Inference from Complex Survey-Embedded Field Experiments

Robert Ashmead[*] Eric Slud[†]

Abstract
Experiments embedded in surveys are a valuable way to test potential survey improvements; however, inference from these experiments can be difficult because one must account for the survey design aspects. Operational issues can restrict possible experimental design choices. In this paper, we discuss a design-based and a randomization-based approach to inference for experiments in an embedded survey. Additionally, some surveys are ongoing and present the option of utilizing data from previous time periods as well as the data from the test period. We discuss and compare an estimator that utilizes data only from the test period with one that compares outcome differences between the study period and a prior period across study treatment groups. Using an example from the American Community Survey (ACS) as well as a simulation, we compare the two inferential approaches and estimators. We find that the permutation-based inferential approach is a flexible and robust method for analyzing embedded experiments and that when the correlation is large between time periods, the difference in time-period differences may provide inferences with greater precision.

Key Words: survey sampling; randomized experiments; design-based estimation; randomization-based estimation

1. Introduction

Randomized controlled experiments are the ideal way to evaluate the impact of interventions, but when an experiment is embedded in a probability sample survey certain survey design and administrative field aspects can make analysis and inference difficult. For example, it may not be possible to completely randomize treatments to individual units because of an administrative structure, making it necessary to instead randomize clusters of units. These complicating factors must be taken into account when conducting the analysis of the experiment and may cause the assumptions for typical approaches to inference to be invalid. In this paper we discuss two aspects of survey-embedded field experiments. The first is how to account for a complicated randomization structure, and the second is the utility of incorporating data from previous time periods in an ongoing survey.

As motivation, we consider an embedded experiment from the American Community Survey (ACS). The ACS is an ongoing national survey that samples approximately 295,000 housing unit addresses per month. It uses a multi-mode collection strategy in which sampled addresses are initially contacted by mail. Then further contact attempts may be made to non-respondents by each of mail, computer-assisted telephone interviewing (CATI), and computer-assisted personal interviewing (CAPI). In each of these modes, multiple attempts may be made. This data collection strategy can be perceived as overly intrusive to some respondents. In response to these concerns, the Census Bureau conducts ongoing research into methods for reducing respondent burden. One of the lines of research is to develop a stopping rule for CAPI based on a cumulative burden score. The idea is that each contact attempt counts as a separate increment of burden, and once the cumulative burden score exceeds a certain threshold, no further attempts are to be made. Each contact attempt is given a numerical score based on its type and outcome. Contact attempts judged to be

[*] U.S. Census Bureau

[†] U.S. Census Bureau & University of Maryland, College Park

more burdensome are given larger scores. For example, an in-person contact attempt that resulted in a firm refusal would contribute a larger amount to the cumulative burden score than a personal visit in which no one was available at the sampled household.

Based on earlier research and development of such a stopping rule (Griffin, Slud, and Erdman 2015), a pilot study was undertaken in the ACS CAPI data collection month of August 2015 to test the effects of the stopping rule on costs, data quality and perceived contact burden (Hughes, Slud, Ashmead, and Walsh 2016). The pilot took place in roughly one quarter of all ACS interviewing areas and randomized clusters of field representatives to one of three groups. The first group was the control group, while the second and third group instituted the burden score stopping rule. While both the second and third group used the burden score stopping rule, only the second group of field representatives could view the cumulative burden score themselves during the pilot. The results from the pilot showed that the stopping rule had a small negative impact on response rates, but was effective at reducing some measures of perceived contact burden (Hughes, Slud, Ashmead, and Walsh 2016).

The analysis of the burden reduction pilot was complicated by the fact that the treatment randomization scheme was constrained by operational concerns and an attempt to balance the interview difficulty between treatments. In the pilot, clusters of field representatives were randomized to three treatment groups. The clusters were defined as Field Supervisory Areas (FSAs) within which all field representatives shared the same supervisor, which was deemed to be an operational necessity. In an attempt to balance the difficulty of interviews between treatments, FSAs were ordered into successive tiers of three by metrics of interview difficulty within each Survey Statistician Field Area (SSFA). Then within each tier of three, each of the three treatments was randomly assigned to exactly one FSA. Therefore, the experimental design used both clusters and strata. Adding to the challenges of the analysis, the number of field representatives in a FSA varied as did the number of cases (sampled households) that a field representative attempted to interview. Lastly, though treatment groups were assigned to field representatives, we were ultimately interested in the treatment effects at the case-level, that is, at the single housing-unit level.

In order to perform a rigorous statistical analysis in an experiment such as the burden score reduction pilot of the ACS, it is necessary to properly account for all the aspects of the experimental design. In this paper we first discuss approaches to inference and analysis to take experimental design into account. Additionally, an ongoing survey such as the ACS presents the opportunity to utilize data from previous data collection periods as well as the pilot period. We also explore the utility of using data from a previous collection period to help estimate the treatment effect.

2. Approaches to Inference

2.1 Designed-based and Permutation-based Inference

In a randomized controlled experiment, imagine that each unit (household in our example) has a potential outcome (Neyman, 1923) for each of two treatments. We are only ever able to observe one of the two treatments for each unit, but we would like to estimate the population-level treatment means in order to estimate the treatment effect. Next, consider that the units selected for the experiment are actually a probability sample from a larger population. After being selected for inclusion in the survey, the units are randomized to the treatment groups using some specified design. We consider two different approaches to inference in this setting, design-based and permutation-based. In general, we can use the same point estimator of the treatment effect for both methods, but the methods to estimate the variance of that treatment effect and make inference are different.

Thinking about the problem from a design-based perspective and drawing from survey sampling theory, we can view the units in each treatment assignment as a random subsample from the main survey sample and then use the inverse probability of treatment selection to estimate population-level quantities. Variance estimates can be made using slight modifications of existing survey variance formulas. This is the approach taken by Van den Brakel and Renssen (1998 and 2005) to estimate variances of mean differences in embedded survey experiments.

In the example of the ACS burden reduction pilot, the primary goal of the pilot was to understand the effects of the intervention. As a result, geographies to be in included in the sample were chosen deliberately, making estimating population-level effects improper. Still, we can use the concept of design-based inference to estimate treatment effects and associated variances where our population of interest is all units in the pilot. Applying this method to our example, the design was a stratified cluster design.

Another way to think about making inference is under a permutation test approach (Zieffler, Harring, Long 2011). Under the null hypothesis that there is no difference between the treatments, the treatment assignments are simply labels assigned at random to observations. Therefore, if we create permutations of the treatment assignments following the same randomization scheme, and apply these permutations of the treatment assignments to the actual data, we get an exact distribution of the test statistic under the null hypothesis. This permutation-based null distribution can be used in conjunction with the observed test statistic to give a p-value from a hypothesis test and a test-based confidence interval. Rather than calculating all possible treatment assignment permutations, we can simply take a Monte Carlo sample of all possible permutations to get an appropriate sample from the null distribution. In our ACS example, we are not interested in applying probability sampling weights, but if we did, it is not clear how to incorporate them with a permutation-based estimator, which is a major disadvantage.

2.2 Difference in Difference Estimator

In an experiment, the typical estimate of the treatment effect is simply the difference in means of the treatment groups. In our example, the difference in means for any two treatment groups during the test period can we written as:

$$\begin{aligned}\hat{\Delta}_D = \frac{\sum_{i=1}^n z_i m_i \bar{y}_i}{\sum_{i=1}^n z_i m_i} - \frac{\sum_{i=1}^n (1-z_i) m_i \bar{y}_i}{\sum_{i=1}^n (1-z_i) m_i} \qquad (1)\\ m_i \text{ for } i = 1, \ldots, n;\\ y_{ij} \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i;\\ z_i \text{ for } i = 1, \ldots, n;\\ \bar{y}_i = \tfrac{1}{m_i} \textstyle\sum_{j=1}^{m_i} y_{ij},\end{aligned}$$

where n is the number of FSAs, m_i is the number of cases (households) in FSA i during the test period, y_{ij} is the outcome for case j in FSA i, $\bar{y}_i$ is the FSA mean outcome during the test period, and z_i is the indicator of treatment assignment in FSA i during the test period. In the ACS example, the outcomes of interest, y_{ij}, are metrics such as the number of contact attempts or firm refusals. For reference, we call $\hat{\Delta}_D$ the difference within test period estimator.

The ACS is an ongoing monthly survey with differences in response behavior observed across months as well as across geographies, so we might try to use the data from prior months to help inform the treatment effect. In comparison to (1), we might use a difference in difference estimator to utilize prior data. In the case of the ACS burden score reduction pilot, we write the difference in difference estimator as:

$$\hat{\Delta}_{DD} = \frac{\sum_{i=1}^{n} z_i(m_i + m_i^*)(\bar{y}_i^* - \bar{y}_i)}{\sum_{i=1}^{n} z_i(m_i + m_i^*)} - \frac{\sum_{i=1}^{n}(1 - z_i)(m_i + m_i^*)(\bar{y}_i^* - \bar{y}_i)}{\sum_{i=1}^{n}(1 - z_i)(m_i + m_i^*)}, \tag{2}$$

$$\begin{aligned} & m_i,\ m_i^* \text{ for } i = 1, \ldots, n; \\ & y_{ij} \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i;\ y_{ij}^* \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i^*; \\ & z_i \text{ for } i = 1, \ldots, n; \\ & \bar{y}_i = \tfrac{1}{m_i} \textstyle\sum_{j=1}^{m_i} y_{ij};\ \bar{y}_i^* = \tfrac{1}{m_i^*} \textstyle\sum_{j=1}^{m_i^*} y_{ij}^*, \end{aligned}$$

where n is the number of FSAs, m_i and m_i^* are the number of cases (households) in FSA i during the test period and previous time period respectively, y_{ij} and y_{ij}^* are the outcomes for case j in FSA i during the test period and previous time period respectively, $\bar{y}_i$ and $\bar{y}_i^*$ are the FSA mean outcomes during the test period and previous time period respectively, and z_i is the indicator of treatment assignment in FSA i during the test period.

While we can use either the design-based or permutation-based approach to estimate the variance of the estimator with the difference within test period estimator (1), it is not clear how to use a design-based approach with the difference in difference estimator (2). However, because of its flexibility we can still use the permutation approach with the difference in difference estimator.

3. Results

In this paper we are interested in the results of two main questions. First, do we get similar results when estimating the variance using the design-based approach compared with permutation-based approach. Second, should we prefer the difference in difference estimator or the typical difference within test period estimator in analyzing our experiment?

The design of the example ACS experiment makes it analogous to a stratified cluster survey; however in the example there was generally only a single cluster in each strata. Therefore we ignored the stratification aspect for the purposes of estimating variance in the design-based approach. Under the design-based method, we estimated the variance of $\hat{\Delta}_D$ by an approximation of the differences between two treatment groups using a standard survey sampling formula for the variance of a ratio estimator in a cluster sample of clusters with unequal size (Lohr, 2009). We estimated the variance of $\hat{\Delta}_D$ under the design-based approach by

$$v\hat{a}r(\hat{\Delta}_D) = \frac{1}{n_1 \bar{m}_1^2} \sum_{i=1}^{n} \frac{z_i m_i^2 (\bar{y}_i - \hat{\bar{y}}_1)^2}{n_1 - 1} + \frac{1}{n_0 \bar{m}_0^2} \sum_{i=1}^{n} \frac{(1 - z_i) m_i^2 (\bar{y}_i - \hat{\bar{y}}_0)^2}{n_0 - 1}$$

$$\begin{aligned} & n_1 = \textstyle\sum_{i=1}^{n} z_i;\ n_0 = \sum_{i=1}^{n}(1 - z_i); \\ & \bar{m}_1 = \tfrac{1}{n_1} \textstyle\sum_{i=1}^{n} z_i m_i;\ \bar{m}_0 = \tfrac{1}{n_0} \textstyle\sum_{i=1}^{n}(1 - z_i) m_i; \\ & \hat{\bar{y}}_1 = \tfrac{\sum_{i=1}^{n} z_i m_i \bar{y}_i}{\sum_{i=1}^{n} z_i m_i};\ \hat{\bar{y}}_0 = \tfrac{\sum_{i=1}^{n}(1-z_i) m_i \bar{y}_i}{\sum_{i=1}^{n}(1-z_i) m_i}. \end{aligned}$$

In order to make inference for the design-based approach we used a p-value calculated from the tail of a normal distribution using the test statistic $\hat{\Delta}_D$ divided by its design-based estimated standard error. This approach relies on having a relatively large sample size to provide justification for the normality assumption. Table 1 shows sample sizes for the three treatment groups in the ACS burden reduction pilot.

In Table 2 we show the permutation-based p-value and the design-based p-value for hypothesis tests of a difference in the outcomes between the control group and the union of

Table 1: Sample sizes for the ACS burden reduction pilot

	Control	Treatment 1	Treatment 2
FSAs	46	46	46
Total cases, test month	4,299	4,135	4,213
Total cases, prior month	4,299	4,147	4,245

Source: Hughes, Slud, Ashmead, and Walsh (2016).

the two treatment groups in the ACS burden reduction pilot. Overall, we see that the two approaches give very similar results, which makes sense because in this case the design-based method is a normal approximation of the permutational approach. We note that the three outcomes with slight differences between the two methods (telephone hours per case, miles per case, and response rate per case) had test statistics with permutation null distributions that were slightly non-normal in the tails. Though in this example we observed that the design-based and permutation approaches gave similar inferential results, in general, the permutation-based approach will be more robust. For a complete discussion of the importance of the results in Table 2 to the ACS burden reduction pilot see Hughes, Slud, Ashmead, and Walsh (2016).

Table 2: P-value comparisons using $\hat{\Delta}_D$, permutation vs. design

Outcome	Permutation-based p-value	Design-based p-value
Attempts per case	0.074	0.071
Contacts per case	0.067	0.073
Firm refusals per case	0.032	0.029
Personal visits per case	0.198	0.208
Telephone attempts per case	0.101	0.053
Interviewing hours per case	0.854	0.831
Miles per case	0.615	0.679
Response rate per case	0.104	0.127

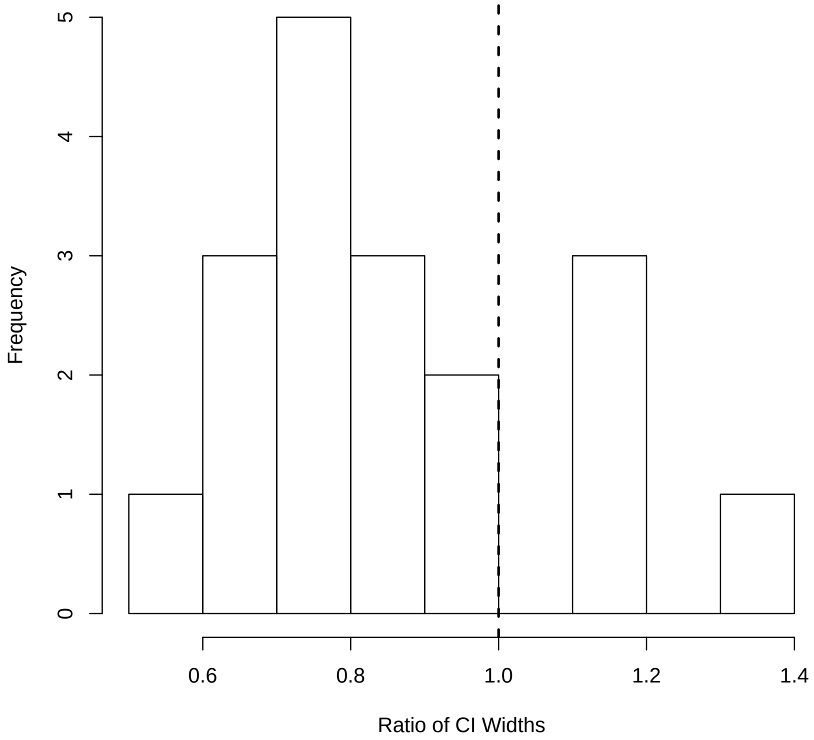
Source: Hughes, Slud, Ashmead, and Walsh (2016).

Next, we explore the results of utilizing additional data from a reference period, comparing estimators (1) and (2). In the ACS burden reduction pilot example, we used the previous month's data (July) as the reference period. While the field representatives in each FSA were not exactly the same in July and August, a large majority of field representatives did not change. Additionally, FSAs are defined geographically, so while the households sampled were not the same between the two months, they should share similar characteristics. Therefore, we think that there will be some level of consistency in the metrics for FSAs across months. This kind of consistency is not guaranteed for all ongoing surveys.

Figure 1 shows the ratio of the widths of 95% confidence intervals for the difference within test period estimators and the difference in differences estimator comparing the control group and the union of the two treatment groups for 18 different outcomes measured in the ACS burden reduction pilot. Ratios less than one indicate that the difference in difference estimator has a smaller 95% confidence interval and is therefore more precise. The figure shows that most, but not all, outcomes have a ratio less than one.

In order to further investigate the differences between the two estimators we created a simulation to mimic the ACS burden reduction experiment. In total, we allowed for $n = 100$ clusters of units over two time periods (test and reference). First, for each cluster

Figure 1: Ratio of the widths of 95% confidence intervals of $\hat{\Delta}_{DD}$ compared with $\hat{\Delta}_D$ for 18 outcomes



Source: Hughes, Slud, Ashmead, and Walsh (2016).

i we randomly selected a number of cases m_i for the test period between 10 and 100, where the number of cases was simulated from a uniform distribution rounded to a whole number. We then chose a related number of cases m_i^* for the reference period by multiplying m_i by a random uniform variable with support $(0.9, 1.1)$, rounding that number and adding 1. Next, a random cluster effect, $\gamma_i \sim \mathcal{N}(0, \sigma^2)$, was simulated for each cluster $i = 1, \ldots, 100$ for a given σ^2 value. For each iteration of the simulation we randomly sampled 50 of the clusters to the treatment group and 50 to the control group using z_i as the indicator of treatment. Next, for a treatment effect Δ, we simulated individual case outcomes in the test period y_{ij} and the reference period y_{ij}^* by

$$y_{ij} = 1 + \lambda_{ij} + z_i\Delta, \text{ where } \lambda_{ij} \sim Poisson(\lambda = 5 + \gamma_i), \text{ and}$$
$$y_{ij}^* = 1 + \lambda_{ij}^*, \text{ where } \lambda_{ij}^* \sim Poisson(\lambda = 5 + \gamma_i).$$

The parameter σ^2 is responsible for inducing correlation between cluster outcomes in the test and reference period. In our simulation we used the values $(0, 0.25, 0.4, 0.5)$ for σ^2 and $\Delta = (0, 0.1)$. For each combination of the parameters σ^2 and Δ, we repeated the simulation $10,000$ times.

Table 3 shows the results of the simulation. The estimated correlation calculated in the table is the correlation among control units in the two time periods. We found that when the correlation was small, then the differences in test period estimator (1) was better in terms of the precision. In those settings with small correlation, using the outcomes from the previous time period only added noise to the estimator. However, as the correlation increased, the differences in differences estimator (2) became more precise, to the point

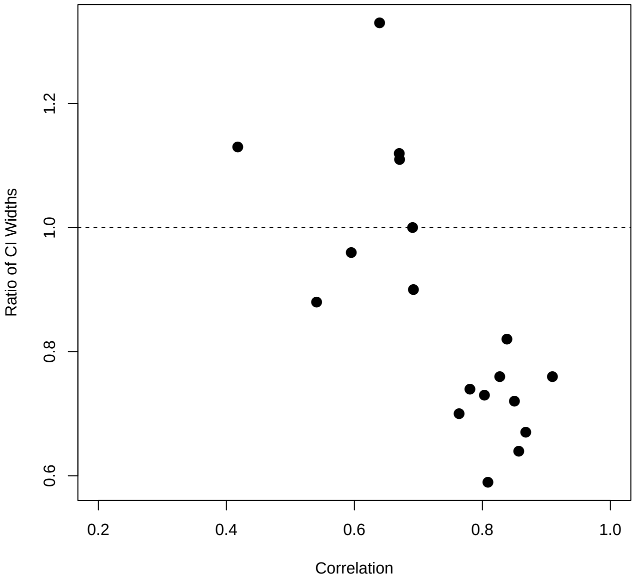
Table 3: Simulation Results

Trt Effect	σ^2	Est. Correlation	MSE Ratio	CI Width Ratio
0	0	0.007	2.157	1.410
0.1	0	0.012	1.933	1.395
0	0.25	0.300	1.088	1.043
0.1	0.25	0.289	1.280	1.084
0	0.4	0.535	0.665	0.817
0.1	0.4	0.541	0.783	0.886
0	0.5	0.640	0.440	0.691
0.1	0.5	0.687	0.410	0.645

where the precision was about twice that of the differences in test period estimator. In the simulation, the bias of each of the estimators was negligible for all cases and therefore not included in Table 3.

Keeping the results from the simulation in mind, we calculated the correlation between time periods among the control units for 18 outcomes in the ACS example. In Figure 2 we plotted those correlations against the 95% confidence interval width ratios from Figure 1. Again, ratios less than 1 suggest that the difference in difference estimator is more precise than the difference in test period estimator. We see from the figure that correlations between time periods are relatively large for most of the 18 outcomes in the example. Further the figure shows that as the correlation increases, the difference in difference estimator does increasingly better compared with the difference in test period estimator. This is exactly the result we would expect given the results from the simulation.

Figure 2: The correlation between time periods plotted against the ratio of the widths of the 95% confidence intervals of $\hat{\Delta}_{DD}$ compared with $\hat{\Delta}_D$ for 18 outcomes in the ACS burden reduction pilot



Sources: Hughes, Slud, Ashmead, and Walsh (2016). American Community Survey Paradata, July & August 2015.

4. Discussion

Experiments embedded in surveys can often have complex designs because of the practical constraints of the field environment, and the necessity of taking design features into account can make statistical analysis difficult. In this paper we discussed two approaches to inference when faced with an embedded survey. We found that both the design-based and permutation-based approaches yielded similar results in our ACS example, but the permutation-based approach was a more flexible and robust method. In the actual analysis of the ACS burden reduction pilot, we decided to use the permutation-based method. However, one downside to the permutation approach is that it is unclear how exactly to incorporate survey weights in the analysis. We also discussed how data from previous time periods might be utilized in an ongoing survey using a differences in differences estimator. We found that when the correlation between outcomes in the two time periods was large, the difference in difference estimator was more precise than an estimator that ignores the data from the previous time period. In our example, the large correlation was due to the fact that a large percentage of the same field representatives were working the same geographic areas across the two time periods.

REFERENCES

Griffin, D., Slud, E. and Erdman, C. (2015). Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation - Phase 3 Results. *American Community Survey Research & Evaluation Rept. Memo. Series* **ACS15-RER-28-R1**.

Hughes, T., Slud, E., Ashmead, R. and Walsh, R. (2016), "Results of a Field Pilot to Reduce Respondent Contact Burden in the American Community Survey's Computer Assisted Personal Interviewing Operation", *American Community Survey Research & Evaluation Rept. Memo. Series* **ACS16-RER-07**.

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

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".

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