

Using an Area Linkage Method to Improve the Coverage of ABS Frames for In-Person Household Surveys

Sylvia Dohrmann and Richard Sigman

Westat

1600 Research Blvd, Rockville, Maryland 20850

Abstract

This paper presents an approach for using address-based sampling (ABS) frames based on U.S. Postal Service (USPS) files in multistage area sample surveys of households. This approach tackles coverage issues by defining two types of segments: area segments and list segments. An area segment is a parcel of land defined by Census geographic boundaries. A list segment is the set of addresses that geocode to a particular area segment, using street-level geocoding. Using the list segments as the survey's secondary sampling units avoids many of the complications associated with inaccurate geocoding; while the geocoding method described in this paper may not place every address accurately, it does assign every address on the USPS list to some list segment, and thus gives all USPS addresses a chance of selection. For in-person surveys, however, the coverage is effectively reduced as a result of addresses that do not correspond to a physical location (e.g., Post Office Box addresses and rural route addresses).

Undercoverage of the ABS frame is addressed by adding an address coverage enhancement procedure. Within a subsample of area segments linked to the survey's sampled list segments addresses present in the area segment but not in the associated list segment are identified. After confirming that the identified addresses do not exist on the ABS frame, they are given a chance of selection for the survey. The aim is to give the addresses identified and sampled by the address coverage enhancement procedure the same selection probabilities as addresses within the same segment sampled from the USPS list. However, practical aspects of the procedure may result in deviations from equal probability to balance the cost of implementation. We describe the address coverage enhancement procedure's sampling concepts and field procedures.

1. Introduction

Historically, traditional listing has been used to create housing unit sampling frames for in-person household surveys. This listing operation requires that field staff canvass the sampled areas in person months before data collection begins in order to list the addresses of all the housing units observed in those areas and to allow adequate time for sample selection. The operation can be both costly and time consuming. Furthermore, given the nature of this task, these lists may be incomplete and contain erroneous listings.

Over the past decade, researchers have been investigating commercially available address lists originating from the United States Postal Service (USPS) as an alternative to traditional listing for creating housing unit sampling frames. The USPS-based lists have some attractive features for sampling housing units for in-person surveys with multistage area sample designs. The lists are electronic, easily input into sample selection software, and are updated monthly. Addresses can be sampled from the lists immediately before data collection, thus avoiding the cost and time needed to physically list the addresses prior to data collection. Sample designs using the postal service lists as housing unit sampling frames have come to be known as address-based sampling (ABS).

While ABS frames are gaining popularity for household surveys, their use with multistage area sample surveys has its limitations; these limitations are described in section 2. Section 3 describes a sampling method that includes an area linkage method for forming segments that mitigates these limitations and facilitates the coverage enhancement procedure which is described in section 4. Section 5 describes the methods for selecting segments for the coverage enhancement procedure and any addresses found through it. Section 6 provides a summary.

2. Limitations of ABS

Two complicating factors impede the use of the USPS-based lists as a complete replacement for the traditional

listing methods. The primary factor is that the lists do not have full coverage of residential addresses, particularly in rural areas. The second factor is that errors occur in grouping the postal addresses into small geographic areas, in particular the secondary sampling units (SSUs or segments) generally used to cluster the sampled households within the sampled primary sampling units (PSUs) for efficient data collection. These two factors are discussed in turn below.

2.1 Noncoverage

There are several reasons why USPS-based residential address lists may not provide complete coverage of the households in an area, especially in the context of area sampling with in-person data collection. Since the USPS files contain only residential addresses along a mail delivery route, addresses of housing units for which mail service is not available are excluded from the list. Residents living in these areas must collect their mail either at a general delivery facility or a Post Office (P.O.) Box. Also, city-style addresses may not be provided for housing units along rural routes. All areas in the U.S. are expected to be eventually converted to city-style addresses for purposes of E-911 location¹; in the interim, unconverted housing units are represented in the ABS frame by only their postal addresses (general delivery or P.O. Box), so that the physical units are not locatable from the mailing address alone. If selected for an in-person survey, such housing units could not be located for interviewing, resulting in noncoverage of housing units that have only a P.O. Box address or collect all their mail from a general delivery facility.

Coverage issues also exist for noninstitutional group quarters such as dormitories, assisted living facilities, halfway homes, and shelters. The presence of these units on the ABS frame depends on how the residents of the facility receive their mail. Some facilities, such as universities, operate their own “post offices,” and thus the USPS does not have information on the individual mailing addresses of the residents. Other facilities, such as assisted living facilities, halfway homes, and shelters may be operated by a business or charitable organization, and may not appear on an address list restricted to residential addresses (Dohrmann, Han and Mohadjer 2006).

Placing the issue of group quarters aside, since not all household surveys include these units in their target population, the highest rates of undercoverage of the address lists have been found to generally be in rural areas (Dohrmann, Han and Mohadjer 2006, Staab and Iannacchione, 2003, and O’Muircheartaigh, Eckman, and Weiss 2002). In fact, many researchers agree that in some very rural PSUs, such as those containing American Indian reservations, USPS mail delivery is not pervasive enough for ABS to be effective. In these PSUs, the best alternative may be to traditionally list the area segments.

Montaquila et al. (2009) found that although coverage rates were generally higher in urban areas, there was variation in coverage rates at the segment level within PSUs, even in very urban PSUs. This variation is such that the USPS-based lists may appear to provide near-complete coverage of some segments and inadequate coverage of others within the same PSU.

Regardless of the extent to which the USPS-based lists are used as housing unit frames, quality control procedures should be considered with ABS frames as with any other area sampling frame. The procedure first presented in Dohrmann et al. (2012) and described further in section 4 can be used not only as a means to assess the quality of the sampling frame, but also to enhance the coverage of the frame.

2.2 Geocoding

For an area sample survey, units need to be located “on the ground” so that they can be grouped into segments for efficient data collection. It also may be desirable to have the ability to associate these segments with Census geographic boundaries so that the associated data (i.e., aggregate Census block-, block group-, and tract-level characteristics) can be used for stratification and in constructing the measure of size. However, the only “geographic” references on the USPS files are the ZIP and ZIP+4 codes associated with the address. Using ZIP or ZIP+4 codes to define segments would not create segments with precise geographic boundaries, however, because the USPS defines these codes simply as collections of delivery points created for mail delivery efficiency. Addresses within the same ZIP code are usually clustered geographically, but not always, especially in rural areas.

¹ E-911 location refers to the ability of 911 emergency service vehicles to locate physical locations based on street addresses.

Geocoding is the process of attaching geographic coordinates to a location. If an address has geographic coordinates attached to it, it can be matched to other geographic entities such as the Census geographic designations. In most geocoding applications, portions of each street, usually called “street segments,” are defined in terms of street number ranges. When an address is geocoded, it is matched to the street segment with the street number range that contains the address, rather than being matched to a specific address in the database. The position of the address is then determined by interpolating within the range along the street segment. With this “street-address geocoding,” the specific geographic coordinates assigned to an address may not be the precise location of the structure at that address, especially when the housing units are not evenly dispersed along a street. The structure may lie some distance down the street or even across the street from the location identified by the geographic coordinates.

When the street name or range containing the street number is not found in the database, two alternative methodologies to the street-address level geocoding process are available. The first assigns central geographic coordinates to the address based on the ZIP+4 associated with the address (e.g., 20850-1118); if the ZIP+4 is not recognized, the second alternative is to assign central geographic coordinates to the address based on the ZIP code associated with the address. The geographic coordinates obtained for an address by using the street segment, ZIP+4 code, or ZIP code can then be used, along with geographic boundary files available from the Census Bureau, to assign the address to a Census block. When geocoding assigns an address to a Census block other than the one it would be assigned to based on its actual physical location, a census-geography geocoding error is said to have occurred. While geocoding to the ZIP+4 or ZIP code centroid would generally be expected to have larger rates of geocoding errors than geocoding to the street segment as described above, employing this hierarchical assignment of geographic coordinates to an address (known as street-level geocoding) ensures that every address is assigned to a Census block.

3. Segment Formation and Selection of Secondary Sampling Units

We describe below an area linkage method that is designed to allow one to use USPS-based lists as ABS frames in nearly all areas. It tackles both geocoding issues and other sources of undercoverage while taking full advantage of the electronic nature of the USPS-based lists. The issue of coverage is addressed by adding a coverage enhancement procedure called Address Coverage Enhancement or ACE²; the issue of inaccurate geocoding is addressed with the use of the geocoding methodology described in section 2.2 to define the segment and the use of a linked area segment for coverage enhancement.

The approach begins with the area segments, i.e., the segments created using Census geographic boundaries. Next, addresses on the ABS frame are linked to area segments using the street-level geocoding method described in section 2. The street-level geocoding ensures that each address on the ABS frame can be associated with one, and only one, area segment. The collection of addresses that geocode to a particular area segment is termed a *list* segment.

The list segments are the secondary sampling units, and are simply collections of addresses with no definitive ordering or geographic reference. Figure 1 provides an illustration of the relationship between an area segment (defined by the blue boundary) and a list segment (grouped for illustration by the light red area) with no distinguishable boundary. Since the area segment and the list segment have a one-to-one linkage, the area segment can be used as the frame of reference for ACE. Thus, we will also refer to the area segment linked to a list segment as a “coverage enhancement area” when it is used for this purpose.

Even though the list segments are the secondary sampling units, segment formation begins in the usual way with the creation of area segments that are collections of Census blocks (or block groups or tracts); data associated with these Census blocks, including population numbers, housing unit counts, and demographic characteristics, may be used in this process. However, since only the addresses that geocode into the area segment will comprise the secondary sampling unit, the associated data may not accurately represent the list segment frame. Yet, these data, even if approximate, may be desirable for use in the stratification and/or calculation of the segment measure of size for probability proportionate to size (PPS) sample selection in addition to the number of addresses that geocode into the area segment.

² In Dohrmann, et al. (2012), this was simply referred to as the coverage enhancement procedure or CEP.

Figure 1: Illustration of Corresponding Area and List Segments



As noted above, the actual physical locations of all the addresses in a list segment may not all fall within the linked area segment. Many if not all of the addresses geocoded at the street-address level would be expected to fall within the area segment boundaries.³ Some portion of the remainder will be geocoded at the ZIP+4 level. Generally, since only a small number of addresses has the same ZIP+4 (e.g., all those on the same side of a city block, or on a floor of an apartment building), geocoding at this level is unlikely to associate a housing unit with an area segment far from its actual location. If it is necessary to geocode at the ZIP code level, a classification that may contain thousands of addresses, the distance between the area segment and the actual location of the housing unit in question could be much larger. While an exact geographic match between the two types of segments would be helpful for data collection, the lack of an exact match presents an efficiency issue, since interviewers may have larger distances between sampled cases than with the usual area segments, not a coverage issue.

The lack of an exact match between the area and list segments also has an implication on the decision to use traditional methods in areas for which the USPS-based lists do not seem to have sufficient coverage. In some PSUs, the coverage rate variation may be such that some list segments contain very few, or zero, addresses. If the decision is to traditionally list these segments, all the listed addresses must be compared to the list segment frames for all the segments in the PSU to ensure the listed addresses are assigned weights that appropriately reflect their probabilities of selection into the sample. For these cases, it may be more cost effective to simply give such cases a higher chance of selection for ACE (see section 3.2) rather than listing multiple segments and comparing them to all the list segment frames in the same PSU.

4. Address Coverage Enhancement (ACE)

The address coverage enhancement procedure, ACE, is comprised of two components: a field procedure and a home-office procedure. The field procedure consists of a canvass of the area segment and the home-office procedure

³ Some of the boundaries of Census blocks were newly created for the 2010 Census, resulting in more Census blocks overall and more Census blocks that contained no residences compared to the 2000 Census. We observed Census-geography geocoding errors in in some very urban area segments when there was a Census block containing no residences adjacent to a street segment. In this situation, some of the addresses associated with the street segment were physically in an adjacent Census block but were incorrectly geocoded into the Census block containing no residences that was adjacent to the street segment.

consists of a quality check of the field work and the identification of any housing units not covered by the list segment frame.

4.1 Field procedure

At the outset of the data collection period, addresses sampled from the list segment for interviewing are transmitted to the assigned data collectors' computers. If a segment is also selected for ACE, all addresses in the list segment (both sampled addresses and those not selected into the sample) are also transmitted to the data collector assigned to work that segment. Generally, the data collector performs the ACE procedure before conducting interviews in the segment; however, the procedure may be performed later as long as there is time remaining in the field period to conduct interviewing in addresses added to the sample as a result of ACE.

Using a specially developed software application, the data collectors access the preloaded list of all addresses in their list segment associated with the area segment chosen for ACE. The data collectors are instructed to travel through their coverage enhancement area segment in a systematic manner. Their task is to determine for each housing unit they encounter on the ground within the boundaries of the area segment whether the address is on the list segment frame. If so, they assign the address a status of "located." If not, they add the address to the list (and the application flags the address as added in the field). To reduce the potential for data entry errors, all the streets in the area segment are preloaded into the application, so the data collector may simply select the street name from a drop-down list rather than typing the name into the application. If the street name is not included in the preloaded list, the data collector adds the street manually. If components of an address are not discernible in the field, for example, if the house number is not visible, the data collector records identifying information associated with the unit and indicates the approximate location of the unit relative to the other nearby units.

For operational reasons, the data collectors are asked to assign a status to all addresses on list segment frame of either located in the area, located outside the area, not located, or located but the data collector was not sure if the unit was within the segment boundaries (see Figure 2). For sampling purposes, however, only the added addresses are of interest. After all addresses in the area are added or given a status, the data collector transmits the full list back to the home office. At this point, the data collector begins or resumes interviewing the sampled cases in the list segment using a separate software application on their computer.

Figure 2: Illustration of Address Statuses



4.2 Determination of non-covered DUs

At the home office, a detailed quality review of each data collector's work is conducted. All added addresses and associated data collector comments are studied and any oddities, such as house/unit numbers appearing out of sequence (based on their time stamps), are investigated. Added addresses are reviewed to confirm they appear to be inside the area segment boundaries, paying special attention to any addresses added to the application with street names that were not preloaded in the application. Additionally any addresses with unknown components are researched and reconciliation attempted using Internet resources including local government GIS websites.

After the home office review, the reconciled addresses added in the field (excluding those with unknown components) are checked against the full ABS frame held by the vendor to determine if the addresses are truly missing from the ABS frame, or if they are present on the ABS frame but geocoded into another list segment and thus given a chance of selection (see Figure 3). While some addresses may be matched to the frame with a simple automated match, others may require more complex processing such as addresses with different secondary unit numbering scheme on the frame than “on the ground” (e.g., apartments designated as A, B, and C versus 1, 2, and 3) or street names that may have aliases (Route 7 versus King St.). Because addresses with unknown components cannot be uniquely matched to addresses on the ABS frame, all these addresses are given a chance of selection.

Following this matching process, added addresses that were found to be missing from the ABS frame and added addresses with unknown components are given a chance of selection. Depending on the probabilities associated with ACE (as discussed in section 5.1) and the numbers of such addresses, either all or a random subsample of these addresses are added to the sample and transmitted to the data collector for interviewing.

Figure 3: Illustration of Addresses in List and Area Segments after Frame Check



5. Sampling Segments and Added Addresses for ACE

5.1 Method

The ACE procedure determines for a subset of segments if there any addresses that would have been identified by traditional listing of the area segment that are not present in the associated list segment. The subset of segments should be selected using a probabilistic approach that allows every segment a chance of being selected for the procedure. Our approach takes into account the fact that USPS-based lists have better coverage in some areas than in others by assigning the probability of area segment selection for the ACE based on expected coverage of the list for the addresses in that area.

There are two sampling operations in ACE: the sampling of segments for assignment to ACE; and the sample of added addresses that cannot be matched to the ABS frame. Segments are selected for the procedure with probability $P(i) = k_i r_i$ where k_i is the under/over sampling factor for ACE segment i , and r_i is the within-segment sampling rate for the ABS frame in segment i . Added address j is selected for the survey in ACE-selected segment i with probability $P(j|i) = 1/(w_j k_i)$ where w_j is the ratio of the sampling weight of a sampled added address in segment i to a sampled listed address in segment i .

If added addresses are to have the same chance of selection as the list-frame addresses, then $w_j = 1$ and it is necessary that $P(j|i)P(i) = r_i$. In this case, $k_i \geq 1$. Setting $k_i = 1$ reduces the number of ACE segments, but requires $r_i = 1$, which is the ratio of the sampling weight for a sampled added address to that of a sampled ABS-frame address.

In rural areas, the number of addresses in the list segment may be very small or even zero. It may be best to select such a segment for ACE with certainty; that is, setting $P(i) = 1$, $P(j|i) = r_i$, and $w_j = 1$. If this is the case for many of

the segments in the same interviewer staffing area - such as, a PSU or a county within a multi-county PSU - it may be more efficient to resort to traditional listing in that area. This is similar to setting $P(i)$, $P(j|i)$, and w_i as above for all sampled segments in area, without a need to distinguish listed addresses from those on the ABS frame.

Having some estimate of area-segment undercoverage, \hat{M}_i , is beneficial, but no definitive threshold is required.

One simple approach is to take the difference between the number of addresses in the list segment and the number of housing units in the associated area segment based on the most recent decennial census. The key issues to consider when using this approach are the age of the decennial census housing unit count, and whether large differences between the number of addresses in the list segment and housing unit count are likely due to Census-geography geocoding error or due to frame noncoverage. Other factors to consider are the number of P.O. Box addresses in the list segment and the potential for large number of group quarters (if the target population includes people living in group quarters). Additional methods of estimating the segment coverage incorporate other characteristics of the area (see Montaquila, Hsu, and Brick 2011 and McMichael et al. 2010 for more details).

If the number of added addresses expected to be found through the procedure is small, it may be best to set $w_i=k_i=1$, and bring all sampled addresses into the sample. If \hat{M}_i is large, then $w_i k_i$ may be set to a value greater than 1 so that the number of actual addresses found that are added to the sample is more manageable. One way to increase $w_i k_i$ is to increase w_i , which increases the weight of sampled added addresses. Another way to increase $w_i k_i$ is to increase k_i , which increases the probability that segment i will be selected for the ACE, since $P(i)=k_i r_i$.

Though in the above discussion, we first specified an expression for $P(j|i)$ and then one for $P(i)$, the planning for ACE first determines the values of $P(i)$ for all the sampled segments. One approach for the planning of ACE is to first determine a maximum number of addresses missing from the ABS frame that can be added to the sample in a segment, say m' , based on the resources available in the field. If $\hat{M}_i > m'$, set $k_i = \hat{M}_i / m'$. If the estimate \hat{M}_i proves to be much smaller than the number of missed addresses actually found through ACE, then the added addresses can be subsampled by making $w_i > 1$, which will increase the weight of a segment's sampled added addresses relative to its sampled listed addresses.

5.2 Practical application

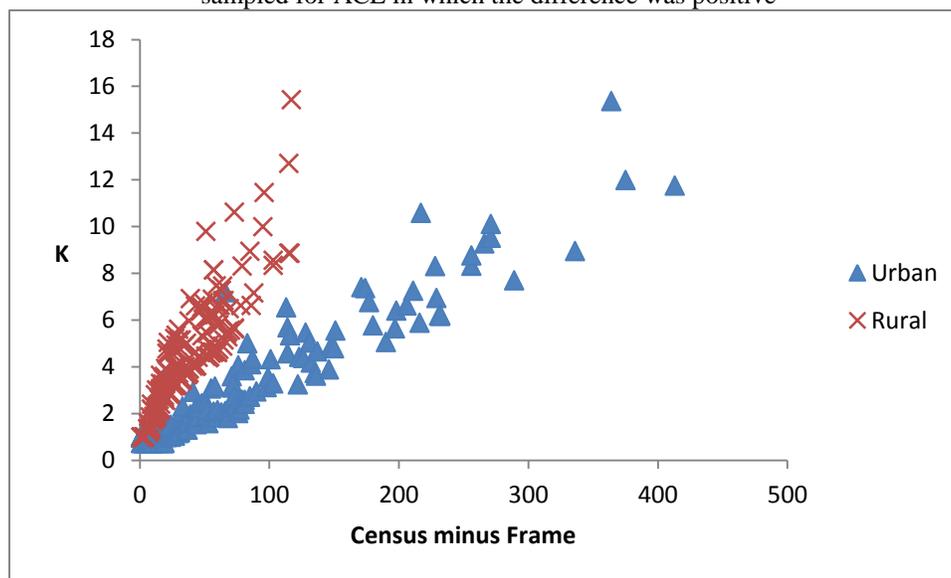
We recently implemented the ACE procedure in a national survey conducted from the spring of 2012 through the summer of 2013. The survey had approximately 7,000 sampled segments partitioned into four approximately equal-sized waves. The ACE for the first wave was conducted in the spring of 2012. The ACE for the fourth wave was conducted in spring of 2013. When planning the four ACE's, we compared the ABS list frame count to 2010 Census housing unit counts. Based on this comparison, we decided to perform traditional listing in 118 segments in three counties. Two of these counties were each a one-county PSU. The third county was a rural county in a two-county PSU. These segments were either in rural areas or on or near Indian reservations where city-style addresses were not used to deliver mail.

The segments where traditional listing was not performed were subjected to sampling for ACE using the formula $P(i)=k_i r_i$. For the first and second waves, when the number of ABS list frame addresses in a segment was greater or equal to the 2010 Census housing unit count we used $k_i=0.5$, which created an expected weight ratio of 2.0 for the sampled added addresses in these segments. After reviewing the results of the first wave, we felt that the design effect incurred from this weight ratio was too great, and k_i was increased to 0.75 for the third and fourth waves. This increase in k_i also increased the number of segments sampled for ACE in this category of segments, which because of resource constraints required that we correspondingly decrease the number of segments sampled for ACE in a different category of segments.

When the number of ABS list frame addresses in a segment was less than the 2010 Census housing unit count, we used $k_i > 1$, except for one category of such segments where we decreased k_i from 1.0 to 0.75 between the second and third waves to compensate for the increase in the number of segments sampled for ACE described in the preceding paragraph. The more the 2010 Census housing unit count exceeded the number of ABS list frame addresses in a segment, the larger k_i was. Figure 4 is plot of k_i versus the difference between 2010 Census counts and ABS list frame in the segments sampled for ACE in which this difference was positive. Different symbols are used to plot

urban and rural segments, where urban segments are defined to be those in a county that is part of a metropolitan statistical area with a county population of at least 250,000 and rural segments are those in all the other counties. Different k_i values as a function of the difference in segment-level counts between the 2010 Census and the ABS list frame were used for urban and rural segments because it was our belief that urban differences are primarily due to Census-geography geocoding errors whereas rural differences are primarily due to ABS undercoverage. From the 7,082 segments subjected to sampling for ACE, 691 were selected.

Figure 4: k_i versus segment differences between 2010 Census count and ABS list frame count for the segments sampled for ACE in which the difference was positive



6. Summary

This paper describes a methodology for using ABS for household sampling that allows full ABS implementation and includes a coverage enhancement procedure that permits representation of all addresses while limiting the resource requirements. Since the address coverage enhancement procedure (ACE) can be implemented immediately before (or concurrent with) data collection, using the same hardware as used for interviewing, data collectors have the flexibility of working initial sample cases while the addresses added through coverage enhancement are reviewed by the home office.

This methodology addresses both the coverage and geocoding issues previously cited as obstacles to using ABS for area surveys with in-person interviewing. The coverage issue is addressed with the inclusion of ACE applied more frequently in areas estimated to have poor coverage and a home-office procedure with strict quality control measures to ensure that all units added to the frame previously did not have a chance of selection. Geocoding inaccuracy, while potentially complicating ACE, does not increase noncoverage. While the geocoding may not place every address in the correct area segment as defined by Census geographic boundaries, every address on an ABS frame is assigned to a list segment that is defined as the collection of addresses that geocode to a specific area segment.

References

- Dohrmann, S., Han, D., and Mohadjer, L. (2006). Residential Address Lists vs. Traditional Listing: Enumerating Households and Group Quarters. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2959-2964.
- Dohrmann, S., Montaquila, J., Good, C., and Berlin, M. (2012). Using address-based sampling frames in lieu of traditional listing: A new approach. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 3729-3741.

- McMichael, J., Ridenhour, J., Shook-Sa, B., Morton, K., and Iannacchione, V. (2010). Predicting the Coverage of Address-Based Sampling Frames Prior to Sample Selection. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 4852-4859.
- Montaquila, J., Hsu, V., and Brick, M. (2011). "Using a 'Match Rate' Model to Predict Areas where USPS-based Address Lists May be Used in Place of Traditional Listing." *Public Opinion Quarterly* 75(2):314-335.
- Montaquila, J., Hsu, V., and Brick, M., English, N., and O'Muircheartaigh, C. (2009). "A Comparative Evaluation of Traditional Listing vs. Address-Based Sampling Frames: Matching with Field Investigation of Discrepancies." *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 4855-62.
- O'Muircheartaigh, C., Eckman, S., and Weiss, C. (2002). Traditional and Enhanced Field Listing for Probability Sampling. *Proceedings of the Social Statistics Section of the American Statistical Association*, 2563-2567.
- Staab, J.M., and Iannacchione, V.G. (2003). Evaluating the Use of Residential Mailing Addresses in a National Household Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 4028-4033.