Choices of Frame Construction on the National Children's Study: Impacts on Address Quality and Survey Results

Ned English[1], Katie Dekker[1], and Colm O'Muircheartaigh[2]

[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
The National Children's Study (NCS) is a large and complex longitudinal study, the initial stage of which is the enrolment of women of childbearing age. The NCS data collection depends on a sample of addresses within selected segments. Addresses could originate from a number of sources, depending on the nature of the environment: traditional listings, enhanced listings, or licensed address lists based on the USPS delivery-sequence file (DSF). The literature suggests that each method has different coverage properties in particular situations. At question are the characteristics of housing units, households, and ultimately women and children included using each method.
Our analysis compares the characteristics of households that were present on the original DSF vs. those that were added through listing. We do so using questionnaire data from a subset of sites to understand the types of women, babies, and potential health outcomes whose inclusion can be influenced by frame type.

Key Words: Address-based samples, area probability, National Children's Study, listing, frame construction

1. Introduction

The National Children's Study (NCS) is an innovative panel survey with the goal of understanding environmental, socioeconomic, and cultural impacts on child development (Montaquila et al. 2009, Montaquila et al. 2010a, Downs et al. 2010). As originally designed, the NCS intends to enrol a nationally-representative panel of 100,000 children follow them from before birth until age 21 for health and environmental testing. So, the NCS represents a study of almost unprecedented scale and scope (Michael and O'Muircheartaigh 2008). It is clear that the breadth of the NCS magnifies any impacts of frame construction and sample design decisions, as panel members will be maintained for a considerable length of time for a myriad of reasons (National Research Council 2008, Montaquila et al. 2010).

The NCS sample design was designed around a housing unit frame generated by traditional listing in selected area probability segments (Michael and O'Muircheartaigh 2008). "Traditional listing" is a method of address frame generation created by field staff known as "listers". Listers systematically record all residential addresses in defined geographies, regardless of occupancy status (Kish 1965, Eckman 2010). This method of frame creation had been considered the "gold standard" in the survey research industry until very recently, when new technologies encouraged the pursuit of less-costly alternatives (O'Muircheartaigh et al, 2003, O'Muircheartaigh et al. 2006, O'Muircheartaigh et al. 2007).

In the past decade it has become possible to license extracts of the United States Postal Service Delivery Sequence File (DSF or CDSF) from a particular set of vendors. The DSF embodies a list of all housing units in the United States that receive mail (O'Muircheartaigh, Eckman, and Weiss 2003, Amaya et al. Forthcoming). Survey research and government organizations have been researching the use of the DSF as a replacement for traditional listing, due to the implications for cost savings (O'Muircheartaigh et al, 2003, Iannacchione et al. 2003, O'Muircheartaigh et al. 2007, Battaglia et al. 2008, Link et al. 2008, Montaquila et al. 2009). The sum total is that the DSF is often adequate itself in urban areas, but may not be so in rural areas with non-city-style delivery (Staab and Iannacchione, 2003, Link et al. 2008, O'Muircheartaigh et al. 2009, Montaquila et al. 2010b). One would need to traditionally list or employ a hybrid approach in such areas to avoid risk for undercoverage (Montaquila et al 2010, Eckman 2010).

One hybrid approach is "enhanced" or "dependent" listing, where listers begin with the universe of addresses believed to be in an area, and then edit and augment the list where necessary (Eckman and Kreuter 2011). Examples of the use of dependent listing include the Census update of the Master Address File (MAF), The National Survey of Family Growth at the University of Michigan, and various NORC studies. Enhanced listing is considered more efficient than traditional listing, due to the presence of a starting list, and carries coverage advantages of the DSF in urban areas, e.g., multi-unit or hard-to-find buildings (Eckman 2010).

Frame construction was at the discretion of the NCS study centers, beginning with the listing process, meaning one could choose to implement traditional listing, enhanced listing, or use the DSF alone. At question is the impact of such decisions. Is enhanced listing always worth undertaking, and can one "get away" with not enhancing? Also, how does enhanced listing impact multi-mode surveys that require both a mail and phone component? We explore the use of enhanced listing to fill-in-the-gaps in rural areas.

In addition, the NCS was originally based on an in-person data collection model, but is currently comparing in-person enumeration with multi-mode and provider-based approaches in an "extended pilot" phase. We examine the kinds of households that were added during enhanced listing, and the impact on both enrolled households and data collection itself.

Our research continues our recent evaluation in one suburban NCS county that showed that addresses not present on the DSF tended to be different than those found on the DSF (English et al. 2009, English et al. 2010). At the time of the evaluation, however, we didn't have screener or interview data for household members. The current research moves beyond the initial work by reporting on the actual households, in a mix of suburban, urban, and rural counties, and considering differences between them.

Analyzing household screener data permits moving the overall discussion beyond rates of undercoverage.

Our study attempts to address a few "big" picture questions. Primarily, we would like to know further about what types of environments suggest that one should enhance the DSF. Moreover, what kinds of households would be at risk of undercoverage if one did not? Second, it would be valuable to know how the omission of such households from the study would affect resulting data. It is possible to ascertain such an effect by analyzing screener data based on the frame origin of each case e.g., if they were present on the original DSF or were added through enhanced listing. Third, we would like to know how mode impacts frame construction decisions, as most research thus far has focused on in-person studies. The NCS presents an opportunity to further understand what may be included or missing on different sampling frames in disparate environments.

2. Methods

The current research is based on four NCS sites (counties) where NORC was responsible for listing and data collection. The four sites for this analysis were Cook County, IL (Chicago); Cumberland County, ME (Portland); Polk County, IA (Des Moines); Westmoreland County, PA (Greensburg). While these four sites are not nationally-representative, they do embody a mix of urban, rural, and suburban areas, both within and across counties. Cook County, IL and Polk County, IA are dominated by urban and suburban areas, while Cumberland, ME and Westmoreland, PA are characterized by rural areas and smaller towns (with the city of Portland, ME, being the exception). Each of the sites was listed in the summer and fall of 2010 using enhanced listing. So, the USPS[1] delivery sequence file from contemporaneous months during 2010 was geocoded and subset to selected segments in each county. Listers than verified each address in-person, and added any addresses found on-the-ground that were not present on the list. The goal of the enhanced listing is to create a list with all housing units present in each segment, both from the DSF or DSF omissions.

We then undertook data collection in the four counties, starting in early 2011. Two of the sites underwent in-person data collection, known as "enhanced enumeration". These sites were Cumberland, ME and Polk, IA. The other two, Cook, IL and Westmoreland, PA, adopted a multi-mode approach known as "high/low". "High/low" essentially employed mailed self-administered questionnaires (SAQ) combined with outbound telephone calls. In all counties (enhanced enumeration and high/low) we fielded the frame resulting from enhanced listing, which contains any address present on the DSF that was verified in the field, plus those added on the ground that were not present on the DSF. Addresses that were not found on the DSF (such as "tear downs") or addresses falsely included due to geocoding error were removed from the frame prior to sampling. We expect "adds" to consist of households that receive their mail at non city-style addresses (PO and RR boxes), new construction, and geocoding error.

[1] NORC licenses the DSF file provided by Valassis; this file is known as the "ADVO" file, and so we describe it as such in this paper.

We collected screener data of interest as part of the NCS questionnaire. 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. We also obtained demographic information, including the race, age, and income of the householder. Lastly, we collected study-specific information related to the pregnancy status of the households. Our main comparison is between households present on the DSF vs. those added during enhanced listing, in order to stimulate the effect of not listing.

3. Results and Discussion

Table 1 shows the outcome of enhanced listing in each site, meaning what happened to the original geocoded DSF lines in each segment once checked in the field. It is important to emphasize that all subsequent tables are preliminary and unweighted, as they were produced during the data collection period. Nonetheless, we do not expect substantial differences between these findings and those in a final report, especially for the in-person sites. Also, any counts in the following tables are rounded to the nearest 50 following National Children's Study protocol.

As shown in table 1, the urban sites Cook, IL and Polk, IA had the highest rates of address confirmation, meaning addresses that were verified to exist in-person. As expected, the two more rural sites had considerably lower confirmation rates. Also not surprising, the rural sites (Cumberland, ME and Westmoreland, PA) had a substantial share of "true" adds, which are addresses not present on the DSF in the segment. Such households would receive mail via non city-style delivery (PO and RR box), and thus would not be covered if we did not enhance the DSF. "True" adds are distinct from "geo" adds, which were addresses on the DSF but geocoding outside a selected segment. So, listers recorded them in the field because the addresses were not present on their list, but they were indeed on the DSF elsewhere. These "geo" adds are counteracted by "removes", which would be addresses on the listing sheets not found in their segment. Many such removes are procedural, especially in rural areas. When a lister "removes" a line, it does not always mean that the line doesn't exist, rather that the line doesn't exist in the expected Census block. Often times, lines that are "removed" from one block in a segment are added back as "geo" adds on another block in the segment. Rural sites had the most cleaning, with "geo adds" often counteracting removes. As expected, urban sites had the most 'pristine' DSF, and thus the least need to employ enhanced listing. Geocoding error is more pronounced in the less urban areas due to less developed street databases (Eckman and English, 2011).

Table 1: Outcome of enhanced listing

Category	*Description*	*Polk, IA*	*Cook, IL*	*Cumberland, ME*	*Westmoreland, PA*
Confirm	On List, found in block	89%	92%	61%	41%
"True" Add	Not on List, Found in Block	1%	2%	18%	15%
Remove	On List, not Found in Block	6%	4%	13%	27%
Geo Add	On List, Geocoded in Different Block in Segment	4%	2%	7%	17%

It is clear from table 1 that enhanced listing can substantially augment or edit the starting list, especially in rural areas. Tables 2 and 3 show the results of enumerating the households in the in-person sites, both the more rural Cumberland, ME (table 2) and the more urban Polk, IA (table 3). Again, it is important to emphasize that these results are from a point in time (preliminary data) and are rounded to the nearest 50. These tables can be used to effectively simulate not enhancing the list, as they show eligibility for the raw DSF, that of the enhanced DSF, and what would happen if one fielded only the addresses that were added on the ground ("true" adds). So, the tables show the kinds of households that tend to be added on the ground.

Table 2 indicates that there were a substantial number of addresses added during enhanced listing, which were housing units at rates similar to the DSF itself. The majority of buildings not appearing as housing units would be group quarters, rather than addresses torn-down or demolished. Occupancy, however, was considerably lower for those added in the field. Experience has shown that many added addresses are to be long-term vacant units removed from the DSF.

Results for Polk County, IA, a mixed urban suburban county, are shown in table 3. It is clear that there were many fewer "true adds" in Polk County, in contrast with Cumberland County.

Table 2: In-Person Eligibility, Cumberland, ME

Category	*DSF (12950)*	*EDSF (16450)*	*True Adds (3450)*
% Housing Units	96%	96%	94%
% Occupied Housing Units	88%	79%	45%
% Eligible Households	53%	52%	43%

Table 3: In-Person Eligibility, Polk, IA

Category	*DSF (9250)*	*EDSF (9350)*	*True Adds (100)*
% Housing Units	100%	99%	85%
% Occupied Housing Units	94%	94%	87%
% Eligible Households	58%	58%	55%

We have described characteristics of addresses added to the DSF during enhanced listing in our in-person sites as those addresses representing households at risk of undercoverage if the DSF alone were used for data collection. Multi-mode approaches have different concerns than in-person, as they depend on processes such as mailing and telephone matching for data collection. Consequently, multi-mode surveys require "cleaner" or more standardized address data than in-person studies. Multi-mode data collection may therefore be expected to be more sensitive to the source of an address e.g., if it were present on the original DSF or were added in the field. The reason for this is that households added in the field during enhanced listing may actually receive mail via PO or RR box delivery, and therefore not be mailable to their apparent address. Furthermore, many rural housing units do not have formal address numbers and so are listed by description. Housing units listed by description are also unmailable and cannot be matched to telephone numbers. We thus can measure the effect of enhanced listing on the multi-mode or "high/low" sites differently than the in-person sites.

Table 4 shows the impact of enhanced listing on measures important to multimode surveys in the rural Westmoreland, PA site, while table 5 does the same in the urban Cook, IL site. Westmoreland, PA had a substantial quantity of "true adds" generated during enhanced listing. Such addresses tended to be more challenging for operations, however, as seen in the overall lower rates of telephone matching and successful mail delivery. Cook County had many fewer "true adds", but those that were added were less successful at matching telephone numbers.

The previously-described tables show that "true adds" are more difficult to match telephone numbers or mail addresses to. Such a fact should be considered when designing multi-mode surveys in rural areas. In addition, we can see that "true adds" behave differently depending on the environment. Specifically, while "true adds" are expected to embody non city-style addresses in rural areas, they represent something different in urban areas e.g., new construction not yet on the list or chronically vacant units. It is important to emphasize that these rates are for addresses that were actually fielded, and so they do not include addresses dropped from the list because they did not exist as housing units in the segment.

Table 4: Multimode Eligibility, Westmoreland, PA

Category	*DSF (14700)*	*EDSF (18600)*	*True Adds (1050)*
% Phone Matching	65%	57%	37%
% Mail Undeliverable	3%	5%	12%
% Eligible Households	30%	30%	28%

Table 5: Multimode Eligibility, Cook, IL

Category	*DSF (10300)*	*EDSF (10550)*	*True Adds (200)*
% Phone Matching	44%	43%	30%
% Mail Undeliverable	5%	5%	5%
% Eligible Households	37%	37%	**

Beyond eligibility and matching rates, it is possible to examine actual characteristics of those screened households. At the time of this analysis, there were only enough screened households from our preliminary analysis in Cumberland, ME, as shown in table 6. Very few addresses were added to the DSF in Polk County, IA, and data collection was still under way in Cook and Westmoreland. As shown the rates White non-Latino, home-owner, and married were higher in added lines than those on the DSF. Of interest to the NCS, the percentage of households containing a woman who was pregnant was also higher, but the share of women attempting to become pregnant was reduced.

Table 6: Household Metrics for Cumberland, ME (in-person)

Metric	*DSF*	*EDSF*	*True Adds*
% White non-Latino	91%	91%	96%
% Own Home	59%	61%	76%
% Ever Married	47%	49%	57%
% English-Speaking	96%	96%	99%
% Pregnant	2.6%	2.7%	3.1%
% “High Trier”	3.7%	3.5%	1.8%

We can see that enhanced listing has varying impact by environment. Urban areas (such as in Cook, IL and Polk, IA) add very few addresses that were not present on the DSF and geocoding inside the segment (< 1% of the total). Enhanced listing does remove demolished units, ineligible units (such as group quarters), and corrects geocoding error. So, we would expect the enhanced list to carry efficiencies over the “raw” list. Rural areas (Cumberland, ME and Westmoreland, PA) saw a considerable share of new adds during enhanced listing (nearly 20% of total). There would therefore be a real risk of undercoverage in such areas if the list were not enhanced. It is clear there is heterogeneity within all the previously described counties, however, with areas urban enough to suffice with the DSF alone.

“True adds” can be shown to exhibit distinct characteristics from those addresses initially present on the DSF. For example, added addresses have lower occupancy and eligibility rates, pointing to vacant units removed from the DSF and newly constructed units not yet occupied. We also observed lower telephone matching and mail delivery rates compared with those present initially on the DSF. Of similar interest to the NCS, addressees added to the DSF during enhanced listing had different demographic and pregnancy characteristics.

It is clear that the highest value of enhanced listing would be in rural segments with non-city style delivery. Such areas also exhibit more geocoding error due to limited GIS street databases, and so even those addresses present on the DSF may appear in the incorrect block and thus necessitate editing. Ultimately, the tolerance for when to e-list depends on the individual study, and may be expected to vary depending on the final use. We would recommend enhanced listing for added efficiency and "clean up" where possible, however, even if coverage is not a concern in urban and suburban areas.

4. Conclusions

It is important to emphasize that the above results are preliminary, and are meant to represent an initial discussion on the kinds of issues related to enhanced listing. We will be updating these results at the completion of field work, which are expected to change the final numbers but not necessarily the overall story. We will also be pursuing follow-up data, including pregnancy results and biomeasures, to differentiate the kinds of women included or excluded from different frames. Going forward, we will be updating our models to predict the kinds of areas that benefit from particular frame construction methods, using Census 2010 data.

References

Amaya, Ashley, Felicia Leclere, Lee Fiorio, and Ned English. Forthcoming. Improving the Utility of the DSF Address-Based Frame through Ancillary Information. *Field Methods.*

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.

Downs, Timothy J., Yelena Ogneva-Himmelberger, Onesky Aupont, Yangyang Wang, Ann Raj, Paula Zimmerman, Robert Goble, Octavia Taylor, Linda Churchill, Celeste Lemay, Thomas McLaughlin, and Marianne Felice. 2010. Vulnerability-based Spatial Sampling Stratification for the National Children's Study, Worcester County, Massachusetts: Capturing Health-Relevant Environmental and Sociodemographic Variability.

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

Eckman, Stephanie and Ned English. Forthcoming. Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates. Field Methods.

Eckman, S. and F. Kreuter. 2011. Confirmation Bias in Housing Unit Listing. *Public Opinion Quarterly* 75 (1): 139-150

English, Ned, Colm O'Muircheartaigh, Katie Dekker, and Lee Fiorio. Qualities of Coverage: Who is Included or Excluded by Definitions of Frame Composition. 2010

Proceedings of the American Statistical Association, Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

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.

Michael, R.T. and O‟ Muircheartaigh, C. 2008. Design Priorities and Disciplinary Perspectives: The Case of the U.S. National Children's Study. *Journal of the Royal Statistical Society: Series A*, 171, Part 2, 465-480.

Montaquila, J. M., J. M. Brick, and L. R. Curtin. 2010a. 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. 2010b. 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): 317-335.

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*.

National Research Council (US) and Institute of Medicine (US) Panel to Review the National Children’s Study Research Plan. 2008 Washington (DC): National Academies Press (US)..

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*.