of the treatment assignments due to reassignments across treatment groups and a failure to include some small FSAs in some SSFAs in the treatment-group assignments. These violations, along with other data errors detailed in Section 3, result in small failures of randomization-based comparability between treatment group assignments.

    c. Deficiencies in the way that the burden score calculation, as implemented by UTS, scores “not attempting contact” CHI entries with indication of a respondent refusal, especially entries that repeat or correct previous entries.

d. Anticipated errors, including both inconsistencies and outright measurement errors, in many of the reported CHI variables. This may be less important in the kinds of contrasts observed in treatment group outcomes than in the overall baseline information collected via CHI, on the basis of which FR behavior with respect to compliance and imposition of respondent burden can be monitored and managed.

e. Low power for several comparisons needed in this report. The low power arises primarily because (a) the burden score threshold was designed to occur relatively rarely and not expected to cause drastic changes in metrics of contact or response, and (b) we are interested in changes in metrics that would have to be large to indicate clear changes in FR behavior between Treatments 2 and 3 to justify whether it is better or not for FRs to see current burden scores displayed on their laptops. For these reasons, we have highlighted a number of apparent treatment-group contrasts that are suggestive but not statistically significant, and it is important that future research continue after nationwide implementation to assist in effective management of FR transmission and CHI reporting compliance.

5. The number of distinct FSAs randomized within this pilot study, 11-12 in each of the six selected SSFAs, is relatively small, and it is difficult to distinguish true effects because of variation caused by:

   a. RO and SSFA policies with respect to monitoring of FR CHI reporting of interim non-interview outcomes, authorization to stop working on a case, and reassignment practice;

   b. Local and geographic conditions;

   c. Imprecision of the single-treatment assignments in light of multiple FRs working single cases; and

   d. Other influences on FR behavior such as initial burden scores derived from case handling in ACS mail and CATI or local differences in the occurrence of vacant sampled units.

6. The large number of research questions and modest number of randomized FSA treatment assignments in the pilot makes this report necessarily exploratory rather than a source of definitive answers to questions of interest to ACS management in preparing a national rollout of the pilot's policy of curtailing repeated follow-up attempts on high-burden cases. The related statistical methodological issue is the problem of multiple comparisons.

7. The exact mode of calculating the burden score in systems during this pilot differed slightly from that of the Phase 3 results detailed in the Griffin et al. (2015) CAPI report. In the pilot a burden increment of 15 was assigned to CHI attempts with an indicator for "not attempting contact" occurring together with a respondent-refusal interim (as distinguished from final) outcome code. This type of burden increment was not used in Phase 3 results. As a result, some of the pilot results may be slightly different than expected. We suggest revisiting this part of the burden score calculation in the future.

8. The analysis of interviewing hours and miles does not reflect the full cost impact of implementing the cumulative burden score stopping rule into production. The additional daily transmission was done at an additional cost to the program during the pilot, and this report does not assess any impact on team leading charges during the pilot.

9. We assigned burden score increment values and the burden score threshold based on our team's assessment of the relative burden of the various contact attempts and not on any empirical evidence or input from ACS respondents to validate their perceptions of the relative contact burden for various contact methods.

10. The short duration of the August pilot did not allow any research questions to be studied related to FR learning of effective approaches to case follow-up when burden is not allowed to move beyond the burden score threshold. FRs may behave differently than in the pilot once they have a better understanding of the burden score threshold. In the pilot, FRs did not know anything about the detailed definition of burden score increments or the threshold of 40 even if (in Treatment 3) they saw the scores on their laptops. Similarly, FRs operating under the maximum burden threshold in the future will be responding to performance and pay incentives different from those in place during the pilot. For example, FRs may behave differently depending on whether they find that they can use working hours saved on lost high-burden cases on other cases in their workload.

11. Due to incomplete data, we were not able to investigate the frequency or characteristics of cases that were pulled from the active workload because of high burden scores but subsequently converted to late mail returns as a result of the respondent completing the survey by telephone, internet, or mail.

5. Results

In the following section, we separate our results by topic into workload, burden, cost, quality, and field operations. We also include a discussion of the significance testing[5] for the given results at the end of this section. All hypothesis test p-values are two-sided.

As discussed in Section 4, our hypothesis tests have low power to detect small differences between treatment groups. For this reason we have highlighted a number of apparent treatment-group contrasts that are suggestive but not statistically significant, and it is important that future analyses continue after nationwide implementation to assist effective management of FR transmission and CHI reporting compliance. In addition, the "multiple comparisons problem" (Casella and Berger, 2002), associated with the large number of hypotheses tested in this research, means that even hypothesis test results that meet the Census Bureau standard of stand-alone significance at level 10 percent may be erroneous simply due to the large number of comparisons being tested. A simultaneous presentation of many of the hypothesis test results in this section can be found in Section 5.6, where a permutationally-based adjustment due to the multiplicity of hypotheses tested is also proposed.

5.1 Workload

Overall, 380 cases were pulled during the pilot (Table 5.1) due to a high cumulative burden score, which included 187 cases assigned to Treatment 2 and 171 assigned to Treatment 3. A small number of cases assigned to Control and Treatment 1 were pulled, three and 19 respectively, which should not have happened but was due to the reassignment of cases across treatment groups. For example, if a case was primarily worked on by a Treatment 1 FR, but was reassigned to a Treatment 2 FR near the end of the month and subsequently pulled, the case would still be assigned to the Treatment 1 FR who contributed the majority of contact attempts. The percent of cases pulled from Treatment 2 and 3 were 4.5 percent and 4.1 percent respectively. In total, 4.3 percent of cases were pulled from Treatments 2–3.

[5] In the results, we make statistical comparisons only between Treatments 1, 2, and 3. This is because in the design, pilot FSAs were randomized between those treatments. We do not statistically test comparisons between the control group and the other treatments because the assignment of SSFAs to the control group was not done entirely at random.

Table 5.1 Pulled cases by treatment group

	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.

The initial burden score for a case was a contributing factor in whether the case would ultimately be pulled. Table 5.2 shows that the percent of cases pulled increased as the initial burden score increased. Among cases with an initial burden score of 0, less than one percent of cases were pulled for both Treatment 2 and Treatment 3. In comparison, 19.7 and 14.0 percent of cases were pulled among those with an initial burden score of 24 to 25, respectively for Treatment 2 and Treatment 3. Note that a majority of cases had an initial burden score of 4 to 6 or 12 to 20, and only about 5 percent of cases have an initial burden score at each of the extremes (Table 3.5).

Table 5.2 Percent of cases pulled among total cases by initial burden score and treatment group

Initial burden score	Treatment 2 % of cases pulled	Treatment 2 Total Cases	Treatment 3 % of cases pulled	Treatment 3 Total Cases
0	0.8%	265	< 0.1%	268
4 to 6	2.4%	2,222	2.5%	2,237
12 to 20	6.5%	1,455	5.7%	1,501
24 to 25	19.7%	193	14.0%	207

Source: American Community Survey Paradata, August 2015.

Figure 5.1 shows the frequency of pulled cases by day of the test month. Pulling of cases occurred mostly during the middle portion of the month, with the largest numbers pulled on August 20 and 21. Treatments 2 and 3 each followed similar patterns (not shown separately) of pulled cases by day of the month. Note that many ROs have a goal of closing out cases either completely or partially by the 21st or 25th of each month, which could explain the pattern.

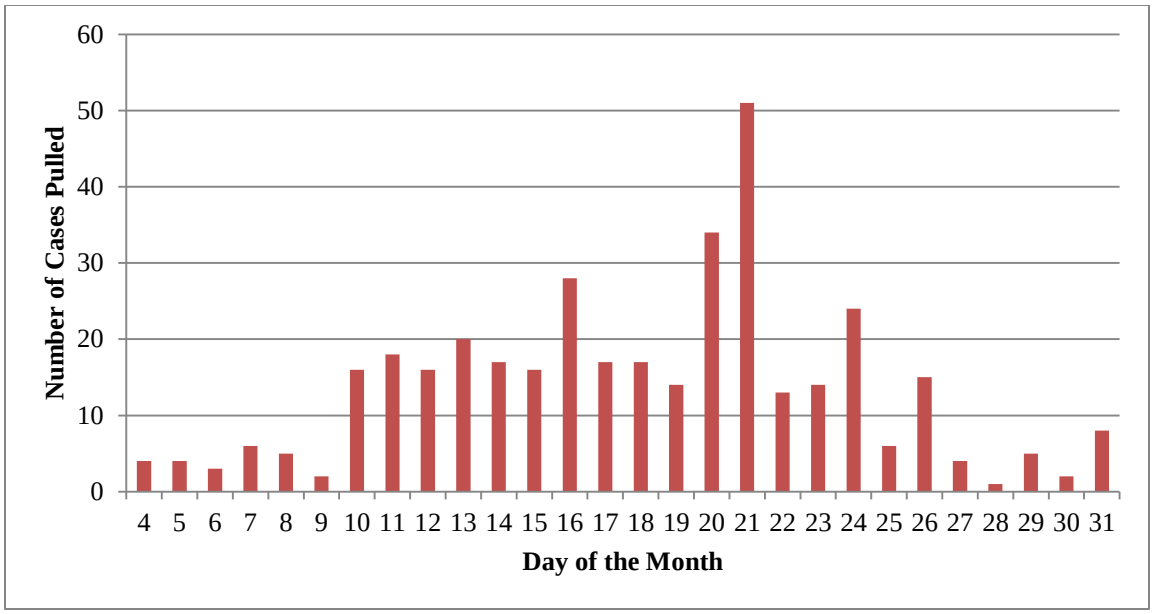


Figure 5.1 Frequency of pulled cases by day of the test month.
Source: American Community Survey Paradata, August 2015.

Table 5.3 summarizes cases pulled by SSFA by treatment group. We find that all SSFAs involved in the pilot experienced pulled cases in both treatment groups 2 and 3. The SSFA percentage of cases pulled in Treatment 2 varied from 2.7 percent to 10.6 percent, and the percentage of cases pulled in Treatment 3 varied from 2.1 percent to 7.0 percent. For reference,

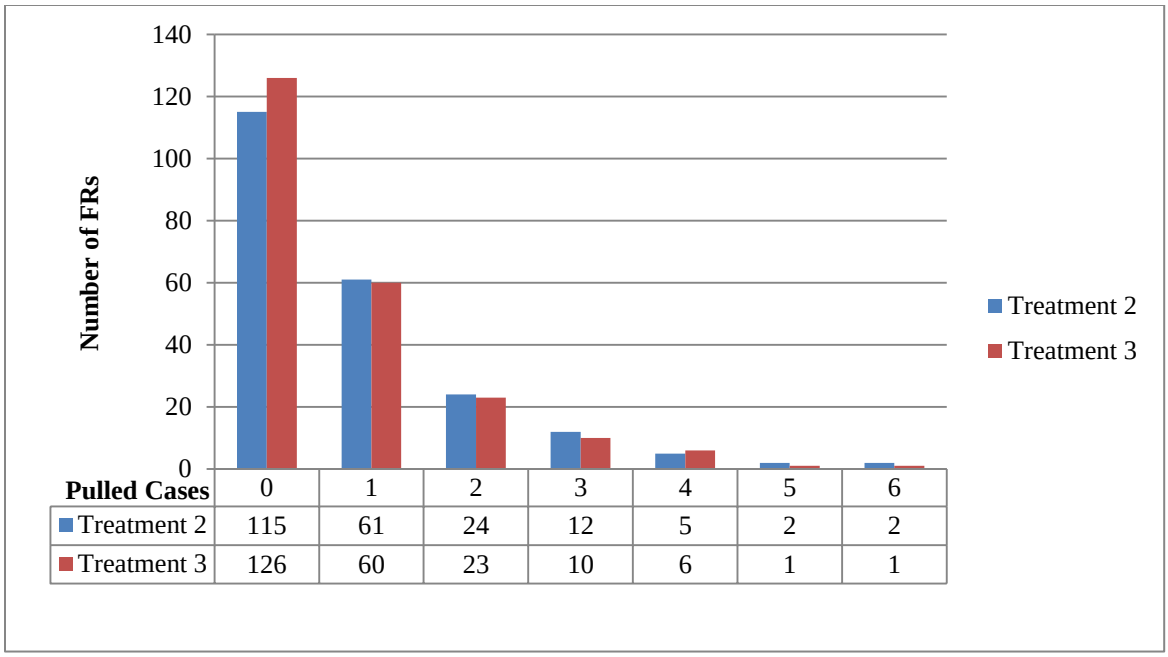
Table 3.3 gives the total number of cases in each SSFA by treatment group. The percentage of cases pulled in Treatments 2 and 3 varies even more widely when the cases are split into their respective FSAs. The percent of cases pulled from Treatment 2 FSAs varied between 0 and 12 percent with a median of four percent, and the percent of cases pulled from Treatment 3 FSAs varied between 0 and 14 percent with a median of four percent.

Table 5.3 Cases pulled by SSFA by treatment group

		Cases Pulled		Percentage of Cases Pulled Among Total Cases	
SSFA	**RO**	**Treatment 2**	**Treatment 3**	**Treatment 2**	**Treatment 3**
2275	NY	13	6	8.1%	3.6%
2277	NY	8	10	3.3%	2.8%
2374	PHI	12	8	3.4%	2.1%
2378	PHI	22	14	10.6%	7.5%
2576	CHI	24	21	3.9%	4.8%
2578	CHI	12	17	2.7%	4.2%
2974	ATL	22	8	4.9%	2.4%
2975	ATL	10	15	2.7%	4.1%
3173	DEN	13	13	3.5%	4.1%
3177	DEN	14	21	4.2%	5.8%
3275	LA	22	24	7.3%	4.4%
3277	LA	15	14	5.4%	4.2%

Source: American Community Survey Paradata, August 2015.

Figure 5.2 shows the number of FRs by pulled cases by treatment. For example, 61 FRs from Treatment 2 and 60 FRs from Treatment 3 had exactly one case pulled. 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 for Treatment 2 and Treatment 3, did have one or two cases pulled. In the pilot, at most six assigned cases were pulled from any FR.



Pulled Cases	0	1	2	3	4	5	6
Treatment 2	115	61	24	12	5	2	2
Treatment 3	126	60	23	10	6	1	1

Figure 5.2 Number of FRs by pulled cases by treatment

Source: American Community Survey Paradata, August 2015.

Performance cluster is a measure of difficulty of a case, with cases in Cluster 1 being the least difficult and those in Cluster 3 the most difficult. In many displays, Clusters 2 and 3 are combined as a single Cluster 2–3. In both Treatments 2 and 3, the percent of cases pulled increases in the performance clusters associated with higher difficulty. In Performance Cluster 1, 3.9 percent of cases in Treatment 2 and 3.5 percent of those in Treatment 3 were pulled. By comparison, 6.1 percent of Treatment 2 and 5.4 percent of Treatment 3 cases were pulled among cases in Performance Cluster 2–3.

Table 5.4 Percentage of cases pulled among total cases by performance cluster and treatment

	Total Cases		**Percent of Cases Pulled**	
Performance Cluster[1]	**Treatment 2**	**Treatment 3**	**Treatment 2**	**Treatment 3**
1	3,017	2,977	3.9%	3.5%
2–3	1,111	1,223	6.1%	5.4%

[1]A small percentage of cases in the pilot had a missing performance cluster (less than 0.2%). These were approximately equally distributed between treatment groups

Source: American Community Survey Paradata, August 2015.

We observe a moderate amount of variation by Regional Office in the percentage of cases pulled by treatment group (Table 5.5) and the percentage of cases pulled by performance cluster cross-classified by treatment group (Table 5.6). A large percentage of cases were pulled from Treatment 2, Performance Clusters 2–3 in Atlanta (9.4 percent) and Treatment 3 Performance

Clusters 2–3 in Chicago (8.1 percent). We do not observe in either table a consistent pattern of which treatment group had a larger percentage of cases pulled.

Table 5.5 Percentage of cases pulled among total cases by regional office by treatment group

Regional Office	Treatment 2	Treatment 3
New York	5.2%	3.1%
Philadelphia	6.1%	3.7%
Chicago	3.4%	4.5%
Atlanta	3.9%	3.3%
Denver	3.9%	5.0%
Los Angeles	6.4%	4.3%

Source: American Community Survey Paradata, August 2015.

Table 5.6 Proportion of cases pulled by performance cluster by regional office by treatment

	Treatment 2		Treatment 3	
Regional Office	**Cluster 1**	**Clusters 2–3**	**Cluster 1**	**Clusters 2–3**
NY	4.7%	6.9%	3.1%	3.0%
PHI	4.9%	7.9%	2.6%	5.8%
CHI	3.6%	2.2%	3.7%	8.1%
ATL	3.3%	9.4%	3.2%	3.9%
DEN	3.4%	4.9%	5.1%	4.8%
LA	6.3%	6.4%	3.0%	5.5%

Source: American Community Survey Paradata, August 2015.

5.2 Perceived Contact Burden

Because of the removal of cases, we expect some reduction in measures of burden such as the number of contact attempts and contacts per case recorded in CHI for Treatments 2 and 3. Note that we measured contact burden using the CHI, making it unclear if observed differences were actually due to cases being removed or if FRs were not correctly recording CHI. In sections 5.5.2 and 5.5.3, we investigate CHI reporting.

Table 5.7 shows the recorded contact attempts per case for each treatment during August and July. We estimate the difference between Treatment 1 and Treatments 2–3 to be 0.234 contact attempts per case [90 percent confidence interval (CI) is (0.019, 0.450), p-value = 0.074]. This represents a 6.0 percent decrease in the number of reported contact attempts per case. There is no evidence of a difference in total contact attempts per case between Treatments 2 and 3 [p-value = 0.992].

Table 5.7 Recorded contact attempts per case by treatment and month

	Total cases				Contact attempts per case				
Month	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Trt 2–3**
August (Test Month)	44,911	4,299	4,135	4,213	3.87	3.90	3.67	3.66	3.67
July	45,016	4,229	4,147	4,245	3.84	3.89	3.93	3.92	3.92
Estimated Change* (August – July)					†	0.01	-0.28	-0.26	-0.27

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

We observe a small decrease in the recorded contacts per case among Treatments 2 and 3 compared with Treatment 1 (Table 5.8). The estimated difference between Treatment 1 and Treatments 2–3 was 0.062 contacts per case [90 percent CI (0.006, 0.117), p-value = 0.067]. This represents a 6.2 percent decrease in the contacts per case. We do not find a statistically significant difference in the contacts per case between Treatments 2 and 3 [p-value = 0.886].

Table 5.8 Recorded sample-person contacts per case by treatment and month

	Total Cases				Contacts per case				
Month	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Trt 2–3**
August(Test Month)	44,911	4,299	4,135	4,213	1.02	1.00	0.93	0.94	0.94
July	45,016	4,229	4,147	4,245	1.01	0.98	0.95	1.00	0.97
Estimated Change* (August – July)					†	0.01	-0.01	-0.06	-0.04

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

An important component of perceived burden is contacts with a firm refusal. Although these events occur in a relatively small proportion of cases, they are a good indicator that a respondent feels burdened. Here we define contact with a firm refusal as when the FR records having made contact with a sample unit member and reported one of the following options from the concerns/behaviors screen in CHI: not interested, hang-up/slams door on FR, hostile or threatens FR, or intends to quit survey. Table 5.9 shows that the recorded sample-person contacts with a firm refusal per case decreased by a small amount for Treatments 2–3 compared with Treatment 1. The estimated difference was 0.017 contacts with firm reluctance per case [90 percent CI (0.004, 0.029), p-value = 0.032]. Though small in magnitude, this represents a 19.4 percent reduction in contacts with a firm refusal. Comparing Treatment 2 and Treatment 3, we estimate a non-significant difference of 0.009 contacts per case [p-value = 0.162].

Table 5.9 Recorded sample-person contacts with a firm refusal per case by treatment and month

	Total Cases				Contacts with firm reluctance per case				
Month	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Trt 2–3**
August (Test Month)	44,911	4,299	4,135	4,213	0.09	0.08	0.07	0.06	0.07
July	45,016	4,229	4,147	4,245	0.09	0.09	0.09	0.08	0.08
Estimated Change* (August – July)					†	0.00	-0.01	-0.02	-0.02

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, July and August 2015.

We are interested in whether the stopping rule decreased specific types of contact burden, specifically personal visit attempts (Table 5.10) and telephone attempts (Table 5.11). Table 5.10 displays the recorded personal visit attempts per case by treatment and month, showing an estimated difference in personal visit attempts per case of 0.095 for Treatment 1 versus Treatments 2–3 [90 percent CI (-0.026, 0.217), p-value = 0.198]. This difference was not statistically different from zero. The observed difference between Treatment 2 and Treatment 3 is -0.115 [p-value = 0.197], which was also not statistically different from zero.

Table 5.10 Recorded personal visit attempts per case by treatment and month

	Total Cases				Personal Visit Attempts Per Case				
Month	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Trt 2–3**
August (Test Month)	44,911	4,299	4,135	4,213	2.23	2.15	2.00	2.11	2.06
July	45,016	4,229	4,147	4,245	2.22	2.12	2.11	2.26	2.19
Estimated Change* (August – July)					†	0.04	-0.12	-0.15	-0.14

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, July and August 2015.

We do not find a significant difference in telephone contacts between Treatment 1 and Treatments 2–3 (Table 5.11). The estimated difference was 0.100 telephone attempts per case [90 percent CI (0.000, 0.201), p-value = 0.101]. We estimate the difference between Treatment 2 and Treatment 3 to be 0.106 attempts per case [p-value = 0.045], which is significantly different from zero. We note that the difference of differences test for this same contrast is not significant [p-value = 0.755] (Table 10.2). This is one of the few places where the tests do not agree.

Table 5.11 Recorded telephone attempts per case by treatment and month

	Total Cases				Telephone Attempts Per Case				
Month	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Cntr**	**Trt 1**	**Trt 2**	**Trt 3**	**Trt 2–3**
August (Test Month)	44,911	4,299	4,135	4,213	0.69	0.79	0.75	0.64	0.69
July	45,016	4,229	4,147	4,245	0.68	0.81	0.85	0.72	0.78
Estimated Change* (August – July)					†	-0.02	-0.10	-0.08	-0.09

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, July and August 2015.

The distribution of total burden score by treatment (

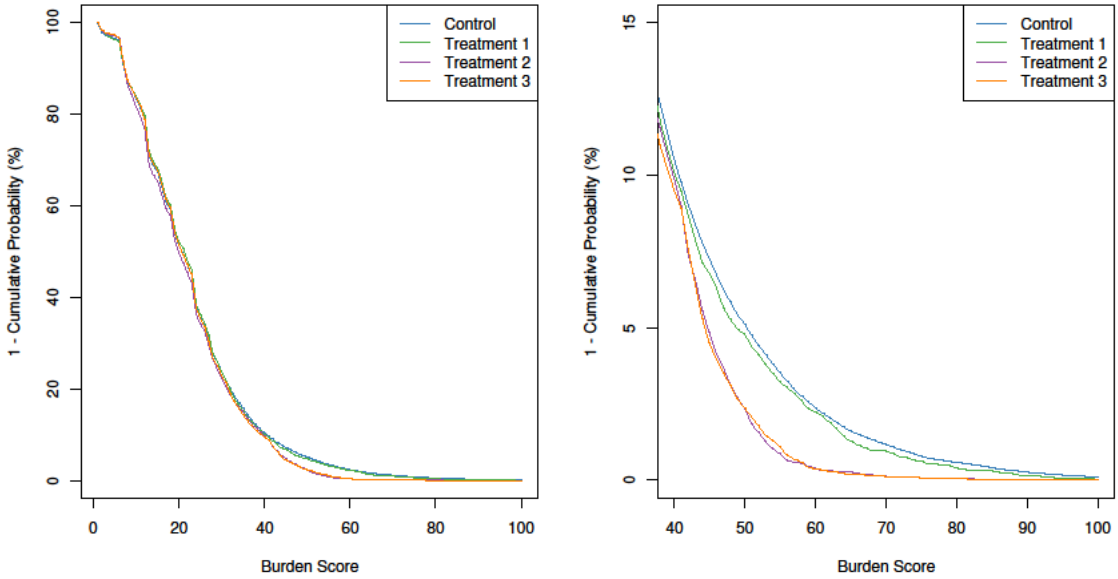


Figure 5.3) shows that the percentage of cases with burden scores above 40 decreased for both Treatments 2 and 3 versus Treatment 1. In

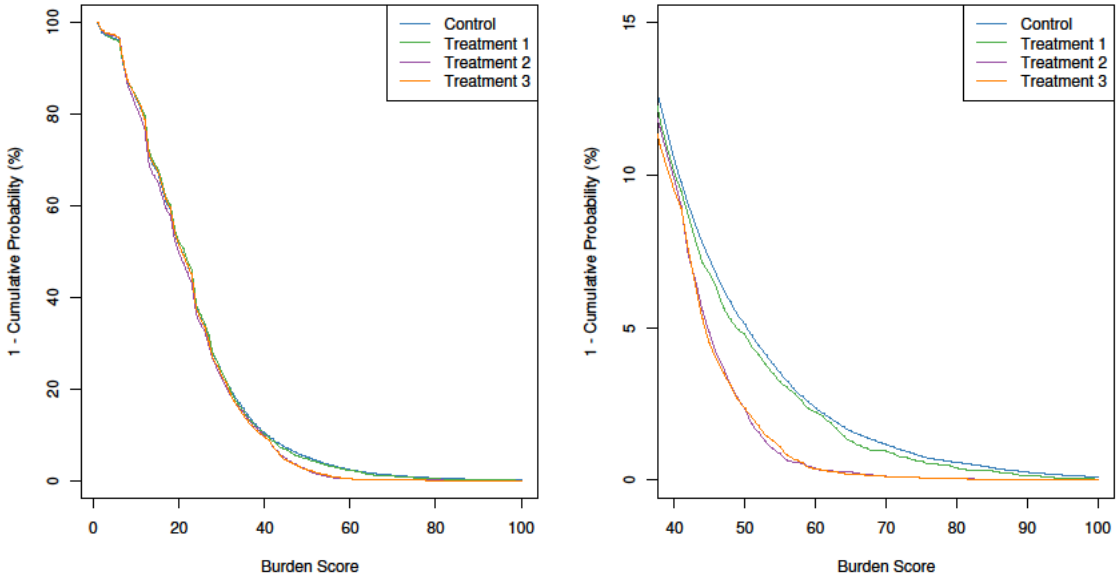


Figure 5.3, the plot on the right hand side is the right tail of the plot on the left. There are similar distributions of cases with burden scores less than 40 for all treatment groups, but Treatments 2 and 3 had only 0.3 percent of cases with burden scores over 60, while control and Treatment 1 had 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. Cases in Treatments 2 and 3 can have burden scores over 40 for a number of reasons:

a) the final burden increment is large (e.g. 38+15 = 53);
b) there are a number of attempts to a unit when it is near the threshold;
c) the case is not pulled because an FR does not transmit; or
d) the case is reassigned to Treatment 1 or Control.

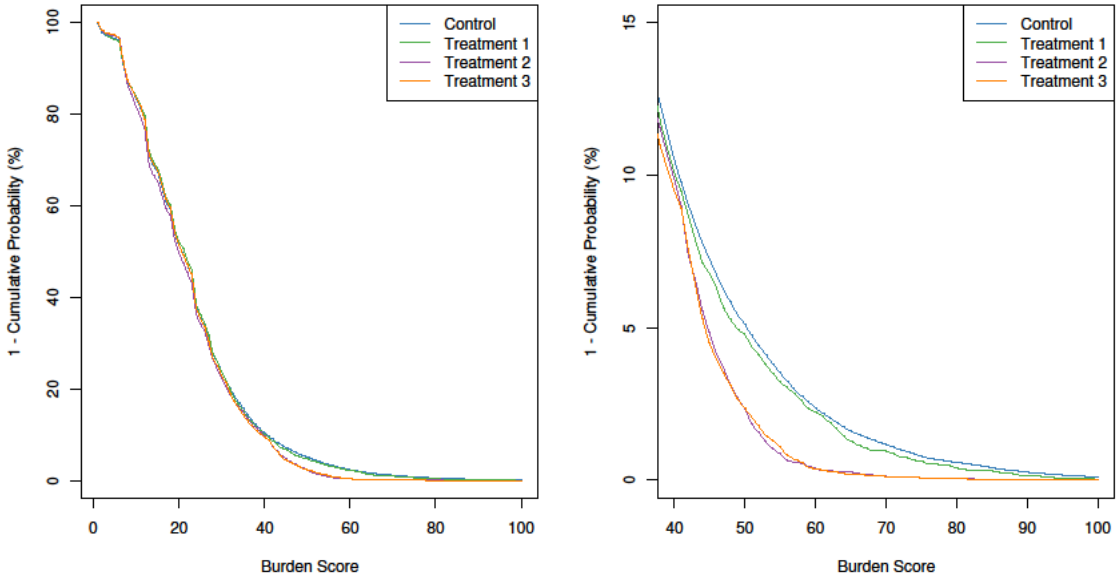


Figure 5.3 Total burden score distribution by treatment, August 2015

Source: American Community Survey Paradata, August 2015.

Table 5.12 shows the percentage of cases reaching a burden score of at least 41, cross-classified by treatment group and initial burden score categories. It is clear that while the initial burden score carried over from mail or CATI contact attempts has a profound effect on the case-rate of occurrence of high cumulative burden scores in CAPI, this effect does not interact with treatment.

Table 5.12 Counts of cases by treatment and initial score, and percentage with burden score >40

Initial Score		Control	Treatment 1	Treatment 2	Treatment 3
0	**Count**	4,406	341	265	268
	% with burden > 40	1.0%	1.2%	0.8%	1.1%
4-6	**Count**	22,527	2,196	2,222	2,237
	% with burden > 40	6.1%	5.4%	4.7%	5.2%
12-17	**Count**	15,947	1,537	1,455	1,501
	% with burden > 40	14.1%	13.2%	13.2%	12.7%
20-25	**Count**	2,031	225	193	207
	% with burden > 40	35.2%	36.0%	38.3%	32.4%

Source: American Community Survey Paradata, August 2015.

5.3 Interviewing Hours and Miles

While it appears that FRs made fewer attempts per case in Treatments 2 and 3 compared with Treatment 1, we do not find any evidence that the interviewing hours per case decreased as well (Table 5.13). The estimated difference between Treatment 1 and Treatments 2–3 is -0.01, which is not significantly different from zero [90 percent CI (-0.099, 0.078), p-value =0.854]. Additionally, we do not find evidence that Treatment 2 is significantly different from Treatment 3 [p-value = 0.353]. Note that for these calculations of hours interviewing per case, all cases (including those determined to be ineligible as ACS sample units) are included.

This finding may be explained in part by the operational impact of the hours per case ceiling that is implemented in ACS data collection operations. Specifically, FRs are given an allocation of hours based on their workload size for all interviewing efforts that they must remain within as a cost controlling measure. These hours per case allocations were not modified during the pilot test to account for cases that were pulled. Therefore, once a case was pulled, the FR could choose to spend more of the time remaining in their allocation on other cases, thereby absorbing any potential reduction in hours interview per case created by pulled cases. Previously released research modeling the effects of the proposed stopping rules on respondent burden, cost, and quality using 2012 paradata estimated a national reduction of 4.4 percent of interviewing hours based on the implementation of a cumulative burden score stopping rule (Griffin, Slud, and Erdman 2015). The findings by Griffin et al. may inform future decisions about appropriate interviewing hours per case allocations during production implementation of these procedures.

Table 5.13 Hours interviewing per case by treatment and month

Month	Control	Trt 1	Trt 2	Trt 3	Trt 2–3
August (Test Month)	1.75	1.70	1.68	1.74	1.71
July	1.76	1.75	1.74	1.86	1.80
Estimated Change* (August – July)	†	-0.05	-0.05	-0.13	-0.09

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

Miles per case is another measure of cost, and, similarly to interviewing hours per case, we do not find any significant differences between Treatment 1 and Treatments 2–3 (Table 5.14). The estimated difference in miles per case between Treatment 1 and Treatments 2–3 was -0.642 [90 percent CI (-2.712, 1.429), p-value = 0.615]. The difference in miles per case between Treatment 2 and Treatment 3 is not statistically significant [p-value = 0.388].

Table 5.14 Miles per case by treatment and month

Month	Control	Trt 1	Trt 2	Trt 3	Trt 2–3
August (Test Month)	21.76	22.78	22.71	24.12	23.42
July	22.33	24.12	24.26	26.04	25.15
Estimated Change* (August – July)	†	-1.52	-1.50	-1.88	-1.69

*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 control FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

5.4 Response Rates

We anticipated that pulling cases would also affect the CAPI response rate.

Table 5.15 shows the case response rate by treatment group and month. 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. See the text around Table 5.42 for additional discussion of the proportion of pulled cases likely to result in complete interviews. We do not find a statistically significant difference in response rate between Treatment 2 and Treatment 3 [p-value = 0.502].

Table 5.15 Case response rate, by treatment and month

	Control	Trt 1	Trt 2	Trt 3	Trt 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 control FSAs appeared in both months.
Source: American Community Survey Paradata, August and July 2015.

Table 5.16 shows the response rate by performance cluster, treatment and month. The table shows modest decreases during the pilot for Treatments 2–3 compared with Treatment 1 for both Performance Cluster 1 and Clusters 2–3, though the difference appears to be larger for Clusters 2–3.

Table 5.16 Response rate by performance cluster, treatment, and month

Month	Performance Cluster	Control	Trt 1	Trt 2	Trt 3	Trt 2–3
August (Test Month)	**1**	94.0%	93.7%	92.9%	92.7%	92.8%
	2–3	92.3%	91.4%	90.1%	88.7%	89.4%
July	**1**	93.9%	94.3%	94.6%	93.9%	94.2%
	2–3	91.9%	91.5%	91.7%	92.0%	91.9%

Source: American Community Survey Paradata, August and July 2015.

The initial burden score is a key factor in the likelihood the case will result in a complete interview. Table 5.17 shows that cases with smaller initial burden scores were more likely to end up as complete interviews. This was true among all treatment groups during the pilot. To see this in terms of statistical tests, note that each of the Treatment 1 and Treatment 2–3 columns of Table 5.17 can be viewed as a 4 x 2 within-group multinomial contingency table. The null hypotheses of row-column independence are rejected by 3 degree of freedom chi-squares for multinomial data at respective p-values of 0.003 for Treatment 1 and < 0.0001 for Treatments 2-3[6].

In Table 5.17, among cases with higher initial burden scores, response rates for Treatments 2–3 were considerably lower than those for Treatment 1. Among cases with high initial burden scores (greater than or equal to 12), the estimated difference in response rate between Treatment 1 and Treatments 2–3 was 3.0 percentage points [90 percent CI (1.1%, 4.9%), p-value =0.009]. This finding makes sense because cases with high initial burden were most likely to be pulled.

Table 5.17 Response rates by initial burden score by treatment group, August 2015

Initial burden score	Control	Trt 1	Trt 2–3	Trt 1 Cases	Trt 2-3 Cases
0	95.6%	94.4%	97.0%	341	533
4 to 6	93.4%	93.4%	93.0%	2196	4459
12 to 20	93.2%	93.2%	90.6%	1537	1956
24 to 25	89.1%	87.1%	80.8%	225	400

Source: American Community Survey Paradata, August 2015.

[6] The increasing trend, within each of Treatment 1 and Treatments 2-3 combined, of the rate of hitting the burden threshold as a function of initial score, can also be documented by logistic regression in terms of an ordered score of 1 to 4 for the respective initial-score categories of 0, 4-6, 12-20, and 24-25. The two-sided p-values for significant (positive) slope are respectively 0.011 for Treatment 1 and < 0.0001 for Treatments 2-3.

Table 5.18 shows the response rate for the pilot for each combination of initial burden category, performance cluster and treatment group. It shows that the largest decrease in response rate for Treatments 2–3 compared with Treatment 1 and control are for units in Performance Clusters 2 and 3 with an initial burden score of 12 to 20 or units with an initial burden score of 24 to 25. See Table 3.5 and Table 3.6 for the distribution of initial burden score and the distribution of initial burden score by performance cluster.

Table 5.18 August, 2015 response rates by initial burden score, cluster and treatment

	Control		**Treatment 1**		**Treatment 2–3**	
	Performance cluster		**Performance cluster**		**Performance cluster**	
Initial burden score	**1**	**2–3**	**1**	**2–3**	**1**	**2–3**
0	95.5%	96.4%	94.4%	†	97.15%	†
4 to 6	94.2%	92.3%	94.1%	91.6%	93.8%	91.2%
12 to 20	93.8%	92.3%	93.4%	92.9%	91.6%	88.2%
24 to 25	89.1%	88.9%	90.6%*	78.1%*	82.8%	74.8%*

†Very small cell size (< 10 total cases), *Small cell size (< 200 total cases)
Source: American Community Survey Paradata, August 2015.

5.5 Field Operations

5.5.1 Transmission Compliance

These research questions relate to individual FRs and to event occurrences for FRs on single ACS workdays. There were 3,147 unique FRs working across all SSFAs in August 2015, with 57,558 distinct cases.

How well do FRs transmit at the start and end of each work day?

The definition of "transmission compliance" requires some explanation. The guideline issued to FRs and supervisors during the pilot was that transmissions should be made after 8:00 AM and before beginning work and making any ACS CHI entries for the day; and again at the end of the work day, before midnight, and after any ACS CHI entries for the day. These guidelines were necessary to ensure that all ACS CHI entries were received in time for the overnight processing of the updated burden scores and to send files to the laptops to pull cases that had exceeded the burden score threshold.

One complication in the definition of transmissions at the start and end of the workday is that there were 144 recorded CHI attempts between 12:00 midnight and 1:00 AM in August 2015, out of a total of 221,014 attempts across all FRs and cases in August 2015. This included 102, 30, 1, and 11 attempts in Control and Treatments 1, 2 and 3 respectively. Similarly, there were 5,383 attempts between 1:00 AM and 5:00 AM in August 2015, of which FRs in Control and

Treatments 1, 2 and 3 respectively accounted for 4348, 381, 307, and 347. Here and throughout this Section, "attempts" refers to all CHI entries.

The definitions we use in the following tables are that "compliant pre-work transmissions" are those made by an FR after 8:00 AM if no CHI entries have been made on that day between 1:00 AM and the transmission; and "compliant post-work transmissions" are those made before midnight and after all CHI entries made the same day.[7] The pre- and post-work and combined pre- and-post-work transmission compliance rates, out of all 38,552 FR-workdays with at least one CHI attempt for the August 2015 pilot, are displayed in Table 5.19. Recall, however, that FRs in the Control were not supposed to follow the twice-daily transmission protocol and were only supposed to transmit once, after their ACS workday. If the analyses in Table 5.19 are restricted to FR-workdays with at least one attempt that increments the burden score on its case, proportions are approximately the same, changing by no more than 0.3 percent in any cell.

Table 5.19 Compliance rates for August 2015 FR-workdays with CHI attempts

	Control	**Treatment 1**	**Treatment 2**	**Treatment 3**
Pre-work Compliance Rate	7.6%	19.9%	23.3%	19.3%
Post-work Compliance Rate	72.9%	83.4%	85.7%	82.0%
Pre- and Post-work Rate	5.2%	16.4%	20.2%	15.4%
FR-workdays	29,880	2,993	2,735	2,944

Source: American Community Survey Paradata, August 2015.

Due to data errors in ROSCO for the first 7 days of August, tentatively ascribed to recording all transmission times as Eastern Daylight Times instead of local times for the FRs, we expect greater (erroneous) degrees of noncompliance in the first 7 days of the August 2015 pilot. To see whether this was actually an issue, we re-created the table for transmission compliance restricting to the days of the pilot after August 7 (Table 5.20). There are some differences of compliance rate percentages in Table 5.19 and Table 5.20. The main difference is that the compliance proportions are slightly higher across the board in Table 5.20, by amounts of the order of 2 percent, but the contrasts between treatment groups are essentially unchanged.

[7] Using 1:00 AM rather than 12:01 AM as the earliest CHI entry time that could render a post-8:00 AM transmission noncompliant turns out to make very little difference to the results, less than 0.1 percent in all table entries.

Table 5.20 Compliance rates for FR-workdays after August 7 with CHI attempts

	Control	Treatment 1	Treatment 2	Treatment 3
Pre-work Compliance Rate	9.0%	23.8%	27.0%	23.6%
Post-work Compliance Rate	76.0%	85.4%	87.0%	84.8%
Pre-and-Post-work Rate	6.5%	20.3%	24.3%	20.0%
FR-workdays	19,191	1,885	1,731	1,854

Source: American Community Survey Paradata, August 2015.

In Treatments 1, 2, and 3, compliance was very poor for the pre-work transmission, 23.8 percent, 27.0 percent, and 23.6 percent respectively (Table 5.20). However, it was much better for the post-work transmission, 85.4 percent, 87.0 percent, and 84.8 percent respectively for Treatments 1, 2, and 3. The overall pre- and post-work compliance rates were 20.3 percent, 24.3 percent, and 20.0 percent for Treatments 1, 2, and 3 respectively. Compliance among the treatment groups was not much greater than for Control who were not instructed to transmit twice a day. Though the compliance was generally poor for all groups, we do find that the combined pre- and post-work compliance rate for Treatment 2 higher than for Treatment 3 by a statistically significant 4.3 percentage points [p-value 0.035, 90 percent CI (0.9%, 7.7%)].

It is not evident from the data what caused the lack of compliance. We do know that during the pilot, an error was identified in the reports that ROs used to monitor transmission compliance, leading to significant overestimates of the percentage of FRs complying with the transmission procedures. Therefore, in those instances where the reports were transmitted to supervisors (FSs), these managers were not able to identify correctly when staff had not complied with the transmission protocol, and to intervene appropriately. Although this error was ultimately remedied during the pilot, we expect that providing the correct reports to managers will likely lead to higher transmission compliance in the future.

What was the effect of poor transmission compliance?

Next, we study the effect that poor transmission compliance had on the objectives of the pilot study. We investigate how transmission compliance affected the calculation of the burden score by UTS, additional attempts and burden for cases reaching the burden threshold, and cases that should have been pulled. By highlighting aspects of excess burden incurred by cases hitting the burden score threshold triggering case removal, we can learn about the extent to which close monitoring of FR transmission compliance matters to the management of burden.

For the analysis, we first ordered attempts by time into an array of (case, day) pairs for the 221,014 attempts recorded in CHI for August 2015. There were 55,578 distinct cases allowing only the days (0 to 34, ranging from July 31 to September 3) for each case on which there were actual attempts. There were a total of 168,861 (case, day) pairs. In addition, the FR transmissions file contains all the time-stamped transmissions of FRs. As mentioned above, the transmission time-stamps from July 31 to August 7 inclusively are in doubt because of ROSCO errors

occurring during that time period related to the time zone. The transmissions file was ordered by FR and day, and transmissions between 12 midnight and 1:00 AM were given modified time-stamps of 1 second before midnight of the previous day. For each case, only the earliest and latest intra-day transmissions by each FR were relevant to overnight processing of the accumulated burden scores and to re-transmission to the FR at the earliest transmission after the next 7:00 AM. Therefore, all FR transmissions other than the earliest and latest in each day were dropped. Using this FR-level information, we calculated for each (case, day) combination the following variables:

BurdenSent = cumulative burden increments transmitted by all FRs by the end of the day.

BurdenReceived = cumulative UTS burden across all FRs received up to the present day in a transmission following an overnight previous transmission to UTS.

Burden = cumulative burden from all attempts by FRs on the case through the end of the day.

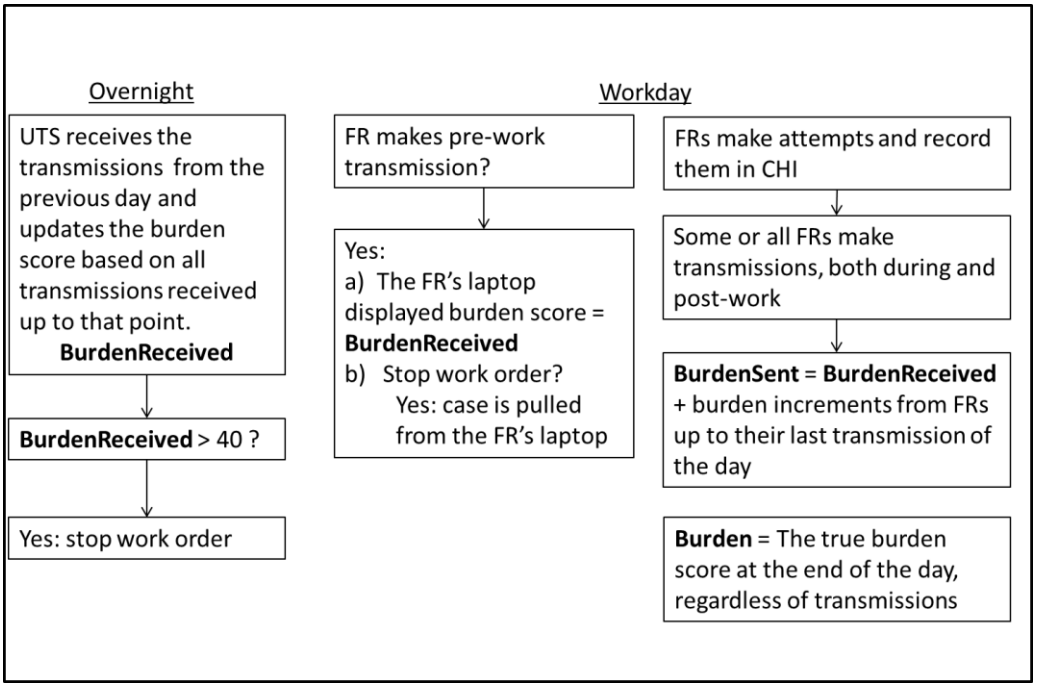


Figure 5.4 Transmission diagram

Source: American Community Survey Paradata, August 2015.

According to these definitions, *BurdenReceived* is no larger than *BurdenSent*, which is less than or equal to *Burden*. *BurdenSent* is the burden score that would be displayed on FRs' laptops if they perform a compliant pre-work transmission the next day, while *Burden* is the true accumulated burden on the case (combining increments from all FRs working the case) at the end of the day. The difference between these burden measures is due to the gaps between transmissions. For example, all three will be equal for a case if an FR has transmitted at the end

of the previous day's work and again before the beginning of the current day's work and has made no attempts in the current day incrementing the burden score. However, if multiple contact attempts are made on the case within the same day, an omission of a post-work transmission on the same or previous day can cause the gaps between *Burden* and *SentBurden* to be large. This often happens near the final attempts on a high-burden case, whether that case is pulled or not.

One can study the gaps between these three burden variables either on typical days or on specific days related to the pilot study, such as the day when *Burden* first exceeds 40. We are most interested in the latter. There are 5,537 cases that attained a cumulative burden of 41 or more in August 2015, out of the total of 57,558 unique cases. The averages of the three burden measures for these cases, subdivided by assigned treatment group, are given in Table 5.21. We find that there were large but remarkably similar gaps between the measures for all treatment groups.

Table 5.21 Average cumulative burden scores and scores transmitted and received by the end of the day on which Burden first exceeded 40.

	Control	**Treatment 1**	**Treatment 2**	**Treatment 3**
Burden	44.91	44.90	44.92	44.59
BurdenSent	29.77	29.98	29.66	30.13
BurdenReceived	18.04	17.76	17.55	18.57
Total Cases	4,382	407	372	376

Source: American Community Survey Paradata, August 2015.

Like the means, the standard deviations of the three variables are very different across type of burden measure but rather constant across group, as shown in Table 5.22.

Table 5.22 Standard deviations of burden scores and scores transmitted and received by the end of the day on which Burden first exceeded 40.

	Control	**Treatment 1**	**Treatment 2**	**Treatment 3**
Burden	3.8	3.7	3.9	3.7
BurdenSent	10.6	9.3	9.3	9.4
BurdenReceived	8.3	8.4	8.1	7.9

Source: American Community Survey Paradata, August 2015.

Table 5.21 and Table 5.22 demonstrate that the gaps in burden quantities monitored by UTS due to lags between sent and received burden scores are large near the burden threshold where cases might be pulled, but that these differences do not vary much by treatment. The tables estimate the magnitude of the excess due to noncompliant transmissions of the actual current burden over that seen in transmissions to and from UTS, at levels of burden up to (but not yet exceeding) the threshold at which cases are removed in the pilot.

To address possible differences in FR behavior across groups, we measure the average number of additional attempts on a case, by treatment, on a case ***after*** its burden reaches 41. The answer is given, in terms of average extra attempts per case and extra burden per case, in Table 5.23.

Both the average extra attempts and average extra burden for Treatments 2 and 3 combined are significantly different from those for Treatment 1. The estimated average difference between Treatments 2–3 and Treatment 1 for extra attempts is 5.87 [p-value < 0.001, 90 percent CI (4.21, 7.53). The estimated average difference for extra burden is 1.27 attempts [p-value < 0.001, 90 percent CI (0.91, 1.63)]. However, there is little or no distinction between Treatments 2 and 3. While it cannot yet be confirmed these contrasts are due solely to the removal of pulled cases, that is our conjecture, since the proportion of cases in which burden increments continue beyond the day when burden reaches 41 is not large. (See discussion following Table 5.23 below.)

Table 5.23 Average number and burden of attempts made after cumulative burden reaches 41 in the August 2015 Pilot, by treatment group.

Treatment Group	**Control**	**Treatment 1**	**Treatment 2**	**Treatment 3**
Average Extra Attempts	2.2	1.8	0.5	0.6
Average Extra Burden	9.2	8.1	2.0	2.4

Source: American Community Survey Paradata, August 2015.

The *BurdenSent* and *BurdenReceived* differences from *Burden* both refer to delays from past transmissions, while the extra attempts and burden – at least in Treatment groups 2 and 3 – relate to the days and additional burden until the next transmission following the crossing of the burden threshold. These time and burden gaps, on average by treatment group, are given in Table 5.24. Table 5.23 andTable 5.24 indicate that in the case-removal environment of Treatments 2 and 3, work continues after the ***attempt*** first exceeding the cumulative burden threshold of 40 for an average of about half an additional attempt and one-quarter day of work.

Table 5.24 Average time, in days, until next transmission after the attempt at which burden first exceeds 41, by treatment group.

	Control	**Treatment 1**	**Treatment 2**	**Treatment 3**
Gap Until Next Transmission	0.38	0.29	0.28	0.28
Gap Since Last Transmission	0.79	0.559	0.48	0.60

Source: American Community Survey Paradata, August 2015.

Next, we investigate how the gaps in burden incurred and the cumulative burdens transmitted to UTS and received from UTS affected whether cases "should have been pulled." There were 748 cases in Treatments 2 or 3 attaining cumulative burden of 41 or more, of which 84 had an additional burden increment on a day subsequent to hitting 41, and of those 84 only 37 were pulled. This represents a surprisingly small rate (11.2 percent = 84/748) of above-40 burden cases that "should have been pulled" but were not, indicating that transmission failures may not have been too serious in compromising the effectiveness of case removal. A further figure of interest is that the median and average number of burden points that the 84 cases generated subsequent to the end of the day on which their cumulative burden reached 41 were respectively 7.5 and 10.0.

A final note on the summarization of burden by the methods described above relates to the "assigned FR" concept. One can ask: is the FR whose attempt first reaches cumulative burden score 41 the same as the assigned FR? The answer is generally yes, but in the 5,537 cases that do attain a burden score of 41, there are 1,229 cases (22.2 percent) in which the assigned FR is not the same as the FR whose attempt brings burden up to 41. This evidence of the frequent reassignment of cases to FRs spreads evenly across treatment groups. The attempt's FR differs from the assigned FR in 22.6 percent of such cases in the Control group, 20.1 percent in Treatment 1, 22.8 percent in Treatment 2, and 18.9 percent in Treatment 3.

How well does case removal work once a stop work order is given?

We found no evidence that FRs were able to continue work on a case once it was pulled, although it often happened that FRs, through multiple attempts on the final day of a case, incremented the burden score by 15 or more. For the 380 pulled cases, we define ***Lag*** as the last attempt-day in CHI entries for the case minus the UTS termination date. This quantity will generally be a small negative integer, but not always. A cross-tabulation, by treatment, is given in Table 5.25. Negative ***Lag*** values imply that the last attempt-day was before the UTS termination date, and positive ***Lag*** values imply that the last attempt-day was after the UTS termination date. There is no discernible difference in ***Lag*** pattern between Treatments 2 and 3, but pulled cases generally have the last day of FR attempts occur one day earlier than the termination date defined by UTS, which is the date of the transmission from UTS when the case is actually pulled from the FR's laptop.

Table 5.25 Frequencies of lags in days by case between UTS termination date and last attempt by treatment.

Lag	Control	Treatment 1	Treatment 2	Treatment 3
< -1	0	2	13	11
-1	3	14	148	142
0	0	2	17	8
1	0	0	3	3
2	0	1	6	7

Source: American Community Survey Paradata, August 2015.

Of the 203 pulled cases that also have a final UTS-calculated burden score defined at the termination date, the burden computed at the last attempt-record on the case is always at least as large. However, in all but a very few cases, the difference is between 0 and 3, but was as large as 39. The final burden in CHI minus the UTS burden score at termination, for the 203 pulled cases

where the latter is defined,[8] occurred with frequencies shown in Table 5.26. The table shows that only 13.3 percent of pulled cases accrued any excess burden because of delayed transmissions.

Table 5.26 Frequency of differences between final burden and UTS-calculated burden at termination.

Difference	**0**	**1**	**2**	**3**	**4+**
Frequency	176	3	5	4	15

Source: American Community Survey Paradata, August 2015.

5.5.2 FR Behavior Differences

Are there systematic group differences in "not attempting contact" behavior?

CHI attempts for which no contact is attempted may indicate activities related to locating the sampled unit, geocoding for ACS cases, correcting previous entries, or closing out a case. Table 5.27 displays the percent of such records in CHI as a fraction of the total, cross-tabulated by treatment group for records in August 2015. For this measure of FR behavior, there is no difference between any of the treatment groups. For instance, comparing Treatment 2–3 with Treatment 1 shows no evidence of a difference in the percentage of "Not attempting contact" attempts [p-value for difference = 0.720].

Table 5.27 Percentage of "Not attempting contact" attempts by treatment.

	Control	**Treatment 1**	**Treatment 2**	**Treatment 3**
Not Attempting Contact	24.1%	24.1%	24.7%	24.5%
New Contact	75.9%	75.9%	75.3%	75.5%
Attempts	173,645	16,751	15,181	15,428

Source: American Community Survey Paradata, August 2015.

What other group differences in FR behavior existed in August 2015?

In response to the removal of cases above the burden score threshold in Treatments 2 and 3, other aspects of FR behavior might have changed. FRs might have changed their pattern of targeting or recording firm refusals, or other respondent concerns, or multiple instances of refusals within single cases. The tables below seek to investigate these possible changes in actual or recorded behavior.

First, Table 5.28 shows the number and proportion of cases in August 2015 with at least one firm refusal cross-classified by treatment group. While Treatments 1, 2 and 3 look as though they

[8] Since UTS-defined burden scores at termination are defined primarily in terms of current FR-display burden scores that are calculated only in Treatment group 2, the great majority of (i.e., 183 of the 203) pulled cases with defined UTS burden score at termination values are in Treatment 2, with only 3 in Treatment 3.

have progressively lower rates of cases with at least one firm refusal, the differences are not significant. The observed difference between Treatment 1 versus Treatments 2 and 3 combined was 0.9 percentage points [p-value = 0.152, 90 percent CI (-0.1%, 2.0%)]. This difference was not statistically different from zero, and, even if it were, we could not distinguish which part of the difference is due to a change in behavior rather than to case removal. The percent with a firm refusal was not significantly different between Treatments 2 and 3 [p-value = 0.269, 90 percent CI for difference (-0.3%, 1.5%)].

Table 5.28 The percent of cases with at least one firm refusal, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
Cases with Refusal	3,086	304	267	246
Percent with Refusal	6.9%	7.1%	6.5%	5.8%
Total Cases	44,911	4,299	4,135	4,213

Source: American Community Survey Paradata, August 2015.

Another approach is to tabulate the proportion of cases by treatment group that exhibit at least two instances of strong reluctance in the form of either firm refusals or interim indication of respondent refusal (Table 5.29). Here the difference between the rate for Treatment 1 and Treatments 2 and 3 combined was 0.8 percentage points, significantly different from zero [p-value = 0.027, 90 percent CI (0.2%, 1.4%)]. There was not a significant difference between rates for Treatment 2 and Treatment 3 [p-value = 0.276, 90 percent CI for the difference (-0.2%, 0.9%)].

Table 5.29 The percent of cases with multiple instances of firm refusals or interim respondent refusal, by treatment.

	Control	Treatment 1	Treatment 2	Treatment 3
Percent with Multiple Firm or Interim Refusals	3.1%	2.7%	2.1%	1.7%
Total Cases	44,911	4,299	4,135	4,213

Source: American Community Survey Paradata, August 2015.

Here the rate of occurrence also tends to be successively smaller from Control to Treatment 1 to Treatment 2 to Treatment 3. The results here are correlated with those of the previous table and cannot - for that reason - be viewed as independent evidence of a trend, but the results suggest a difference between Treatments 2 and 3 behavior. However, the greater (and significant) contrast between Treatment 1 versus Treatments 2–3 accords with the idea that the reduction in multiple refusals is due to the removal of cases before the later refusals could occur.

A similar tabulation of the mean number of attempts with CHI records of “other concerns” shows no difference across groups (Table 5.31) with 28.0 percent in the Control group, 29.4 percent in Treatment 1, 27.5 percent in Treatment 2, and 28.0 percent in Treatment 3. There is no significant difference between Treatment 1 and Treatments 2–3 [p-value for difference = 0.393].

Table 5.30 The percent of cases with at least one "other" concern, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
Percent with at least one "other" concern	28.0%	29.4%	27.5%	28.0%
Total Cases	44,911	4,299	4,135	4,213

Source: American Community Survey Paradata, August 2015.

How did FR behavior change, by treatment, with respect to "observed household from vehicle," type and time of day of attempt, strategy, etc.?

In the following tables, each of these types of attempts is cross-classified by treatment group in August 2015. In Table 5.32 through Table 5.34, the universe is restricted to the 166,555 attempts in which contact was attempted. (See the discussion of this category preceding Table 5.27 above.) In Table 5.31, the universe is restricted to only attempts for which personal contacts were attempted. Throughout these tables, the treatment is not "assigned" uniquely to the case, but is regarded as an attribute of the attempt made by a specific FR.

Overall, Table 5.31 shows small percentages of attempts for which the FR noted "observing the household from vehicle" as the reason for noncontact. The percentages were 1.5, 1.5, 2.3, and 1.0 percent respectively for Control, Treatment 1, Treatment 2, and Treatment 3. The difference between Treatment 2 and Treatment 3, 1.3 percentage points, was borderline statistically significant [p-value = 0.101, 90 percent CI = (0.0%, 2.6%)].

Table 5.31 Percentage of personal visit attempts for which the FR categorized the attempt as "observing the household from vehicle", by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
All other strategies	98.5%	98.5%	97.7%	99.0%
Observing household from vehicle	1.5%	1.5%	2.3%	1.0%
Attempts	99,920	9,253	8,294	8,890

Source: American Community Survey Paradata, August 2015.

In Table 5.32, the difference in the percent of telephone attempts between Treatments 2–3 and Treatment 1 was not statistically significant [p-value =0.390 , 90 percent CI (-1.5%, 4.8%)]. However, the difference between Treatment 2 and Treatment 3 was significantly different from zero [p-value = 0.041, 90 percent CI (0.7%, 6.7%)]. We do note the discrepancy between this finding and that of the test of difference of differences in Table 10.2.

The next three tables display the proportions of attempts, characterized by telephone versus personal visit, time of day, and weekend versus weekday. (In Table 5.33, AM denotes an attempt (telephone or personal visit) made before 12noon; PM1 an attempt between 12noon and 3pm; PM2 an attempt between 3pm and 6pm; and NIGHT an attempt made after 6pm.) In all three tables, there are no discernable differences among the treatment-group percentages, although

formal significance tests were not done. The tentative conclusion is that FR contact strategy was not seriously affected by the pilot study interventions.

Table 5.32 Percentage of telephone versus personal visit contacts, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
Telephone	23.7%	26.8%	27.0%	23.3%
Personal Visit	76.3%	73.2%	73.0%	76.7%
Attempts	130,936	12,647	11,368	11,604

Source: American Community Survey Paradata, August 2015.

Table 5.33 Percentage of personal visit attempts by time of day, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
AM	16.2%	16.5%	17.4%	15.2%
PM1	30.4%	29.7%	30.7%	31.9%
PM2	43.1%	42.1%	40.5%	41.0%
NIGHT	10.4%	11.7%	11.4%	11.8%
Attempts	130,936	12,647	11,368	11,604

Source: American Community Survey Paradata, August 2015.

Table 5.34 Percentage of weekend vs. weekday visit attempts, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
Weekend	29.9%	29.4%	29.5%	30.2%
Weekday	70.1%	70.6%	70.5%	69.8%
Attempts	130,936	12,647	11,368	11,604

Source: American Community Survey Paradata, August 2015.

The FR may indicate on a "strategy" screen in CHI the approach being followed in attempted contacts. From this screen, we defined FR selections of the strategies "left advance-letter," "note/appointment card," or "packet/brochure" as "low burden;" the strategies "scheduled appointment," "called household," or "left phone-message," as "medium burden;" and the strategies "waited for respondent," "checked with neighbors," or "contacted other family" as "high burden." Table 5.35 shows the percentage of each among the 108,369 total attempts for which a "strategy" was recorded. We find that the difference in the percentage of low burden attempts for Treatment 1 minus Treatments 2–3 was 3.8 percentage points, not significantly different than zero [p-value = 0.157, 90 percent CI (-0.6%, -8.2%)]. Additionally, the difference between Treatments 2 and 3 was not statistically significant [p-value = 0.110, 90 percent CI (-10.9%, 0.2%)]. Lastly, the difference in the percentage of high burden attempts between Treatment 1 and Treatments 2–3 was -2.5 percent, was also not significantly different from 0 [p-value = 0.129].

Table 5.35 Percentage of personal visit attempts by strategy and treatment

Strategy	Control	Treatment 1	Treatment 2	Treatment 3
High Burden	21.1%	17.6%	20.4%	20.0%
Medium Burden	27.7%	30.3%	33.9%	29.0%
Low Burden	51.2%	52.1%	45.7%	51.1%
Attempts	85,751	7,886	7,473	7,259

Source: American Community Survey Paradata, August 2015.

It also makes sense to ask whether strategies are selected differently across treatment groups when the universe of case attempts is restricted to those with previous cumulative burden of 30 or more, a high burden score but not high enough to trigger case removal in the pilot. Table 5.36 shows the strategy proportions restricting to cases with high cumulative burden. Though Treatment 2 appears to have had a smaller percentage (33.5 percent) of low burden attempts compared with Treatment 3 (37.8 percent), this difference is not statistically significant [p-value = 0.248].

Table 5.36 Percentage of visit attempts by strategy and treatment among cases with previous cumulative burden of 30 or more

Strategy	Control	Treatment 1	Treatment 2	Treatment 3
High Burden	13.9%	10.2%	10.0%	10.7%
Medium Burden	47.1%	52.1%	56.5%	51.5%
Low Burden	39.0%	37.7%	33.5%	37.8%
Attempts	11,714	998	720	691

Source: American Community Survey Paradata, August 2015.

None of these tables supplies convincing evidence that FR behavior or strategy is changing systematically as a result of the pilot study interventions (twice-daily transmissions and case removals).

5.5.3 Payroll/CHI Inconsistencies

As noted in Section 3.1, FRs submitted daily work records to payroll through the WebFRED system, which is later displayed in the CARMN system. One way to assess FR reporting accuracy is to compare the days and categories of reported interviewing payroll hours with the corresponding entries under the CHI system documenting their contact attempts and results. Among FR-workday records, 34,674 FR-workdays are associated with the August 2015 interview period. Of these, none showed CHI entries without corresponding payroll data, and of the workdays with interviewing hours, 27,769 showed both CHI and payroll entries ("matching" FR-days), while 6,128 showed payroll but no CHI entries of any kind.

Table 5.37 shows the breakdown of FR-workdays according to whether interviewing hours were submitted to payroll and, if so, whether there were CHI entries. (Those without CHI entries cannot be classified by treatment.) The FR-workdays with no interviewing hours recorded are

fairly balanced among the three treatment groups and are more numerous among the treatment groups than among the control because these are essentially all, 570 out of 575, days in which FR transmission activity associated with the pilot were the only activity submitted to payroll.

Table 5.37 FR-workdays with matching (both CHI and payroll) and payroll-only entries with interviewing hours, and FR-workdays with no interviewing hours, by treatment group

		No CHI	Control	Trt 1	Trt 2	Trt 3
With Interviewing Hours	Matching FR-workdays	0	21,284	2,235	2,001	2,249
	Payroll-only FR-workdays	631	4,209	421	450	417
With no Interviewing Hours	All FR-workdays	20	30	214	264	249

Source: American Community Survey Paradata, August 2015.

Table 5.38 shows that 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%)].

Table 5.38 Counts and rates of payroll-only FR-days with interviewing hours, by treatment

	Control	Treatment 1	Treatment 2	Treatment 3
Matching	83.5%	84.1%	81.6%	84.4%
Payroll-only	16.5%	15.9%	18.4%	15.6%
FR-days	25,493	2,656	2,451	2,666

Source: American Community Survey Paradata, August 2015.

The analogous table restricted to FR-workdays in which both interviewing hours and miles were reported to payroll is Table 5.39. Here the payroll-only rates are much smaller and the possible group differences much less clear.

Table 5.39 Counts and percentages of payroll-only FR-workdays with interview hours & miles

	Control	Treatment 1	Treatment 2	Treatment 3
Matching	96.7%	98.1%	96.5%	97.9%
Payroll-only	3.3%	1.9%	3.5%	2.1%
FR-days	17,339	1,730	1,610	1,789

Source: American Community Survey Paradata, August 2015.

Finally, we display in Table 5.40 the proportion of FR-workdays in which both CHI and payroll activities are reported with interview hours (and on all such FR-workdays, miles were also reported to payroll) where there was at least one personal visit (PV) attempt reported in CHI. Although FR behavior with respect to personal visits might have varied across treatment groups, the table shows equal rates of PV versus non-PV attempts across all treatment groups.

Table 5.40 Counts of FR-days with payroll and CHI and interviewing hours and percent of such FR-days with Personal Visit (PV) CHI attempts

	Control	Treatment 1	Treatment 2	Treatment 3
No PV Attempts	19.7%	22.1%	22.1%	21.0%
PV Attempts	80.3%	77.9%	77.9%	79.0%
FR-days	21,284	2,235	2,001	2,249

Source: American Community Survey Paradata, August 2015.

5.5.4 Distribution of Final Case Status

Another way to see the effects of the pilot interventions in Treatments 2 and 3 with case removals is to compare the distribution of final outcome codes by group. "Completed" cases are the eligible cases with final outcome codes indicating interviews (i.e. complete, partial, temporarily-occupied or vacant-unit interviews) were obtained. "Noninterview" cases are the eligible cases with non-interview outcomes. Late mail returns (LMRs) are cases removed from workload, and "Pulled" cases are noninterviews removed in the pilot due to a high burden score. For context, the breakdown of all eligible[9] August cases by final outcome within the treatment groups is given in Table 5.41, with final column indicating the overall percentage of total cases with the final outcome status of the listed types. For comparison, the breakdown of outcomes by treatment group for the cases whose burden score was 41 or more is displayed in Table 5.42. Note that the response rate as calculated in Section 5.2, is the percentage of cases that are either in the "Completed" or "LMR" categories below.

Table 5.41 The percentage of final outcome status by treatment group among eligible August cases

	Control	Trt 1	Trt 2	Trt 3	Total cases	Overall percentage
Completed	86.1%	86.8%	83.6%	82.9%	49,337	85.7%
LMR	7.3%	6.3%	8.5%	8.6%	4,254	7.4%
Pulled	0.0%	0.4%	4.5%	4.1%	381	0.7%
Noninterviews	6.6%	6.5%	3.4%	4.4%	3,586	6.2%
Cases	44,911	4,299	4,135	4,213	57,558	

Source: American Community Survey Paradata, August 2015.

[9] While this report has dealt almost exclusively with eligible cases, we also investigate whether the burden reduction rules had an impact on classifying cases as ineligible. It is possible that some cases that were pulled and subsequently classified as eligible non-interviews, would have been classified as ineligible if not pulled. This would result in a reduction in the percentage of cases found to be ineligible among Treatment 2 and 3. We found no significant difference in the percentage of ineligible cases between Treatment 1 and Treatment 2–3 [p-value= 0.800]. The difference between Treatment 2 and 3 is borderline significant [p-value = 0.103]; however, the comparison with the percentage in July suggests that percentage of ineligible cases in Treatment 3 increased under the burden reduction rules (rather than decreased as expected). Further research should be considered on this topic.

The removed cases due to the pilot study intervention are the "pulled" cases in Treatments 2 and 3, and it seems logical that these subtract from the Completed and Noninterview categories. The majority of non-interviews are "respondent refusals"; other noninterviews include cases categorized as "unable to locate", which were actually more frequent in Treatments 2–3 than in Treatment 1. The percentage of cases with a LMR outcome status in Treatments 2–3 is significantly different than in Treatment 1 [p-value = 0.025, 90 percent CI (-4.0%, -0.6%)]. However, the conclusion of this test disagrees with that of the difference of differences test (Table 10.1). The reason for the possible increase in LMRs in Treatments 2 and 3 over Treatment 1 is likely a result of the policy that potential respondents who had their cases pulled due to surpassing the burden score but initiated further contact were instructed to complete the interview on the telephone or themselves, by mail or online.

In Table 5.42, note that in Control and Treatment 1, more than half of the cases that accumulate burden > 40 end as completed cases. While this appears to disagree with our previous observation from the discussion preceding Table 5.15 that most pulled cases would not have ended up as interviews had they not been pulled, it however does not. The discrepancy is because cases that are pulled from the workload due to surpassing the burden score threshold have the opportunity to be converted to late mail returns. The statement that "most pulled cases would not have ended up as interviews had they not been pulled" applies to cases that are pulled and end as non-interviews.

Table 5.42 Cross-tabulation of percentage of case final outcomes within treatment groups for cases that eventually accumulate burden > 40

	Control	Treatment 1	Treatment 2	Treatment 3	Total cases	Fraction
Completed	56.3%	56.8%	37.9%	41.8%	2,998	54.1%
LMR	8.6%	4.7%	7.3%	7.4%	450	8.1%
Pulled	0.1%	4.7%	50.3%	45.5%	380	6.9%
Noninterviews	35.0%	33.9%	4.6%	5.3%	1,709	30.9%
Cases	4,382	407	372	376	5,337	

Source: American Community Survey Paradata, August 2015.

One may conjecture that certain types of case outcomes and FR behaviors would be more understandable if cases are cross-classified by initial burden score or by occupied-versus-vacant status. To examine this, the proportion of cases reaching a burden score of 31, cross-classified by treatment-group and vacant status, is given in Table 5.43. In this table only, the threshold 30 is used in place of 40, because so few vacant cases accumulate high burden, less than 1 percent reaching burden score of 40 in each treatment group.

Table 5.43 Percentage and counts of cases with burden score >30 by treatment and vacancy status

	Control	Treatment 1	Treatment 2	Treatment 3
Total Occupied Cases	32,778	3,012	2,750	2,935
% with burden score > 30	29.1%	30.1%	29.8%	29.6%
Total Vacant Cases	12,133	1,287	1,385	1,278
% with burden score > 30	3.2%	3.4%	2.5%	2.9%

Source: American Community Survey Paradata, August 2015.

Very few cases classified as vacant housing units receive burden scores over 30. This may be because FRs do not code likely vacants in CHI with strong burden score increments or because the FR activities in likely vacant cases genuinely do not impose burden on respondents. However, other CHI entries for these cases, not included in the burden score, may better reflect burdens that FRs impose on the public in handling vacant cases.

5.5.5 FR Hours Worked

FRs may be concerned that the implementation of the burden score will cause their hours to decrease if cases are removed that reach the burden threshold. Table 5.44 compares the distributions of hours worked by FRs in August and July by treatment group. While there was a nominal decrease in hours worked from July to August for FRs in both Treatment 2 and 3, FRs in Treatment 1 and Control also experienced a small decreases in hours.

Table 5.44 Distribution of hours worked by FRs in August & July 2015 by treatment and month

Treatment Group	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Control, August	0.25	21.50	33.25	35.95	47.75	155.00
Control, July	0.25	21.25	34.25	36.30	48.75	151.00
Trt 1, August	0.25	20.75	31.25	32.87	44.06	84.00
Trt 1, July	0.25	19.75	33.12	34.28	47.56	100.00
Trt 2, August	0.25	19.94	31.12	33.30	45.61	106.50
Trt 2, July	0.25	18.50	33.00	33.72	47.25	95.25
Trt 3, August	0.25	22.75	32.75	34.73	44.44	118.00
Trt 3, July	0.50	24.56	35.50	36.85	46.88	111.00

Source: American Community Survey Paradata, August 2015.

Restricting to only FRs who worked in both July and August, Table 5.45 shows the median and mean change in hours from August and July 2015 by treatment group. It shows that FRs in all treatment groups experienced some decline in hours worked from July to August. Applying Wilcoxon rank-sum tests, we find no significant differences between Treatment 1 and Treatment 2–3 [p-value = 0.8347] or Treatment 2 and Treatment 3 [p-value = 0.1743].

Table 5.45 Change in monthly hours worked per FR, July to Aug. 2015, by Treatment, for FRs working both months

	Control	Treatment 1	Treatment 2	Treatment 3
Median change in hours worked	-0.50	-1.50	-0.13	-1.25
Mean change in hours worked	-0.35	-1.42	-0.46	-2.12

Source: American Community Survey Paradata, August 2015.

5.6 Discussion of Significance Testing

Table 5.46 gives a summary of all the hypothesis tests between Treatment 1 and Treatments 2 and 3 combined in the above sections. Positive observed differences imply the value for Treatments 2 and 3 is lower than the value for Treatment 1. We find several variable contrasts to be statistically significant at the 0.10 level and several near significance, though we note that these p-values were not adjusted for multiple comparisons. Because of the large number of tests conducted, the results should be considered with additional scrutiny. For a summary of results of Treatment 1 compared to Treatments 2–3 using the difference of differences statistic, see Table 10.1 in the appendix.

Table 5.46 Summary of hypothesis tests and confidence intervals for the difference between Treatment 1and Treatments 2–3

	Table Number	Observed Difference	p-value	90% CI	95% CI
Total attempts per case	Table 5.7	0.234	0.074	(0.019, 0.450)	(-0.023, 0.491)
Contacts per case	Table 5.8	0.062	0.067	(0.006, 0.117)	(-0.004, 0.128)
Firm refusals per case	Table 5.9	0.017	0.032	(0.004, 0.029)	(0.001, 0.032)
Personal visit attempts per case	Table 5.10	0.095	0.198	(-0.026, 0.217)	(-0.049, 0.240)
Telephone attempts per case	Table 5.11	0.100	0.101	(0.000, 0.201)	(-0.019, 0.219)
Interviewing hours per case	Table 5.13	-0.010	0.854	(-0.099, 0.078)	(-0.115, 0.095)
Miles per case	Table 5.14	-0.642	0.615	(-2.712, 1.429)	(-3.098, 1.815)
Response rate	Table 5.15	1.3%	0.104	(0.0%, 2.6%)	(-0.3%, 2.8%)

	Table Number	Observed Difference	p-value	90% CI	95% CI
Response rate among cases with high initial burden	Table 5.17	3.0%	0.009	(1.1%, 4.9%)	(0.7%, 5.3%)
AM&PM transmission compliance rate	Table 5.20	-1.7%	0.286	(-4.3%, 0.9%)	(-4.8%, 1.4%)
Average extra attempts	Table 5.23	5.87	<0.001	(4.21, 7.53)	(3.91, 7.82)
Average extra burden	Table 5.23	1.27	<0.001	(0.91, 1.63)	(0.85,1.70)
Percent of attempts "not attempting contact"	Table 5.27	-0.5%	0.720	(-3.0%, 2.0%)	(-3.5%, 2.4%)
Percent of cases with at least one firm refusal	Table 5.28	0.9%	0.152	(-0.1%, 2.0%)	(-0.3%, 0.2%)
Percent of cases with multiple firm refusals	Table 5.29	0.8%	0.027	(0.2%, 1.4%)	(0.1%, 1.5%)
Percent of cases with at least one "other" concern	Table 5.30	1.6%	0.393	(-1.4% 4.6%)	(-2.0%, 5.2%)
Percent of attempts that were by telephone	Table 5.32	1.7%	0.390	(-1.5%, 4.8%)	(-2.1%, 5.4%)
Percent of low-burden strategy attempts	Table 5.35	3.8%	0.157	(-0.6%, 8.2%)	(-1.4%, 9.0%)
Percent of high-burden strategy attempts	Table 5.35	-2.5%	0.129	(-5.3%, 0.2%)	(-5.8%, 0.7%)
FR payroll-only days	Table 5.38	-1.0%	0.605	(-4.2%, 2.2%)	(-4.8%, 2.7%)
Late Mail Returns	Table 5.41	-2.3%	0.025	(-4.0%, -0.6%)	(-4.3%, -0.3%)

Source: American Community Survey Paradata, August 2015.

Table 5.47 gives a summary of all the hypothesis tests of differences between Treatment 2 and Treatment 3 in the results. We find only three variables to be statistically significant, though the same multiple comparisons issues apply as above. Positive observed difference values occur when the Treatment 2 statistic was larger than the statistics for Treatment 3. For a summary of results using the difference of differences statistic, see Table 10.2 in the appendix.

Table 5.47 Summary of hypothesis tests and confidence intervals for the difference between Treatment 2 and Treatment 3

	Table Number	Observed Difference	p-value	90% CI	95% CI
Total attempts per case	Table 5.7	0.001	0.992	(-0.244, 0.247)	(-0.291, 0.293)
Contacts per case	Table 5.8	-0.006	0.886	(-0.068, 0.057)	(-0.079, 0.068)
Firm refusals per case	Table 5.9	0.009	0.162	(-0.002, 0.019)	(-0.003, 0.021)
Personal visits attempts per case	Table 5.10	-0.115	0.197	(-0.262, 0.031)	(-0.289, 0.058)
Telephone attempts per case	Table 5.11	0.106	0.045	(0.018, 0.194)	(0.002, 0.21)
Interviewing hours per case	Table 5.13	-0.059	0.353	(-0.162, 0.044)	(-0.181, 0.063)
Miles per case	Table 5.14	-1.418	0.388	(-4.079, 1.243)	(-4.570, 1.733)
Response rate	Table 5.15	0.6%	0.502	(-0.8%, 1.9%)	(-1.0%, 2.1%)
Pre-and post-work transmission compliance rate	Table 5.20	4.3%	0.035	(0.9%, 7.7%)	(0.3%, 8.3%)
Percent of cases with at least one firm refusal	Table 5.28	0.6%	0.269	(-0.3%, 1.5%)	(-0.5%, 1.7%)
Percent of cases with multiple firm refusals	Table 5.29	0.4%	0.276	(-0.2%, 0.9%)	(-0.3%, 1.0%)
Percent of attempts "observing household from vehicle"	Table 5.31	1.3%	0.101	(0.0% 2.6%)	(-0.2% 2.8%)
Percent of attempts that were by telephone	Table 5.32	3.7%	0.041	(0.7%, 6.7%)	(0.2%, 7.2%)
Percent of low-burden-strategy attempts	Table 5.35	-5.4%	0.110	(-10.9%, 0.2%)	(-11.9%, 1.1%)
Percent of low burden strategy attempts when burden at least 30	Table 5.36	-4.4%	0.248	(-10.6%, 1.8%)	(-11.7%, 3.0%)
FR payroll-only days	Table 5.38	2.9%	0.187	(-0.7%, 6.5%)	(-1.3%, 7.1%)

Source: American Community Survey Paradata, August 2015.

The permutation methods used in the report to make inferences can easily be extended to cover simultaneous testing of multiple hypotheses. When permuting the treatment groups, we can calculate not only the marginal null distribution of each test statistic but also the joint distribution

of all test statistics. The joint null distribution of test statistics is not known to the researcher in most situations, which accounts for the use of a technique like the Bonferroni correction.

Assume that we wish to test *k* hypotheses. To calculate an adjusted p-value that controls the family-wise (also called experiment-wise) error rate (the probability of at least one type-I error), we propose the following method that can readily be implemented in a permutational testing framework:

1) Calculate the joint null distribution of the *k* tests
2) For a given hypothesis, calculate the percentile of the observed test statistic compared with the null distribution. Call the percentile *x*.
3) Calculate the probability (over all permutations) that at least one null distribution has a value greater than or equal to its *x* percentile.
4) This probability is the adjusted p-value that controls the family-wise error rate. Two-sided p-values are obtained by applying this procedure to the absolute value of the test statistic.
5) Repeat for all *k* hypotheses

For example, we apply this method to six of the hypothesis tests from Table 5.46. We show both the unadjusted and adjusted p-values in Table 5.48. Note that the more hypothesis tests we include the more conservative the adjusted test becomes and the larger the adjusted p-values will be.

Table 5.48 Multiple-comparison adjusted p-values for the difference between Treatments 2–3 combined and Treatment 1

	Unadjusted p-value	Adjusted p-value
Total attempts per case	0.074	0.339
Contacts per case	0.067	0.308
Firm refusals per case	0.032	0.164
Personal visits attempts per case	0.198	0.680
Telephone attempts per case	0.101	0.425
Response rate	0.104	0.436

Source: American Community Survey Paradata, August 2015.

6. Summary

Based on the results observed during the pilot, the cumulative burden score stopping rule was effective at reducing some of the metrics we used to evaluate the perceived contact burden

associated with the contact efforts in the ACS CAPI operation. In summary, the key results[10] below were observed.

Workload:

The percent of cases pulled from Treatments 2 and 3 were 4.5 percent and 4.1 percent respectively. The specific percent of cases in the pilot pulled from Treatments 2 and 3 in each SSFA ranged from 2.1 percent to 10.6 percent. The majority of FRs in treatment groups 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. In the pilot, at most six assigned cases were pulled from any FR. Cases located in areas associated with response rate Performance Clusters 2 and 3 (i.e. areas with historically lower response rates) were pulled at higher rates than those associated with response rate Performance Cluster 1 (i.e. areas with historically higher response rates). However, this may be partly due to Performance Clusters 2 and 3 having a higher proportion of cases with non-zero initial burden scores.

Perceived Contact Burden:

We observed a 6.0 percent decrease in average reported contact attempts per case and a 6.2 percent decrease in the reported contacts per case in Treatments 2 and 3 over Treatment 1. We also observed a 19.4 percent reduction in reported sample-person contacts with a firm refusal for Treatments 2 and 3 over Treatment 1. However, we note that the number of such contacts is overall quite small.

Implementing a stopping rule based on the cumulative burden score reduced the number of cases with high burden scores. There are similar distributions of cases with burden scores less than the threshold of 40 for all treatment groups, but only 0.3 percent of cases in Treatments 2 and 3 had 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 in the Control group more than 0.5 percent of cases had burden scores over 80.

Measures of perceived contact burden are based on paradata reported by FRs, therefore we cannot be certain in all instances whether these measures reflect actual changes in attempts made or, instead, reflect changes in reporting by some FRs. However, many indicators of FR reporting

[10] In the results, we only make statistical comparisons between Treatments 1, 2 and 3. This is because, in the design, pilot FSAs were randomized between those treatments. We do not statistically test comparisons between the control group and the other treatments because the assignment of SSFAs to the control group was not done entirely at random.

behavior did not demonstrate evidence of changes in their reporting behavior (see "Field Operations" below).

Interviewing Hours and Miles:

While FRs made fewer reported contact attempts per case in Treatments 2 and 3 during the August pilot, we did not find evidence that interviewing hours or miles per case decreased. This finding may be explained in part by the hours per case ceiling in place for ACS data collection operations. If FRs still have hours remaining within the allocation provided for their workload, then they may make more attempts on remaining cases after other cases are removed from their workload for exceeding the cumulative burden score threshold. These hours per case allocations were not modified during the pilot test to account for cases pulled from FR workloads. Previously released research modeling the effects of the proposed stopping rules on respondent burden, cost, and quality using 2012 paradata estimated a national reduction of 4.4 percent of interviewing hours based on the implementation of a cumulative burden score stopping rule (Griffin, Slud, and Erdman 2015). The findings by Griffin et al. may inform future decisions about appropriate interviewing hours per case allocations during production implementation of these procedures.

Response Rates:

Lower response rates were observed in Treatments 2 and 3 versus Treatment 1, due to cases being removed for exceeding the cumulative burden score threshold. We estimate that the response rate difference was borderline-significant at 1.3 percentage points lower (two-tailed p-value =0.104) for Treatments 2 and 3 versus Treatment 1. Comparing this estimate (1.3 percent) to the percentage of cases pulled (4.3 percent), one can argue that most pulled cases would not have resulted in completed interviews even if the burden stopping rules were not in place. The largest decrease in response rate for Treatments 2 and 3 compared with Treatments 1 and Control are for units in Performance Clusters 2 and 3 with higher initial burden scores.

Field Operations:

FR compliance with the twice-a-day transmission guidelines was uneven across SSFAs. In the three treatment groups, compliance was on average 20.9 percent for the start of the work-day transmissions and 83.7 percent for the end of the work-day transmissions. FRs in Treatment 2 were more compliant (20.2 percent overall) than in Treatments 1 and 3 where the respective overall compliance rates were 16.4 and 15.4 percent. Compliance rates for Treatment 2 and 3 were significantly different (p-value = 0.035). Transmission compliance was lower than needed to ensure burden scores were updated accurately each day and cases were pulled in a timely manner, and therefore some objectives of this pilot were not fully realized. As an indication of this, out of 748 cases in Treatments 2 or 3 attaining cumulative burden of 41 or more, 84 had an additional burden increment on a day subsequent to hitting 41, and of these, only 37 were

eventually pulled. These 84 cases generated an average of 10.0 burden points subsequent to the end of the day on which their cumulative burden reached 41.

Errors were identified, and later corrected, during the pilot in the reports that managers had available to monitor compliance. These errors reduced managers' ability to intervene when FRs were not following procedures, which potentially contributed to these low rates. Correctness of these reports must be confirmed prior to implementation in production of case removal based on the burden score, and managers must give transmission compliance significant attention, to ensure that burden scores are accurately updated each day and cases pulled in a timely manner.

The cumulative burden score calculation relies heavily on the paradata FRs record in the Contact History Instrument (CHI), and the quality of these CHI entries are affected by FR compliance with procedures requiring they 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 CHI reporting behavior during the pilot. Indicators of FR CHI-reporting behavior were found to change little across control and treatment groups during the pilot study. These included the proportions of CHI entries corresponding to:

- not attempting contact;
- observing the household from the vehicle;
- personal visit versus telephone attempts;
- attempts made before noon, early afternoon, later afternoon and post-6 p.m.;
- weekday versus weekend attempts;
- attempts in which low-, medium- or high-burden "strategies" were reported;
- FR-days in which only payroll and no CHI entries were reported;
- FR-days in which both payroll and CHI were reported with some interviewing hours and miles recorded in which personal-visit attempts were made.

FRs may have changed their behavior in some instances. There were small reductions for Treatments 2 and 3 compared to Treatment 1 and Control in the proportion of cases with at least one firm refusal (i.e. the FR indicated the respondent was hostile, not interested, hung-up or slammed the door, or intended to quit the survey) or with two or more interim outcomes reflecting either a firm refusal or other respondent refusal.

7. Next Steps

In consideration of these results 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. We recommend that the

Census Bureau continue to make preparations for a nation-wide implementation of a cumulative burden score stopping rule in the Spring of 2016.

Additionally, we note the following areas to be considered for future research:

1. In the pilot, we were not able to analyze the survey responses of the units to determine if there were differences between treatment groups with respect to their demographic characteristics. Such differences could have important consequences on response bias of ACS estimates. Research on possible national and subnational response biases due to the protocol of case removal by exceeding burden score thresholds will be difficult after national implementation of the case removal policy but should be undertaken either before or after implementation.

2. We assigned burden score increment values and the burden score threshold based on the assessment by Griffin et al. (2015) of the relative burden of the various contact attempts and not based on any empirical evidence or input from ACS respondents to validate their perceptions of the relative contact burden for various contact methods. Cognitive research on the relation between contact attempt characteristics and respondent perceptions of burden would be beneficial to ensure that the case removal strategy of the pilot has the desired effect of reducing actual respondent burden.

3. Additional work is needed to adjust FR performance standards to reflect the burden score stopping rule. Because the burden score stopping rule is contingent on FR use of the CHI, it is imperative to consider ways to reduce any motivation to misrepresent CHI entries or not comply with transmission policy.

4. Additional analysis of CHI data might provide FRs helpful guidance in the best data-driven strategies to employ in making successful attempts and reducing burden. Using historical CHI data, the best strategies could be identified and shown to FRs in relation to specific cases they are working. A tool doing this would have the added benefit of providing FRs an incentive to record contact attempts accurately in CHI. However, if the burden score case-removal rule is implemented nationwide, then research on optimal FR strategies will also require new data on FR CHI entries and case outcomes after implementation.

5. During the pilot, when a case was pulled from the workload due to surpassing the burden score threshold, sometimes the potential respondent either had a scheduled appointment with an FR or called the telephone assistance line. In this situation, the potential respondent was instructed to complete the interview on the telephone or themselves, by mail or online. Assuming that similar procedures would be adopted when case removals for excess burden are later implemented nationwide, then an increase of the rate of LMRs

should be expected. Further research should be conducted to analyze the percentage of pulled cases that are converted to LMRs. To facilitate this research, it is necessary for the data concerning how final outcome codes change over time to be saved. Currently this is not the case.

6. Research will be needed to inform future revisions of the burden score, including possible alterations to of the value of initial scores received from mail and CATI modes, the handling of different types of reluctance expressed by the contacts in noninterview attempts, and the value of contact attempt records in which no personal contact is attempted.

7. Geographic variation between sample cases, and variations across RO and FSA in administering FR case handling and reporting compliance, require further study in relation to the occurrence of high burden cases and the interaction between the case removal policy and other FR incentives. The objective here would be to document best practices in case management at the SSFA and FSA level.

8. Acknowledgments

The authors acknowledge the huge contributions to this pilot of the members of the interdivisional "ACS CAPI Burden Reduction Research" team – Todd Hughes (team leader), Eric Slud, Robert Ashmead, Rachael Walsh, Gina Walejko, Padraic Murphy, Elizabeth Poehler, Mary Frances Zelenak, Donna Daily, Fern Bradshaw, Yorlunza Brown, Alexandria Fraley, Michelle Wiland, Paul Ehmann, Gerson Morales, Joanne Pascale, and Chandra Erdman. In addition, we received comments from other critical reviewers assigned to this project, specifically: Tony Tersine, Beth Tyszka, Asaph Young Chun, and Joe Schaefer.

9. References

Bates, N., Dahlhamer, J., Phipps, P., Safir, A. & Tan, L. 2010. Assessing Contact History Paradata Quality across several Federal Surveys. Proceedings of American Statistical Association, Survey Research Methods Section, 91-105.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (Vol. 2). Pacific Grove, CA: Duxbury.

Compton, E. & Bentley, M. 2012. 2010 Census Nonresponse Followup (NRFU) Contact Strategy Experiment Report. 2010 Census Planning Memoranda Series No. 174. http://www.census.gov/2010census/pdf/2010_Census_NRFU_Contact_Strategy_Experiment.pdf

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application* (Vol. 1). Cambridge University Press.

Frankel, J. & Sharp, L. 1981. Measurement of Respondent Burden. Statistical Reporter. January 1981.

Griffin, D. 2014. Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 2 Results. 2014 American Community Survey Research And Evaluation Report Memorandum Series #ACS14-RER-07. http://www.census.gov/library/working-papers/2014/acs/2014_Griffin_01.pdf

Griffin, D. & Hughes, T. 2013. Analysis of Alternative Call Parameters in the American Community Survey's Computer Assisted Telephone Interviewing. ACS13-RER-11. http://www.census.gov/library/working-papers/2013/acs/2013_Griffin_03.html

Griffin, D. & Nelson, D. 2014. Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 1 Results (Part 2). ACS14-RER-22. http://www.census.gov/library/working-papers/2014/acs/2014_Griffin_02.pdf

Griffin, D., Slud, E. & Erdman, C. 2015. Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 3 Results. ACS15-RER-28-R1. http://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Griffin_01.pdf

Lohr, S. (2009). *Sampling: Design and Analysis*. Cengage Learning.

McCarthy, J. 2011. Using Predictive Models with Measures of Burden Reporting History, and External Data to Identify Likely Establishment Survey Nonrespondents. BLUE-ETS Conference on Burden and Motivation in Official Business Surveys.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Roczniki Nauk RolniczychTom X* [in Polish]; translated in *Statistical Science*, 5, 465–480.

Olson, R. (2015), "2015 ACS-HU CAPI Field Pilot to Reduce Respondent Burden", ACS-HU Regional Office Memorandum No. 15-16, distributed to all Region RO Directors.

Poe, T. 2011. "The Pros and Cons of Making the American Community Survey Voluntary." Prepared Testimony of Ted Poe (TX-02) Before the House Oversight and Government Reform Committee's Subcommittee on Health Care, District of Columbia, Census, and National Archives. March 6, 2011.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sears, J. 2011. Response Burden Measurement and Motivation at Statistics Canada. 2011. BLUE-ETS Conference on Burden and Motivation in Official Business Surveys.

Van den Brakel, J.A. and Renssen, R. H. (1998). Design and analysis of experiments embedded in sample surveys. Journal of Official Statistics, 14(3), 277.

Van den Brakel, J. A., & Renssen, R. H. (2005). Analysis of experiments embedded in complex sampling designs. Survey Methodology, 31(1), 23-40

Van Den Brakel, J. A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. Journal of the Royal Statistical Society: Series A (Statistics in Society), 171(3), 581-613.

Virgile, M. (2015) Measurement Error in American Community Survey Paradata and 2014 Redesign of the Contact History Instrument. American Community Survey Research And Evaluation Report Memorandum Series #ACS15-RER-12.

Zelenak, M.F. (2014). Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation – Phase 1 Results (Part 1). 2014 American Community Survey Research And Evaluation Report Memorandum Series #ACS14-RER-06. http://www.census.gov/library/working-papers/2014/acs/2014_Zelenak_01.pdf

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R*. John Wiley & Sons.

10. Appendices

10.1 Permutation tests for difference of differences

In Table 10.1, positive observed differences imply the estimated value for Treatments 2 and 3 is lower than the value for Treatment 1. The Table gives the formal assessments of significance of the Treatment group differences.

Table 10.1 Summary of hypothesis tests and confidence intervals for the difference of differences between Treatment 1 and Treatments 2–3

	Table Number	Observed difference	p-value	90% CI	95% CI
Total attempts per case	Table 5.7	0.280	0.009	(0.101, 0.459)	(0.067, 0.493)
Contacts per case	Table 5.8	0.053	0.032	(0.012, 0.094)	(0.004 0.101)
Firm refusals per case	Table 5.9	0.015	0.084	(0.001, 0.029)	(-0.002, 0.031)
Personal visits attempts per case	Table 5.10	0.176	0.001	(0.083, 0.268)	(0.066, 0.285)
Telephone attempts per case	Table 5.11	0.074	0.182	(-0.017, 0.165)	(-0.033, 0.182)
Interviewing hours per case	Table 5.13	0.039	0.414	(-0.038, 0.116)	(-0.052, 0.130)
Miles per case	Table 5.14	0.169	0.837	(-1.166, 1.504)	(-1.417, 1.755)
Response rate	Table 5.15	1.2%	0.109	(0.0%, 2.5%)	(-0.3%, 2.7%)
Response rate among cases with high initial burden	Table 5.17	2.7%	0.034	(0.6%, 4.9%)	(0.2%, 5.3%)
Percent of cases with at least one firm refusal	Table 5.28	1.1%	0.086	(0.0%, 2.1%)	(-0.1%, 2.3%)
Percent of cases with multiple firm refusals	Table 5.29	0.8%	0.102	(0.0%, 1.6%)	(-0.2%, 1.7%)
Percent of attempts that were by telephone	Table 5.32	0.7%	0.592	(-1.5%, 2.9%)	(-1.9%, 3.3%)
Percent of low-burden strategy attempts	Table 5.35	2.8%	0.076	(0.2%, 5.4%)	(-0.3%, 5.9%)
Percent of high-burden strategy attempts	Table 5.35	-2.9%	0.010	(-4.7% -1.0%)	(-5.1% -0.7%)
Percent of attempts "not attempting contact"	Table 5.27	-0.4%	0.726	(-2.3%, 1.5%)	(-2.7%, 1.9%)
Percent of cases with at least one "other" concern	Table 5.30	2.1%	0.111	(-0.1%, 4.2%)	(-0.5%, 4.6%)
Late Mail Returns	Table 5.41	-0.6%	0.408	(-1.9%, 0.6%)	(-2.1%, 0.8%)

Source: American Community Survey Paradata, July and August 2015.

In Table 10.2, a positive observed difference implies the estimated difference between August and July for Treatment 3 is larger than that of Treatment 2, thus favoring Treatment 3. Again, the Table provides the formal assessment of significance of those estimated differences.

Table 10.2 Summary of hypothesis tests and confidence intervals for the difference of differences between Treatment 2 and Treatment 3

	Table Number	Observed difference	p-value	90% CI	95% CI
Total attempts per case	Table 5.7	-0.013	0.920	(-0.226, 0.199)	(-0.265, 0.238)
Contacts per case	Table 5.8	0.048	0.140	(-0.005, 0.100)	(-0.014, 0.109)
Firm refusals per case	Table 5.9	0.007	0.408	(-0.007, 0.021)	(-0.009, 0.024)
Personal visits attempts per case	Table 5.10	0.035	0.622	(-0.080, 0.150)	(-0.101, 0.171)
Telephone attempts per case	Table 5.11	-0.017	0.755	(-0.106, 0.072)	(-0.123, 0.089)
Interviewing hours per case	Table 5.13	0.079	0.185	(-0.019, 0.177)	(-0.036, 0.194)
Miles per case	Table 5.14	0.378	0.703	(-1.252, 2.009)	(-1.552, 2.309)
Response rate	Table 5.15	0.0%	0.954	(-1.2%, 1.3%)	(-1.4%, 1.5%)
Percent of cases with at least one firm refusal	Table 5.28	0.0%	0.957	(-1.1%, 1.1%)	(-1.3%, 1.3%)
Percent of cases with multiple firm refusals	Table 5.29	0.5%	0.421	(-0.5%, 1.4%)	(-0.6%, 1.5%)
Percent of attempts "observing the household from vehicle"	Table 5.31	0.5%	0.245	(-0.2%, 1.3%)	(-0.3% 1.4%)
Percent of attempts that were by telephone	Table 5.32	-1.1%	0.414	(-3.4%, 1.1%)	(-3.8%, 1.5%)
Percent of low-burden strategy attempts	Table 5.35	0.5%	0.761	(-2.1%, 3.0%)	(-2.5%, 3.5%)

Source: American Community Survey Paradata, July and August 2015.

10.2 Comparison of results from alternative inference approaches

We have discussed in Sections 2.5.2 – 2.5.5 three alternative approaches to the calculation of p-values and confidence intervals for the treatment group contrasts we found in Section 5. Recall that what we called the "design-based" approach in Sec. 2.5.3 actually involved an approximation of the sampling design by a Simple Random Samples of FSAs within selected SSFAs to assign Treatment 1 and then another Simple Random Sample from the FSAs not assigned to Treatment 1 to assign Treatment 2. The actual design of block-randomization of FSAs to treatments by tier (where tiers of three FSAs were first ordered by increasing the proportion of workload coming from Performance Cluster 1) is most accurately reflected by the

"permutational" description of group-contrast statistics described in Section 2.5.4. What we called in Section 2.5.5 the "nonparametric" approach to the distribution of test statistics was to treat all of the fixed finite population of FSA values (under the null hypothesis) as though they came from identically distributed (or at least exchangeable) continuously distributed random variables. The main difference between the nonparametric and the other approaches was that in Section 2.5.5, the test statistics for group differences were rank tests (the two-sample rank-sum or Wilcoxon statistic) referred to their approximately normal null-hypothesis distributions. The "design-based" analysis was rigorously derived subject to the SRS approximation of the FSA randomization. As was discussed in Section 2.5.4, the permutational calculation of null-hypothesis variances and statistic distributions precisely reflects the actual random treatment assignment mechanism in the pilot experiment, making that the analytical method of choice. Note that the distributions of test statistics can my calculated by Monte Carlo estimation and thereby checked for approximate normality within all three methods of analysis. This has been done carefully only for the permutational method of analysis. For both test statistics and all comparisons, the permutational distributions were extremely close to normal, and p-values calculated from normal percentage points using permutationally calculated variances were essentially indistinguishable from those derived from the quantiles of the Monte-Carlo permutational distribution.

Tables 10.3 and 10.4 compare the p-values for selected treatment-group differences analyzed in this report. Although the different methods of analysis rest on slightly different assumptions and approximations, the p-values are seen to be so close that the choice of method does not appear to have been critical to the conclusions about significance in this report, although the precise test-based permutational confidence intervals reported are slightly different than the corresponding intervals for the other methods.

Table 10.3 P-values from permutation, design-based, and rank-based tests of the difference between the outcomes of Treatment 1 vs. Treatment 2 –3

	Permutation p-value	**Design-based p-value**	**Rank-based p-value**
Total attempts per case	0.074	0.071	0.077
Contacts per case	0.067	0.073	0.062
Firm refusals per case	0.032	0.029	0.074
Personal visits attempts per case	0.198	0.208	0.168
Telephone attempts per case	0.101	0.053	0.070
Interviewing hours per case	0.854	0.831	0.321
Miles per case	0.615	0.679	0.813
Response rate	0.104	0.127	0.082

Source: American Community Survey Paradata, July and August 2015.

Table 10.4 P-values from permutation, design-based, and rank-based tests of the difference between the outcomes of Treatment 2 vs. Treatment 3

	Permutation p-value	**Design-based p-value**	**Rank-based p-value**
Total attempts per case	0.992	0.993	0.916
Contacts per case	0.886	0.896	0.278
Firm refusals per case	0.162	0.153	0.485
Personal visits attempts per case	0.197	0.150	0.388
Telephone attempts per case	0.045	0.090	0.891
Interviewing hours per case	0.353	0.218	0.253
Miles per case	0.388	0.365	0.717
Response rate	0.502	0.377	0.891

Source: American Community Survey Paradata, July and August 2015.