Qualities of Coverage: Who is Included or Excluded by Definitions of Frame Composition?

Ned English[1], Colm O'Muircheartaigh[2],
Katie Dekker[1], Lee Fiorio[1]
[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603
[2]Harris School of Public Policy Studies at the University of Chicago, 1155 E. 60th Street, Chicago, IL 60637

Abstract

Motivated by the high costs associated with traditional listing, survey organizations have looked to the United States Postal Service Delivery Sequence File (DSF) as a replacement for traditional listing, particularly in urban areas. However, several different vendors offer DSF-derived datasets and we do not know how they compare with respect to coverage. Working with the National Children's Study (NCS), we continue our work to understand the coverage properties of these databases. After matching traditionally-listed addresses with three versions of the DSF, we returned to housing units excluded from one or more source to collect additional data. We use GPS-enabled photography and NCS screening and interview data to uncover "patterns of missingness" describing the types of housing units and households that may be included or excluded by traditional or DSF-based lists.

Key Words: Address-based samples, delivery sequence file, National Children's Study

1. Introduction

The National Children's Study (NCS) is a large new in-person panel survey which attempts to understand environmental impacts on child development (Montaquila et al. 2009, Montaquila et al. 2010). The NCS will enrol a panel of 100,000 children before birth and follow them until age 21 for health and environmental testing. The NCS will be one of the largest and most complex surveys ever undertaken, and so the impact of frame construction and sample design decisions are manifold (Michael and O'Muircheartaigh 2007, Montaquila et al. 2010).

The NCS sample design is based on a housing unit frame generated by traditional listing in selected segments. "Traditional listing" is a method of address frame generation created by field staff known as "listers". Listers record all residential addresses in defined geographies in a systematic manner whether they are occupied or not (Kish 1965, Eckman 2010). This method of frame creation has been considered the "gold standard" in the survey research industry since the early days of in-person studies (O'Muircheartaigh et al. 2006, O'Muircheartaigh et al. 2007).

However, cost and time concerns have lead survey researchers to consider alternatives to traditional housing unit listing. In recent years it has become possible to license extracts of the United States Postal Service Delivery Sequence File (or DSF), a list of all housing units in the United States that receive mail (Kennel and Li 2009). Survey research organizations have been researching the use of the DSF as a replacement for traditional listing in recent years (Iannacchione et al. 2003, O'Muircheartaigh et al. 2007, Battaglia et al. 2008, Link et al. 2008, Montaquila et al. 2009). One of the first uses of the DSF was in Dallas County, TX, and suggested that coverage would be adequate for an urban sample (Iannacchione et al. 2003, Staab and Iannacchione, 2003). NORC then began an assessment of the coverage properties of DSF-derived frames with an evaluation in a subset of segments across a number of projects (O'Muircheartaigh, Eckman, and Weiss 2003, O'Muircheartaigh et al. 2007). Other organizations have since undertaken their own research into where it is appropriate to use the DSF, where it is necessary to list, and where it may be possible to employ a hybrid approach (Montaquila et al 2010).

The current thinking is that the DSF performs at least comparably to traditional listings in urban and suburban areas, especially those with regular block-patterns and relative housing stability. Rural areas, however, are known to contain a larger share of non-geocodable addresses, including PO and rural route box addresses. Because survey research organizations are usually interested in targeting small areas, non-geocodable addresses are usually considered undercoverage for in-person studies. Consequently, the coverage of the DSF in rural areas is not yet adequate for in-person surveys which require a housing unit address for sampling purposes (Eckman et al. 2010).

The National Children's Study (NCS) as originally designed, intended to use traditional listing to create an in-person frame in all segments (Montaquila et al. 2009). Because the NCS sample was nationally-representative, many segments included places where previous research has shown the DSF to be comparable or superior. A recent direct evaluation of the NCS sampling frame in the Waukesha County, WI site demonstrated the potential for using the DSF in mixed areas, with a high rate of overlap between DSF coverage and "reality" (English et al. 2009). Our research did show a somewhat larger share of housing units on the traditionally-listed frame than on DSF-based frames in both rural and urban areas (English et al. 2009). In addition, a small percentage of housing units were missing from the DSF or from traditional listings.

Of course, what is important in terms of survey quality is not the relative coverage rates of each method, but a better understanding of "who" is missing from either list. Are the households at risk of undercoverage via one method or another eligible for the NCS? Would including them in the survey frame change critical estimates? Because the NCS has attempted in-person interviews with all households in the frame, it presented an opportunity to compare the properties of those that tend to be captured by traditional listing with those DSF addresses through direct evaluation.

Our current research continues to explore key aspects of frame construction and coverage, with implications for in-person surveys beyond the National Children's Study. Our primary goal has been to determine if traditional listing is the ideal method for National Children's Study frame construction by comparing the collected listings to those from the USPS DSF. In so doing we extend previous research that focused on the overlap between traditional and DSF-derived lists in common areas. We use three sources of the DSF in the current research, from the Valassis (formerly ADVO), MSG, and CIS vendors. We examine the categories of housing units that are missing from given lists, and thus the types of households that would be expected to be under or over-covered if only one method were used. We compare eligibility between lists (traditional and DSF) at the housing unit and household level.

2. Methods

Our current evaluation and analysis was based on the Waukesha, WI National Children's Study site, contracted to NORC. Waukesha County was categorized as a "vanguard site" by the National Children's Study, as it was part of the first set of pilot counties. Waukesha County is located in west-suburban Milwaukee, and has a population of 378,372 (American Community Survey 2006-2008 three-year estimate). Waukesha is a suburban county with a predominantly white non-Hispanic population and contains a mix of urban and rural environments.

NORC field staff traditionally listed a representative sample of 17 segments across Waukesha County during the fall of 2008, containing approximately 13,000 housing units (English et al. 2009). NORC then geocoded the USPS delivery sequence file (DSF) provided by the Valassis Corporation for Waukesha County in November, 2008, identifying those addresses that were within the 17 selected segments. The DSF file provided by Valassis is known as the 'ADVO' file and so we describe it as such in this paper. We matched the traditional listing and the ADVO addresses using LinkPlus probabilistic matching software package which permits "fuzzy" matching (and so tolerates differences in format, variations in spelling, etc.). (We omitted non-city style addresses, such as PO BOXes and rural-route boxes, from the ADVO file prior to matching, as they do not allow for block-level geocoding.)

We then repeated the matching process with another version of the delivery sequence file acquired from the vendor CIS, with a data vintage of August, 2008. The CIS file was provided to us by Westat, as they had employed it for quality-checking of the original listing. Following the second set of matching we had a three-way match, with addresses being present on any of the traditional listings, the ADVO DSF, or the CIS DSF. For purposes of notation we can describe addresses from the traditional listings as being in the "T" frame, those from the ADVO file as the "A" frame, and those from the CIS file as the "C" frame.

While at this point in the process we had a composite list of addresses within the selected segments in Waukesha County. We knew the frame or frames on which each address was included. However, we had no way to resolve any differences between lists (i.e., we did not know which list was correct in instances of disagreement). To determine which addresses were actually present, we had trained staff "check" or "enhance" the merged frame in late 2008 and early 2009 (Eckman and Kreuter forthcoming). The goal of the field checking was to determine which addresses were actually present, and to identify any addresses in selected segments not present on the lists (such as new construction). We then created an edited list, considering all frames and additions, which is known as the best or "B" frame.

We subsequently matched the B frame to a third DSF extract from the MSG vendor, vintage August 2009. The purpose of doing so was to further understand between-vendor variability, recognizing the vintages were off-set between the MSG (or "M") and other lists. In addition, because we did not field-check the M list, we know less about its properties than the A, C, or T frames.

The motivation behind the current research, however, was to understand the qualities of housing units and households missing from the above frames. Our first goal was to describe characteristics of the housing units that tend to be present on one or more sampling frames. To that end we captured n = 1064 GPS-enabled photographs of those units that were missing from at least one frame (T, A, or C). We then coded each housing unit in question using descriptions that could influence list inclusion or exclusion. Example descriptions included: whether the housing unit was single-family or multi-unit; whether the housing unit was "recently constructed"; whether the housing unit was in "poor condition"; whether the housing unit had readily visible house numbers; whether the housing unit was occupied; whether the housing unit was derelict; whether the housing unit was integrated with a commercial unit.

Screener variables of interest were then obtained during data collection during 2009. Such variables included eligibility status both at the housing-unit and household levels. *Housing-unit eligibility* required that an address exist, and be both residential and occupied. *Household-level eligibility* for the NCS required that a household contain one or more women aged 18-49.

Merging in these two sources of housing unit and household level variables help us to move beyond comparisons of coverage, and to understand the households and housing units included or excluded by one or more frame.

3. Results and Discussion

Overall, as shown in table 1, 97 % of the B frame was represented by the traditional listings, 94 % by ADVO, and 94 % by the CIS address vendor. The post-hoc match to MSG (M) had a 94 % overlap with B frame. These results are a few percentage points higher than those presented in 2009 as these are in every selected segment, rather than being limited to blocks with imperfect matching (English et al. 2009). It is also important to note that the vintage of the CIS list was somewhat older than the other lists, and so the differences in coverage are more likely due to age rather than any other reason. Note that while these results are unweighted, the National Children's Study sample design introduces very little variation at the segment level, and so the weighted results are essentially identical.

Table 1: Intersection of B with Individual Frames

Intersection	*Percent of B*
Traditional Listings (T)	97 %
ADVO (A)	94 %
MSG (M)	94%
CIS Addresses (C)	92 %

Table 2 below shows the overlap between all lists and reality in different environments in Waukesha County, WI. We can see that the T list appears to have somewhat more coverage than any DSF source in both urban and rural areas, and so would appear on the surface to be the optimal source in Waukesha County, WI. While we would expect the T list to have advantages in rural areas due to the prevalence of non city-style addresses, post-office or rural-route boxes, we would not in urban areas. The question is if the T list is adding vacant or derelict buildings in urban areas that have not had mail delivery in recent months, resulting in their loss from the USPS list. In addition, we would like to know if any members present on one list and not on another are "different" enough to affect survey results.

Table 2: Intersections of Each Frame with the Best Frame by Segment Urbanicity

Source	*Urban (n=5)*	*Suburban (n=8)*	*Rural (n=4)*	*Overall (n=17)*
Traditional (T)	98%	98%	95%	*97%*
ADVO (A)	95%	97%	88%	*94%*
MSG (M)	95%	96%	89%	*94%*
CIS (C)	96%	95%	84%	*92%*

We took three primary analytical approaches to examine the question of the importance of frame inclusion or exclusion. First, we explored housing-unit level eligibility via photographic and screener data, primarily through vacancy rates. Second, we examined household-level qualities through screener and interview data to understand if those households added from a given list were different enough to impact survey results. Third, we analyzed member-level data to determine if the qualities of the women within the households were distinct in terms of measures important to the National Children's Study.

Our motivation behind the acquisition of photographic data was to describe housing units in Waukesha County using factors that may theoretically influence list inclusion or exclusion. For example, the presence of high gates, unclearly-marked units, or threatening animals could reduce the likelihood an address will be correctly included on a traditional listing. Due to the relatively homogenous nature of housing in Waukesha County, being mostly single-family owner-occupied housing units, relatively few variables were of actual importance. These were as follows: whether a housing unit appeared to be "new" or "recently constructed" and whether a housing unit appeared to be "vacant". It is also important to mention that we only photographed housing units that were missing from one or more list on the initial matching between A, T, and C. So, the photographs examined may be expected to have more problematic housing units than the county as a whole.

Table 3 shows the source of housing units that were coded as "new" or "recently constructed"; 558 of the 1064 housing units photographed were described as such. One can see that the source of "new" homes was dominated by the traditional frame. Such homes tended to be overrepresented by those very recently constructed (often still for sale) and therefore not receiving mail. So, they may be expected to have lower occupancy rates than those from other sources.

Table 3: Source of "New" homes (558 of 1064)

Source	*n "New"*	*n not "New"*	*% "New"*
Traditional (T)	497	61	89%
ADVO (A)	322	236	58%
MSG (M)	332	226	59%
CIS (C)	55	503	10%

Table 4 shows the source of housing units coded as "vacant", of which 96 of the 1064 were coded. Relatively few housing units were described as being "vacant" due to the difficulty of this description from a photo; housing units would generally be more derelict than average to be coded this way. Photo-derived vacancy rates were heavily weighted to those housing units on the T list, with 82% of "vacant" homes being on that list. Both of the “new” and “vacant” descriptions support the narrative that additional coverage on the T list is not necessarily eligible households.

Table 4: Source of “Vacant” homes (96 of 1064)

Source	*n ”Vacant”*	*n not ”Vacant”*	*% ”Vacant”*
Traditional (T)	79	17	82%
ADVO (A)	56	40	58%
MSG (M)	62	34	65%
CIS (C)	12	84	13%

The next theme in our analysis concerned housing unit-level eligibility derived from actual NCS screener data, rather than coded photographic information. Specifically, we were interested in the percentage of housing units from a particular frame were occupied, "existed", and "not a business" during field work. While all lists themselves were approximately 99% eligible at the housing unit level, table five shows those from particular intersections of interest. All DSF vendors were very similar to each other in terms of coverage rates and so the ADVO-derived addresses are contained in the below tables. One should note that the figures in the below tables are for those housing units for which eligibility could be ascertained during the enumeration process. One can see that the “traditional only” and new adds (meaning addresses not present on any list prior to field-evaluation) had rates considerably lower than average, as these tended to be new construction or derelict buildings not on the DSF. Housing units missing from one or more lists had also well below average eligibility. Those housing units missing from the T list were often ineligible due to geocoding error placing the housing units erroneously inside a segment or from duplication inherited from list matching. Ineligible housing units missing from the DSF also indicated a tendency to be caused by geographic error, this time error by listers on the ground.

Table 5: Overall Housing Unit-Level Eligibility

Source	*Eligibility Rate*
Traditional (T)-Only	78% (n = 610)
New Adds	43% (n = 40)
Not on Traditional	68% (n = 277)
Not on ADVO	80% (n = 836)

Table 6 demonstrates the impact of urbanicity on eligibility for the frames we are focusing on, with cell sizes below 30 suppressed. Clearly, T appeared to add vacant addresses in urban areas that were "dropped" from the DSF, as indicated by the relatively low eligibility rate (36%) of such addresses. The DSF was not subject to much “true undercoverage" in urban areas as indicated by the low eligibility (41%) of those addresses not on ADVO; in fact, the DSF appears to add hidden, eligible housing units missed by T listing. Suburban housing units only on the T list had rather high eligibility, suggesting that traditional listing may capture housing units in suburban areas that are subject to DSF geocoding error due to irregular block patters. Rural areas behaved as expected, with the DSF missing some eligible cases due to the relative prevalence of non city-style addresses.

Table 6: Housing Unit-Level Eligibility by Urbanicity

Source	*Eligibility Rate- Urban*	*Eligibility Rate- Suburban*	*Eligibility Rate- Rural*	*Eligibility Rate- Overall*
Traditional (T)-Only	36% (n = 132)	85% (n = 196)	94% (n = 282)	78% (n =610)
New Adds	...	...	...	43% (n = 40)
Not on Traditional	95% (n = 44)	54% (n = 128)	74% (n = 105)	68% (n = 277)
Not on ADVO	41% (n = 144)	84% (n = 239)	92% (n = 453)	80% (n = 836)

Table 7 examines the household-level eligibility for addresses from the frames of interest. Eligibility in the National Children's Study entails containing a women age 18-49. While the overall household-level eligibility rates were 45-46% on all lists, those only on the traditional list were somewhat lower (40%). The above rates were exaggerated somewhat in urban areas.

We should note that we also acquired vendor provided demographic information (race/ethnicity and household income), but the results are not included in this paper due to their limited coverage.

Table 7: Household-Level Eligibility

Source	*Eligibility Rate- Urban*	*Eligibility Rate- Suburban*	*Eligibility Rate- Rural*	*Eligibility Rate- Overall*
Traditional (T)-Only	...	41% (n = 107)	41% (n = 153)	40% (n = 288)
New Adds	...	..	...	...
Not on Traditional	...	41% (n = 34)	43% (n = 30)	40% (n = 81)
Not on ADVO	33% (n = 36)	43% (n = 131)	48% (n = 271)	45% (n = 438)

4. Conclusions

Our preliminary results have shown that the traditional frame missed the fewest addresses overall in Waukesha County, WI. These results do confirm the belief that the T frame tends to contribute new homes that aren't getting mail, or derelict/long-term vacant homes that have been dropped from the DSF. As such we realized lower eligibility rates at the housing-unit and household levels for the addresses added by the T frame. Consequently, the marginal coverage benefit of the T list in Waukesha County, WI is muted, especially in urban areas. The other lists appeared to be comparable on most measures, with vintage being the most important factor. We therefore argue that the “true undercoverage” of the DSF may be less than what has been suggested by previous coverage comparison research.

It is important to emphasize that the DSF and T frames have advantages in different environments. For example, DSF-based frames may capture more hidden units in urban areas, while the T frame has advantages in areas with non city-style delivery. This observation argues for a combined DSF and traditional listing approach, where segments are tailored based on key characteristics, especially urbanicity.

One should also recognize the challenges and shortcomings in any frame evaluation work. First, matching address lists from disparate sources is always imperfect, with false-negatives and false-positives introducing some error to coverage rates. Second, vintages of lists can be off-set with themselves and with the timing of field data collection resulting in lists being “too long” or “too short” relative to each other.

Moving forward we would like to further characterize “patterns of missingness” in our frame sources. In so doing we will be analyzing new data from the NCS filed work, including household-level characteristics such as income and demographics. Additionally, we will be synthesizing our results from our photographic, screener, and interview data.

References

Battaglia, M. P., M. W. Link, M. R. Frankel, L. Osborn, and A. H. Mokdad. 2008. An Evaluation of Respondent Selection Methods for Household Mail Surveys. *Public Opinion Quarterly*, 72(3), 459-469.

Eckman, S. 2010. Errors in Housing Unit Listing and Their Effects of Survey Estimates. Ph.D. Thesis, University of Maryland.

Eckman, S., N. English and C. O’Muircheartaigh. 2010. The Use of Geocoding to Construct Survey Frames from Address Databases. Under Review.

Eckman, S. and F. Kreuter. Forthcoming. Confirmation Bias in Housing Unit Listing. *Public Opinion Quarterly*.

English, N., C. O’Muircheartaigh, K. Dekker, M. Latterner, and S. Eckman. 2009. Coverage Rates and Coverage Bias in Housing Unit Frames. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Iannacchione, V. G., J. M. Staab, and D. T. Redden. 2003. Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey. *Public Opinion Quarterly,* 67(2), 202-210.

Kennel, T. L. and M. Li. 2009. Content and Coverage Quality of a Commercial Address List as a National Sampling Frame for Household Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley and Sons, Inc.

Link, M. W., M. P. Battaglia, M. R. Frankel, L. Osborn, and A. H. Mokdad. 2008. A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, 72(1), 6-27.

Montaquila, J. M., J. M. Brick, and L. R. Curtin. 2010. Statistical and Practical Issues in the Design of a National Probability Sample of Births for the Vanguard Study of the National Children’s Study. *Stat Med*, 29(13), 1399-90.

Montaquila, J. M., V. Hsu, and J. M. Brick. 2010. Using a Match Rate Model to Predict Areas Where USPS-Based Address Lists May Be Used in Place of Traditional Listing. Under Review.

Montaquila, J., V. Hsu, J. Michael Brick, N. English, and C. O'Muircheartaigh. 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, American Statistical Association*.

O'Muircheartaigh, C. A., S. A. Eckman, and C. Weiss. 2003. Traditional and Enhanced Field Listing for Probability Sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

O'Muircheartaigh, C., English, N., Eckman, S., Upchurch, H., Garcia, E., and Lepkowski, J. 2006. Validating a Sampling Revolution: Benchmarking Address Lists against Traditional Listing. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

O'Muircheartaigh, C., English, N., Eckman, S. 2007. Predicting the Relative Quality of Alternative Sampling Frames. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Staab, J. M. and V. G. Iannacchione. 2003. Evaluating the Use of Residential Mailing Addresses in a National Household Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*.