On Moving to the Use of a Hybrid Sampling Frame in the National Survey on Drug Use and Health: Motivations and Challenges

Arthur Hughes[1], Jonaki Bose[1], Bonnie Shook-Sa[2], and Katherine Morton[2]

1. Substance Abuse and Mental Health Services Administration, 1 Choke Cherry Road Rockville, MD 20857
2. RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract

The sampling frame currently used in the National Survey on Drug Use and Health (NSDUH) is constructed by Listers who visit each selected area segment and conduct a complete enumeration of dwelling units that are eligible for selection. The half-open interval (HOI) procedure is used by Field Interviewers during the data collection phase to identify and include dwelling units that were missed during the listing operation. The use of Address-Based sampling (ABS) is increasingly becoming a viable alternative to traditional approaches for developing sampling frames and has the potential to significantly reduce costs associated with the listing operation. ABS relies solely on the use of a computerized listing of addresses (maintained and updated by the U.S. Postal Service) instead of field enumerated listings to select samples of dwelling units.

Research was conducted recently to evaluate the coverage, cost, and implementation of a hybrid sampling frame for NSDUH, where ABS would be employed in most area segments while the traditional listing method would continue in areas where ABS coverage is low. A new Check for Housing Units Missed (CHUM) procedure was proposed to identify missed dwelling units to serve as a replacement for the HOI. This conference paper will summarize the current NSDUH sample design including the sampling frame procedure, present a potential alternative hybrid sampling frame based on prior research, and discuss motivations, challenges, and current decisions associated with implementing a hybrid frame in NSDUH.

Keywords: Address-Based Sampling, hybrid frame, coverage, in-person surveys

Introduction

Federal household surveys currently face a number of significant challenges including declining response rates and increasing costs of data collection (National Research Council, 2011). To combat increasing costs, the use of Address-Based Sampling (ABS) was investigated as an alternative to the use of traditional Field Enumeration (FE) methods in National Survey on Drug Use and Health (NSDUH) sampled area segments where the coverage of the target population was high. The use of FE in area segments would continue to be used in areas where coverage was low, all leading to the proposed construction of a hybrid sampling frame for NSDUH. Ideally, a hybrid frame consisting of a combination of ABS and FE segments would reduce survey costs compared to an FE-only frame without any loss in coverage of the target population.

2005-2013 NSDUH Sample Design

The National Survey on Drug Use and Health (NSDUH) is a primary source of statistical information on use of illicit drugs, alcohol, tobacco, and mental health problems among the U.S. civilian non-institutional population aged 12 or older. It is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), U.S. Department of Health and Human Services, and is planned and managed by SAMHSA's Center for Behavioral Health Statistics and Quality (CBHSQ). Data collection

and analysis are conducted under contract with RTI International. NSDUH is an ongoing survey that administers a variety of questions on substance use and associated behaviors in a face-to-face setting via computer-assisted interviewing at the respondents' place of residence.

The 2005-2013 NSDUH sample design and data collection was carried out as follows:

1. stratify each State and the District of Columbia into a specified number of regions of approximately equal population,
2. select census tracts within regions,
3. select area segments (consisting of a collection of adjacent census blocks) within selected tracts,
4. field enumerate all dwelling units (DUs) within sampled area segments,
5. select dwelling units (DUs) within sampled area segments,
6. screen each sampled DU in person by collecting roster data and selecting 0, 1, or 2 persons, and
7. conduct an in-person interview with each selected person.

The within-state stratification and selection of census tracts and area segments in steps 1, 2 and 3 were completed well in advance of data collection using data available electronically from the U.S. Census Bureau and other sources. Each segment in step 3 consists of a set of adjacent census blocks that, in combination, satisfy the minimum requirement of 100 DUs in rural areas or 150 DUs in urban areas. This minimum number was determined to ensure that there would be a sufficient number of eligible DUs and subsequent completed interviews. Field Enumeration (FE) of DUs is completed by NSDUH field staff called "Listers" who physically travel to each sampled segment to manually count and list all DUs. This information is hand written on a segment listing form and provided to statisticians who select a sample of DUs.

Field interviewers visit sampled DUs in person where they use their hand-held computers to select 0, 1, or 2 persons for the NSDUH interview. Before the interviewer leaves the sampled DU, she/he asks the screening respondent the following question:

Are there any other living quarters within this structure or on this property, such as a separate apartment with a separate entrance?

Any such living quarters identified by the screening respondent are added to the sample. In addition, the half-open interval (HOI) frame-linking procedure (Kish, 1965) is used to identify and add units that may not have been included on the frame by the Lister. Currently, NSDUH Interviewers are trained on using area segment maps and lists, and following the correct path of travel so they can implement HOI while in the field conducting screening and interviewing. The HOI states that if a dwelling unit is selected and an interviewer observes any new or missed dwelling units between the sample dwelling unit (SDU) and the dwelling unit appearing right after the SDU on the segment listing form, then all new or missed dwelling units between the SDU and the next dwelling unit on the listing form will be selected. The HOI is a standard procedure used by survey researchers and practitioners to reduce bias in survey estimates introduced by list frame undercoverage.

More details about the 2005-2013 sample design can be found in the NSDUH Sample Design Report (Morton et al, 2012).

2014 NSDUH Sample Design

Faced with the need to update and improve the NSDUH survey on a variety of fronts, a new sample design was implemented beginning in 2014 followed by a redesigned questionnaire in 2015. Details about

the 2014 design are provided in the 2014 NSDUH Office of Management and Budget (OMB) clearance materials (OMB, 2013). The new sample design includes the use of larger cluster sizes in some States and changes to the sample allocation by State and age leading to improved precision of most national estimates. Beginning in 2014 the HOI procedure was dropped. Dropping HOI was justified because a very small percentage of dwelling units are picked up by this method. In fact, of all added DUs in the fourth quarter of 2010 (a total of 381 added DUs), 272 (71.4 percent) were added as a result of the interviewer asking the screening respondent about additional units within or on the property of the SDU. The remaining 109 added DUs were found within the HOI and were usually either new construction or missed during the FE process. This implies that identifying missed units based solely on HOI accounted for only 0.2 percent of all 48,568 SDUs selected for the quarter. There are advantages to eliminating the HOI from the NSDUH. For example, FIs would no longer need to be trained on the path of travel so that portion of the training can be eliminated. It would also simplify the preparation of the field materials because there would be little need to generate a list of SDUs along with the next listed dwelling unit. However, FIs would still need to carry their segment maps and lists with them while in the field during screening and interviewing as they would be needed to locate the segments and dwelling units.

To accommodate the potential use of ABS in future NSDUHs (i.e., after 2014), the 2014 design includes the selection of census block groups after the selection of census tracts. This additional stage of selection was included in order to reduce geocoding error in ABS segments since ABS segments would be geocoded for sampling purposes and to help interviewers locate the DU sample. Results from a NSDUH mailing list study field test showed that 89.9 percent of addresses from ABS frame geocoded into the correct area segment. However, 99.3 percent of those same addresses geocoded into the correct census block group. Compared to geocoding at the area census block level, geocoding accuracy improves significantly at the census block group level in both rural and urban areas. Thus, in an ABS design, it was determined that census block groups would serve as geographic clusters in areas with sufficient mailing address coverage. Since the use of census block groups as ABS segments will result in larger land areas, the within segment intraclass correlation is expected to decrease.

Address-Based Sampling

ABS in the household survey setting involves the use of mailing lists obtained from a database with (ideally) complete coverage of residential addresses that are eligible for selection. The primary source of these mailing lists is the Computerized Delivery Sequence (CDS) file maintained by the U.S. Postal Service (USPS, 2011) and made available to the public through licensing agreements with qualified vendors. The CDS file includes locatable city-style addresses consisting of a street number and name, city, state, and nine-digit ZIP Code for approximately 96 percent of unique residential addresses[1] on the file (Iannacchione 2011). No names are associated with this file. Examples of survey organizations that use the CDS are: the U.S. Census Bureau for updating their Master Address File which in turn is used as a sampling frame in a number of household surveys including the American Community Survey (U.S. Census Bureau, 2011) and the National Center for Health Statistics which use the CDS to update the sampling frame for the National Survey of Family Growth (Lepkowski et al, 2010).

Similar to conducting HOI on FE area segments, procedures to identify and add missed units in ABS segments may be needed. Such procedures are provided below.

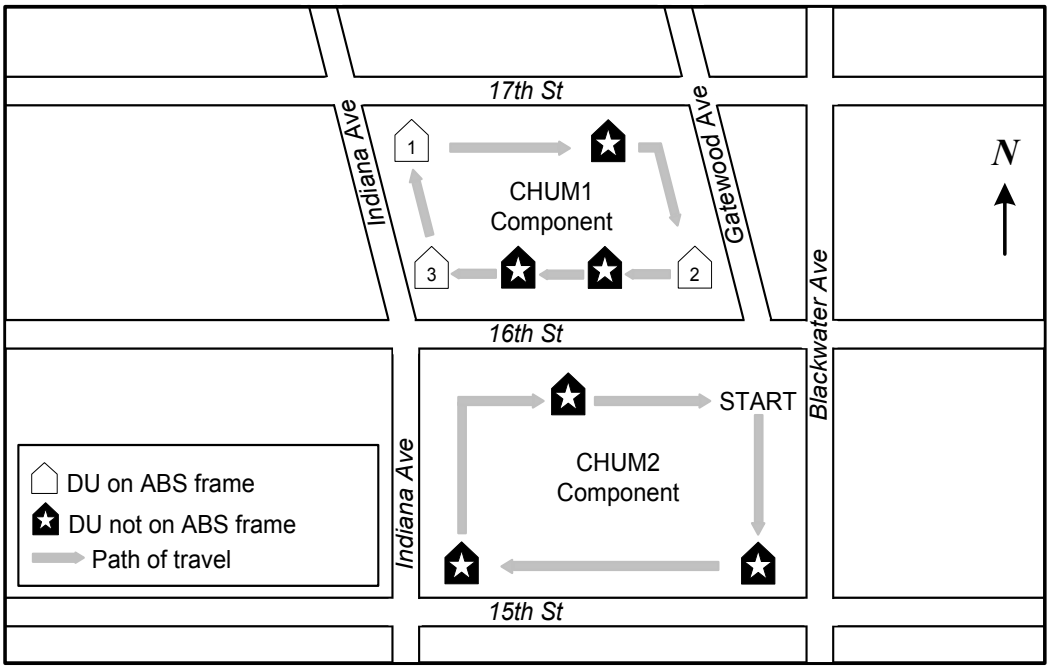
[1] This estimate excludes PO Box addresses that are not the residents' only way to receive mail.

Identification of Missed Addresses in ABS

Cost savings are the primary advantage of using ABS over FE for in-person surveys of the civilian, non-institutionalized population. However, prior research indicates that ABS coverage of the household population is incomplete, especially in rural areas (Iannacchione et al. 2010). A hybrid sampling frame was developed for the NSDUH with the goal of lowering costs without sacrificing coverage levels of the current FE sampling frame. Using an expected ABS coverage rate, segments with a coverage rate at or above a predetermined coverage threshold will use ABS supplemented with the Check for Housing Units Missed (CHUM) frame-linking procedure to identify and record DUs that are not included on the ABS list (McMichael et al. 2008). All other segments will be scheduled for FE supplemented with the half-open interval (HOI) frame-linking procedure (Kish 1965).

The CHUM procedure is designed to be used in segments where ABS coverage is adequate but not complete. As Figure 1 demonstrates, the CHUM procedure uses two components to supplement the coverage of an ABS frame. The CHUM1 component enables FIs to establish a *path of travel* from the sampled DU to the "next DU." This methodology requires segments with discernible boundaries so that DUs can be identified as being within a sampled segment. Facing the located sampled DU, the FI travels clockwise around the block, without crossing a street to find the next DU. Street crossings are avoided to ensure that each path of travel is non-overlapping.

Figure 1: Examples of CHUM Components



CHUM1 Component: Path of travel moves clockwise from the selected DU. If DU 1 is selected, one missed DU would be added to the ABS frame. If DU 2 is selected, two missed DUs would be added. No missed DUs would be added if DU 3 is selected. **CHUM2 Component**: First, a block or area is randomly selected within a segment. Then, FIs are provided a segment map that enables the entire block to be field enumerated. In this case, three missed DUs are added to the frame. Notice that the three missed DUs have no chance for inclusion solely using CHUM1.

After the address of the next DU is found, it is checked against the ABS frame to determine whether it was missed. This check requires access to the entire ABS frame within each selected segment. If the address of the next DU is not on the ABS frame, the DU is sampled and assigned a probability of selection that is linked to the probability of selecting the sampled DU. If multiple missed DUs are picked up by the CHUM1, then subsampling may be used to control the sample size. Otherwise, a missed DU will be included in the sample and assigned the same design weight as the sampled DU. This step is repeated until either the address of a DU on the ABS frame is found or the block is circumnavigated.

Because CHUM1 is restricted to blocks associated with a sampled address, DUs in blocks with no addresses on the ABS frame will be missed. CHUM2 mitigates this source of undercoverage by adding the "missed blocks" and their associated DUs to the frame. For segments that contain multiple missed blocks, efficient sampling methods can be used to select a subsample of missed DUs.

In addition to employing the CDS file and implementing a missed DU procedure such as the CHUM, the U.S. Postal Service also provides a supplemental address file that has the potential for increasing coverage. It is called the No-Stat file and is discussed next.

Expanding the Coverage of ABS with the No-Stat File

Address-based sampling frames are traditionally based on the USPS CDS file, which USPS makes available to qualified vendors through nonexclusive licensing agreements. Since 2009, USPS has also made available the No-Stat file, which supplements the CDS file with approximately seven million predominantly rural addresses not found on the CDS file.

The No-Stat file is a supplemental file also maintained by the U.S. Postal Service that contains addresses that are classified as inactive or that are not currently receiving mail delivery at their household. Examples of these addresses are new housing developments, rural delivery points that have been vacant for 90 days or longer and gated communities that do not receive delivery by USPS personnel to individual units (USPS, 2011).

An evaluation of the No-Stat file based on 2010 NSDUH data found that it includes the addresses of approximately 1.3 million occupied housing units and provides household coverage gains of 1.2 percent overall and 3.8 percent in rural areas. While these coverage gains are modest, including the No-Stat file is a cost effective strategy for boosting rural coverage because its inclusion does not significantly reduce the efficiency of the ABS frame (Shook-Sa et al. in press).

Hybrid Sampling Frame

Iannacchione et al. (2010) evaluated the feasibility of implementing a hybrid sampling frame based on a probability sample of 200 segments and 3,878 eligible dwelling units (DUs) from the NSDUH. Here, a hybrid frame is defined as one where area segments are classified into FE and ABS to maximize coverage while reducing costs and interviewer burden. Various ABS coverage thresholds were evaluated in terms of cost savings and coverage. As the ABS coverage threshold decreases, fewer segments rely on FE and, therefore, the estimated cost savings increases (Table 1). As shown in Table 2, there is not much difference in DU coverage in urban segments. However, there is a mild increase in coverage of rural DUs at higher ABS coverage thresholds. Theoretically, the hybrid sampling frame should provide 100% coverage of the target population. However, as with any frame supplementation method, a loss of coverage occurs when field interviewers neglect to follow the correct protocols when implementing the CHUM procedures.

Table 1: Estimated Annual Cost Savings[1] of the Hybrid Frame

ABS Coverage Threshold[2]	FE Segments			ABS Segments		Estimated Cost Savings[3]
	#	%		#	%	
20%	384	5.5%		6,624	94.5%	62.3%
50%	1,076	15.4%		5,932	84.6%	55.4%
80%	2,565	36.6%		4,443	63.4%	40.1%
FE Only	7,008	100.0%		0	0.0%	0.0%

ABS = address-based sampling; FE = field enumeration.

[1] Cost savings include costs incurred after the selection of the segment and before the start of field interviewing. Costs include the half-open interval (HOI) for FE segments and the Check for Housing Units Missed (CHUM) for ABS segments. Excluded are the costs associated with the 192 NSDUH segments in Alaska and Hawaii.

[2] Ratio of the number of city-style mailing addresses in a segment to the Census Bureau/Claritas dwelling unit (DU) estimate.

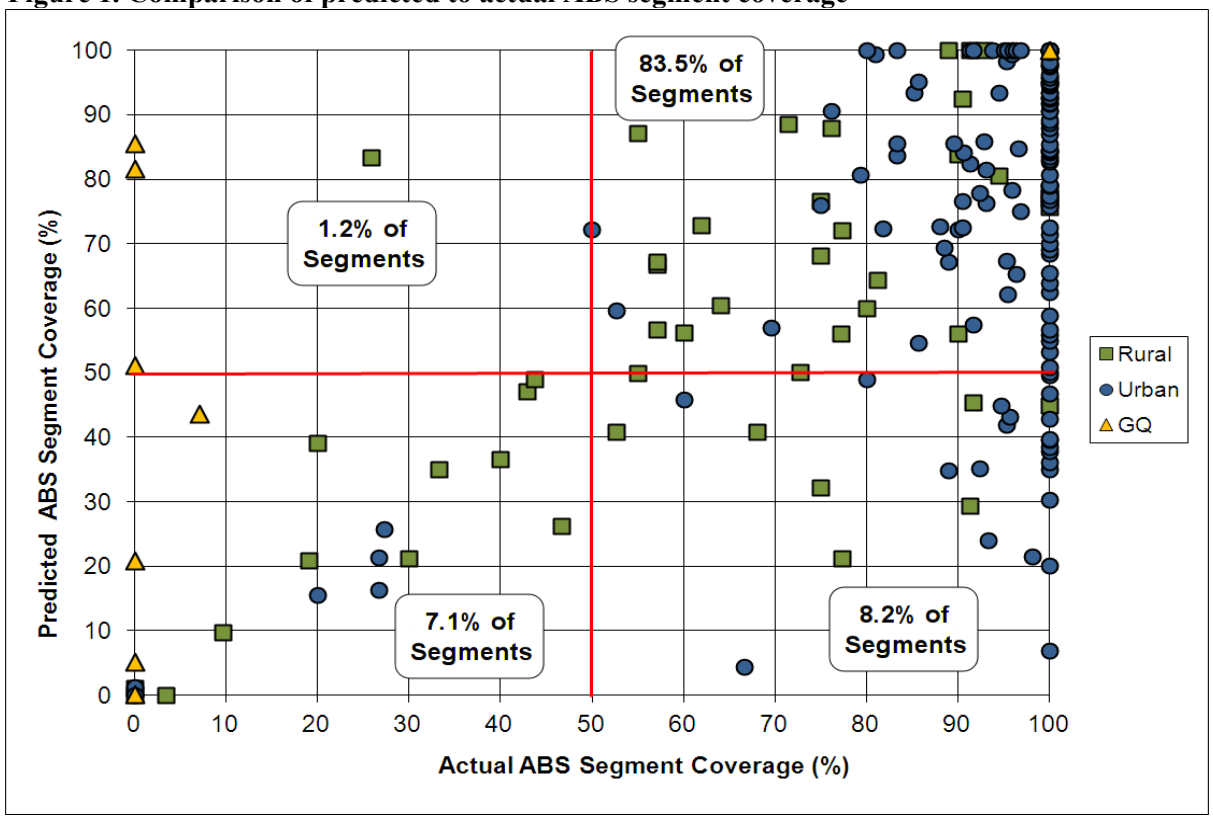
[3] Compared with doing FE in all segments.

Table 2: Estimated Coverage of NSDUH Dwelling Units of the Hybrid Frame

	ABS Coverage Threshold		
	20%	**50%**	**80%**
Urban Segments	98.9%	99.0%	99.1%
Rural Segments	92.3%	93.4%	95.6%
Overall	**97.5%**	**97.8%**	**98.4%**

How closely the estimated ABS coverage comes to predicting the actual ABS coverage greatly influences cost and coverage. By accurately predicting ABS coverage, low coverage segments will always be field enumerated, and fewer segments with marginal ABS coverage will rely on the CHUM. Therefore, accurately predicting ABS coverage at the area-segment level prior to sample selection simultaneously lowers costs and improves frame coverage. Using the field study sample of 200 NSDUH segments, McMichael et al. (2010) compared predicted ABS coverage to actual ABS coverage. Using a coverage threshold of 50%, eighty-four percent of NSDUH segments were correctly allocated to the ABS frame and only one percent was incorrectly assigned to ABS. Similarly, seven percent of NSDUH segments were correctly allocated to FE but eight percent had sufficient ABS coverage and were incorrectly allocated to FE. Several of the segments that were incorrectly allocated to ABS when they should have been field enumerated had a high proportion of group quarters. Thus, information about the presence of group quarters in segments can be used to more accurately predict segment coverage and allocation to ABS or FE. (Figure 1)

Figure 1. Comparison of predicted to actual ABS segment coverage



Some of the error in estimated ABS coverage is attributable to geocoding error. Geocoding here is the process by which an address is assigned latitude and longitude coordinates and thereby allocated to an area segment. Two types of geocoding error can occur at the segment level. Under-coverage geocoding error occurs when addresses are misallocated out of the selected segment. Addresses that are present in the segment are not included on the segment's mailing list. Over-coverage geocoding error occurs when addresses are misallocated into the selected segment. Addresses that are not present within the segment boundaries are included on the segment's mailing list. Shook-Sa et al. (2010) examined both types of geocoding error using mailing addresses and field enumerated addresses from the 200 field study segments. Geocoding accuracy at the segment level was found to be quite poor and varied significantly by urbanicity (23.4 percent under-coverage for rural segments; 7.5 percent for urban segments). Geocoding accuracy improves significantly at the census block group level for both rural and urban segments, with 99.3 percent of addresses geocoding into the correct census block group (99.8 percent urban, 96.5 percent rural). These findings should be considered when designing ABS studies that are based on census geography.

The concern with under-covering the target population is that it will introduce bias in the survey estimates. Within the probability sample of 200 NSDUH segments, Morton et al. (2010) compared prevalence estimates for a number of key NSDUH outcomes based on the current FE frame with those based on the hybrid frame. None of the differences were substantive at the 20 percent ABS coverage threshold. Because the estimates based on the NSDUH frame and the estimates based on the hybrid frame share a large portion of their cases, these comparisons have the statistical power to declare very small differences in the overall prevalence estimates statistically significant (e.g., 0.0002).

Challenges, Current Decision, and Future Directions

SAMHSA has decided not to pursue the use of a hybrid sampling frame in NSUDH at this time for the following reasons:

Potential Increased Field Interviewer Burden - In the NSDUH mailing list field test study described earlier, interviewers reported difficulty performing the CHUM procedure at apartments and trailer parks located in ABS assigned segments. Part of the reason was that, in some situations, the interviewer didn't have a good grasp of how addresses were to be sorted so that the next dwelling unit can be identified and compared to the addresses on the ABS list.

Different Procedures Required to Identify Missed Dwelling Units in ABS and FE Area Segments – By going with a hybrid frame, interviewers assigned to both FE and ABS area segments will need to learn and retain two different procedures for checking for missed units (where one is the CHUM in ABS segments and the other is either the HOI or no procedure in the FE segments). One possibility that was discussed in Iannacchione (2012) was the use of Global Positioning System (GPS) devices to capture the latitude/longitude coordinates along with the street name or description of each dwelling unit in FE segments. In ABS segments, latitude/longitude coordinates would be purchased along with the mailing addresses. Having common dwelling unit location information in the form of GPS coordinates can lead to the development of a single procedure for identifying missed dwelling units in both segment types.

Limited ABS Coverage in Rural Areas - As previously mentioned, the use of supplemental files such as the CDS No-Stat file from USPS results in only modest gains in the coverage in rural areas. Thus, the issue remains that ABS may not be cost effective in rural areas. However, this limitation justifies the development and use of a hybrid frame where FE is used in area segments where the ABS coverage is low.

Limited Coverage of the Group Quarters Population - There is some uncertainly about where to obtain an accurate source that can be used to identify segments with high concentrations of group quarters units such as college dormitories. In Iannacchione (2012), it was determined that the only feasible source for identifying segments with high concentrations of group quarters is the 2010 decennial Census. There is some concern about the quality of this data over time (section 5.2 of Iannacchione, 2012).

It is for the reasons stated above that the current plan is to continue to use field enumeration (FE) for all selected segments. While the use of ABS will not be incorporated into the new 2014 NSDUH sample design, it is still under consideration for future years. Additional field studies may be conducted to further test uniform procedures for identifying and adding missed dwelling units in both ABS and FE area segments. Doing so has the potential of simplifying the interviewer's workload and reducing errors. The use of GPS devices to check for missed units in both area segment types in a consistent and uniform manner is an approach also worthy of testing.

References

Iannacchione, V. G., Morton, K., McMichael, J., Shook-Sa, B., Ridenhour, J., Stolzenberg, S., Bergeron, D., Chromy, J., & Hughes, A. (2010). The best of both worlds: A sampling frame based on address-based sampling and field enumeration. *Proceedings of the American Statistical Association, Section on Survey Research Methods* 3153–3165.
[Available as a PDF at http://www.amstat.org/sections/SRMS/proceedings/]

Iannacchione, Vincent. (2011). The Changing Role of Address-Based Sampling in Survey Research. *Public Opinion Quarterly* 75:556-575.

Iannacchione, V.G., McMichael, J.P., Shook-Sa, B.E., and Morton, K. B. (2012). *A proposed hybrid sampling frame for the national survey on drug use and health*, (prepared for the Substance Abuse and Mental Health Services Administration under Contract No. 283-2004-00022). Research Triangle Park, NC. [Available as a PDF at http://www.samhsa.gov/data/NSDUH/NSDUHMethodsRptHybrid2012.pdf]

Kish, L. (1965). *Survey Sampling* (p. 56). New York: John Wiley & Sons.

Lepkowski J.M., Mosher W.D., Davis K.E., Groves R.M., Van Hoewyk J. (2010). The 2006–2010 National Survey of Family Growth: Sample design and analysis of a continuous survey. National Center for Health Statistics. Vital Health Stat 2(150). Retrieved February 7, 2014, from http://www.cdc.gov/nchs/data/series/sr_02/sr02_150.pdf

McMichael, J.P., J.L. Ridenhour, and B.E. Shook-Sa (2008). A robust procedure to supplement the coverage of address-based sampling frames for household surveys. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association. 4329-4335. [Available as a PDF at http://www.amstat.org/sections/SRMS/proceedings/]

McMichael, J. P., Ridenhour, J. L., Shook-Sa, B. E., Morton, K. B., & Iannacchione, V. G. (2010). Predicting the coverage of address-based sampling frames prior to sample selection. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association. 4852–4859. [Available as a PDF at http://www.amstat.org/sections/SRMS/proceedings/]

Morton, K.B., Martin, P.C., Shook-Sa, B.E., Chromy, J.R., and Hirsch, E.L. (2012). *2011 National Survey on Drug Use and Health Sample Design Report*, (Prepared for the 2011 Methodological Resource Book for the Substance Abuse and Mental Health Services Administration under Contract No. HHSS283200800004C, Deliverable 8). Research Triangle Park, NC. [Available as a PDF at http://www.samhsa.gov/data/2k12/NSDUH2011MRB/NSDUHmrbSampleDesign2011.pdf]

Morton, K. B., McMichael, J. P., Ridenhour, J. L., & Bose, J. (2010). Address-based sampling and the National Survey of Drug Use and Health: Evaluating the effects of coverage bias. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association. 4902–4907. [Available as a PDF at http://www.amstat.org/sections/SRMS/proceedings/]

National Research Council. (2011). *The Future of Federal Household Surveys: Summary of a Workshop.* K. Marton and J.C. Karberg, rapporteurs. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Office of Management and Budget (2013). *2014 National Survey on Drug Use and Health (NSDUH)* Information Collection Request. Retrieved February 7, 2014, from http://www.reginfo.gov/public/do/PRAViewICR?ref_nbr=201307-0930-002

Shook-Sa, B.E., J.P. McMichael, J.L. Ridenhour, and V.G. Iannacchione (2010). The implications of geocoding error on address-based sampling. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association.
[Available as a PDF at http://www.amstat.org/sections/SRMS/proceedings/]

Shook-Sa, B. E., Currivan, D. B., McMichael, J. P., & Iannacchione, V. G. (2013). Extending the coverage of ABS frames: Beyond the USPS Computerized Delivery Sequence File. *Public Opinion Quarterly* 77: 994-1005.

U.S. Census Bureau (2011). 2010 Census Demographic Profile Summary File Technical Documentation, *2010 Census of Population and Housing,* prepared by the U.S. Census Bureau, 2011. Retrieved February 7, 2014, from http://www.census.gov/prod/cen2010/doc/dpsf.pdf

United States Postal Service. (2013). *CDS user guide*. Retrieved February 7, 2014, from http://ribbs.usps.gov/cds/documents/tech_guides/CDS_USER_GUIDE.PDF