Coverage Implications of Targeted Lists for Rare Populations

Ned English, Ipek Bilgen, and Lee Fiorio

NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract
One challenge to using a multi-mode address-based sampling (ABS) design to target minority populations can be low eligibility. At issue is the relative cost efficiency of ABS as an alternative sampling design is inversely related to household eligibility. One potential way to improve operational efficiency would be to enrich the original address frame by using race/ethnicity-targeted lists. Such lists can be used to stratify a general population address list by indicating households likely to contain members of a targeted racial/ethnic group. This paper provides an initial investigation into the impact the use of targeted lists may have on household coverage and resulting survey data in a health survey. Using a binary logistic model, we find that the coverage of race/ethnicity targeted lists declines in dense, urban areas with large populations of renters and low priority group density. List coverage is less likely to be adequate for African American households compared to Asian or Hispanic households due to the use of surnames.

Key Words: Address-based samples, targeted lists, hard-to-reach populations, minority populations, frame construction, modeling coverage

1. Introduction

Survey research has witnessed a transition over the past decade from random-digit dial surveys (RDD) to multi-mode studies based on the United States Postal Service delivery-sequence file (DSF or CDSF) (Brick et al., 2011; Link et al., 2009; Iannacchione et al., 2003; O'Muircheartaigh et al., 2003). This shift has been motivated by the degradation of RDD coverage and response rates, coupled with the theoretical promise of nearly-universal household coverage of the DSF (Link et al., 2009). Of concern to many studies, however, are the cost implications of multi-mode surveys in situations targeting households rarer than the general population (Link et al., 2008). Multi-mode studies, often employing a combination of telephone, mail, and in-person methods, have been shown to be cost-equivalent or superior to RDD on a per-case basis for general household surveys (Amaya & Ward, 2011). One observation has been that as eligibility decreases, costs may become prohibitive for ABS due to associated per-unit costs necessitating alternative approaches including RDD. The discipline would thus benefit from a means to increase eligibility for sparse populations such as priority race and or age groups, to take advantage of the coverage benefits while remaining economically feasible.

REACH U.S., an acronym for "Racial and Ethnic Approaches to Community Health across the U.S.," is a Centers for Disease Control and Prevention (CDC) program designed to eliminate racial and ethnic health disparities by funding local health interventions. NORC at the University of Chicago conducts a risk-factor survey, where we monitor health indicators in 28 communities; each community has defined geographic

areas, with specific priority races/ethnicities. While the previous contract from 2001-2005 was conducted as RDD, *REACH U.S.* adopted multi-mode ABS in 2008 employing telephone, mail, and in-person methodologies.

The REACH U.S. communities are highly diverse with respect to the priority races/ethnicities, as well as the expected eligibility rates, which range from 10-96% per Census. For communities with low or lower eligibility rates, there are potentially two ways to gain sampling efficiencies and lower survey costs: (1) area stratification and (2) race targeted lists. While area stratification may increase overall eligibility by adjusting selection probabilities depending on target population density, the effectiveness is limited in areas with evenly-distributed target populations. Conducting household-level stratification would be a potentially more efficient alternative, if one could reliably identify households likely to contain eligible members. At question is if one may use "targeted lists" of households "flagged" as having at least one member of a particular race/ethnicity to consistently identify eligible households. Such lists are created by vendors for marketing purposes, and contain demographic and purchasing information beyond race/ethnicity.

Our current research considers two questions related to the use of race-targeted lists. First, we consider how well targeted lists cover the population of interest as compared to Census 2010. In our analyses, communities of interest included African-American, Hispanic/Latino, and/or Asian households. Second, we examined the area-level characteristics that might be associated with high or low coverage of targeted lists, as pursued through modeling. This research is relevant to the current state of survey research due to the proliferation of address-based studies, and the constant need to remain cost competitive.

2. Background

Vendors such as InfoUSA, Marketing Systems Group (MSG), Targus, Survey Sampling International (SSI), Valassis, and others provide household-level data containing demographic "flags." Household data may include information about household members' gender, age, race/ethnicity, education, and other market-oriented information. These data are generated from proprietary sources, at least in part through models, including surname lists, consumer data (warranty cards, periodical subscriptions, etc.), as well as Census data (InfoGroup 2012).

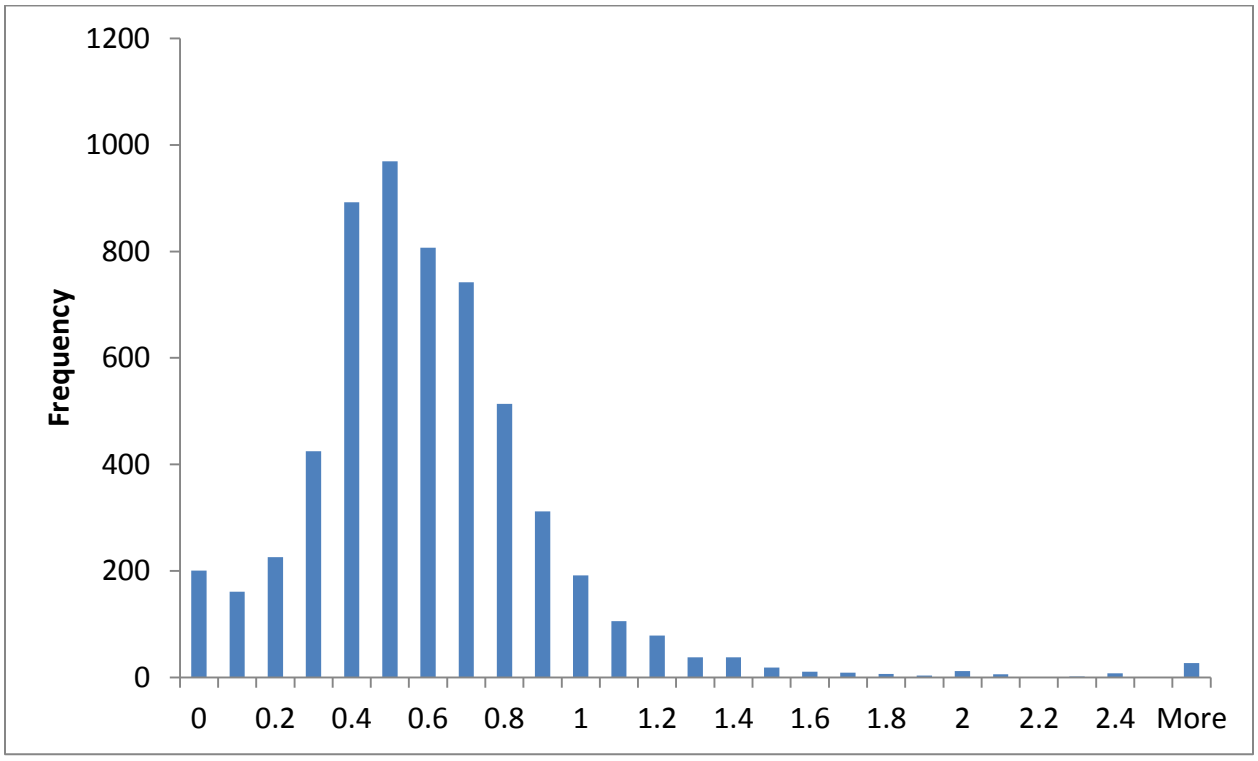
To stratify the frame using the race/ethnicity lists, we first licensed a list of all households expected to contain members of a priority racial/ethnic group in a community's survey geography from InfoUSA. Then, we matched each address on the targeted list to the USPS delivery-sequence file (DSF or CDSF) provided by Valassis. The DSF has been evaluated to contain essentially complete household coverage in urban areas, such as those associated with *REACH U.S.* target neighborhoods O'Muircheartaigh (O'Muircheartaigh, Eckman, & Weiss, 2003; Kennel & Li, 2009; Amaya et al., Forthcoming). Because of the coverage limitations of the targeted lists, in each *REACH U.S.* community there were households present on the DSF that were not on the race/ethnicity targeted list. Our sample design was based on sampling households at differential rates depending on their presence or absence from the race-targeted lists, which we can refer to as "household-level stratification."

3. Data and Methods

At question is how well the targeted lists cover the population of interest; if such lists account for only a small share of the actual population believed to exist in a given area we may be concerned with the risk of coverage bias. We define the measure "target ratio" as the ratio of the vendor count (InfoUSA) of households containing members of a given race/ethnicity, divided by the expected number of said households from Census 2010. Target ratio is an analog for coverage, as we assume the risk of bias decreases as the frame size approaches the population size. It is important to note that higher ratios do not necessarily mean that the same addresses are on both InfoUSA and Census lists, rather that they had similar numbers of addresses. Similarly, low ratios do not always indicate coverage bias, which would be dependent on the differences between households themselves.

To make our analysis of race/ethnicity flag coverage more robust, we calculate the target ratio at the Census block group level. This finer level of analysis allows us to investigate coverage disparities within communities. Figure 1 below shows the distribution of target ratio by block group across *REACH U.S.* communities, with a median target ratio of .53 and a range of 0 to 2.8. Ratios below 1.0 indicate that there were fewer households from the targeted list than those expected according to Census 2010, while those above 1.0 show the opposite; we may interpret the latter as areas of new construction or potentially as error on the part of either data source. Of note, the target ratio varies considerably when calculated at the community level.The ratio of priority race/ethnicity flagged households to Census tabulated priority race/ethnicity households ranges from .35 in the Bronx to .73 in Richmond, VA.

Figure 1. Target Ratio by Block Group across REACH U.S. Communities



A binary logistic regression is used to model the likelihood that the target ratio of a given block group is in one of two categories: at or above .5, implying "adequate" coverage, or below .5, implying "inadequate" coverage.

The independent variables included in the model were derived using data from Census 2010 or ACS '06-'10 five year estimates. In particular, we were interested in the relationship between the demographic and housing characteristics of each block group from Census 2010 and its target ratio. We hypothesized that elevated levels of poverty, foreign born populations and rented housing units would be associated with lower coverage of the targeted lists, while higher rates of occupied housing and housing unit density would be associated with higher coverage. Our reasoning was that the targeted lists were created at least partially from consumer activity and active credit accounts, which would be less-visible among renters, the foreign-born, and those living in poverty. To simplify the model, housing unit density was coded as a binary variable in which a Census block group featuring more than 6,000 HUs per square mile was considered high density and anything at or below 6,000 HUs per square mile was not considered high density; this threshold was defined as approximately the mid-point of the density distribution. As an indicator of the rurality of each block group, we used Census Type of Enumeration Area (TEA) code; TEA is used by the Census Bureau to indicate if a block group is sufficiently urban to enumerate via mail-out/mail-back, or requires in-person address updating and data collection. If any of the component blocks in a block group were not enumerated by the Census using the USPS then we deemed it rural and hypothesized that the coverage of targeted race/ethnicity flags would be limited due to more difficult address matching and household association. We controlled for DSF coverage using a categorical variable based on the DSF-to-Census ratio, calculated as the DSF housing unit count divided by the Census occupied housing unit count. From previous research we know the DSF is most effective as a sampling frame in areas where it can account for 90% to 110% of the occupied housing units according to Census (English et al., 2009; O'Muircheartaigh et al., 2009; O'Muircheartaigh et al., 2007). If the ratio was much above or below that threshold, we would anticipate either under or over-coverage on the part of the DSF; hence, we included three categories, DSF-to-Census ratio below 90%, between 90% and 110%, or above 110%.

To take into account racial/ethnic diversity and the concentration of the priority racial/ethnic group in each *REACH U.S.* community, we calculated the percentage of each block group made up by the priority group and percentage of each block made up by non-Hispanic whites. Places with high concentrations of the priority group should be easier to flag than heterogeneous places where the priority group is distributed among other races/ethnicities. Finally, we controlled for Census region with a categorical variable: Midwest, Northeast, South, and West, and for priority group with three indicator variables: (1) Asian and Pacific Islander, (2) Hispanic/Latino, and (3) African American. Because several of our communities target multiple races/ethnicities, it is possible for multiple indicator variables to equal one for a community.

4. Results

Results from our logistic regression model are presented in Table 1. Not every variable had a significant effect on the model's ability to predict the coverage of race/ethnicity lists. While we correctly hypothesized that rented housing units would be negatively associated with coverage, as expressed by the target ratio, we were incorrect in our assumption that higher HU density could result in better coverage. In our model, elevated

levels of block-level poverty decrease the likelihood that coverage will be adequate, but this finding is only approaching statistical significance. Block groups in highly urban and impoverished communities like those in Bronx, NY and Chicago, IL experienced lower levels of targeted list coverage despite relatively high concentrations of the priority group. Similarly, percent rented housing is demonstrated to decrease the likelihood of adequate coverage; the target ratio is more likely to be above .50 in block groups in *REACH U.S.* communities with stable housing – high occupancy and low percentage of households renter occupied.

Our model also demonstrates that priority group density will increase the likelihood of adequate coverage. This finding is indicative of the fact that many targeted lists are known to be at least partially based on Census data; for example, households living in highly-concentrated areas according to Census may be flagged on the targeted list as the most likely race/ethnicity. It stands to reason that one would be more successful in identifying Asian households, for example, in a block group that is 90% Asian according to the Census than a block group that is only 10% Asian. What is interesting from our model is that both priority group density and non-Hispanic white density significantly increase the likelihood of adequate targeted list coverage (see Table 1). Specifically, the probability of adequate coverage in areas with high target density and high non-Hispanic density is five times more likely than in areas with low target density and low non-Hispanic white density. While these two findings seem contradictory, they indicate that coverage of targeted lists increases in areas that are segregated rather than those that are diverse. Communities that feature both low priority group density and low densities of non-Hispanic whites are heterogeneous making it more difficult to identify households of a particular race/ethnicity.

Interestingly, coverage tends to be higher in communities that target Asians/Pacific Islanders; communities that target African Americans had relatively lower coverage. Because these race/ethnicity flags are at least partially created using surname lists, it is not surprising that African Americans would be more difficult to identify. What is curious, however, is that Hispanic/Latino households are not similarly easy to identify as Asians according to the model. Our belief is that this finding has to do with the nature of the *REACH U.S.* communities targeting Hispanics which tend to be in areas that are poorer, urban and more diverse.

At question is how well our logistic model did at predicting the target ratio for each block group. Because we know the actual target ratio for each block group, we can evaluate how well the model performs. Table 2 below demonstrates the rates at which we correctly categorized a block group as having a high target ratio (1,1) or low target ratio (0,0), as well as instances we were incorrect in either direction (0,1 and 1,0 respectively). The model correctly classifies the target ratio 77% of the time, performing best in larger urban areas such as Chicago, Los Angeles, or Seattle. Such communities had more block groups that others, and thus had a higher influence on the model as a whole. Smaller communities with fewer block-groups, such as Hawaii, were the least-predictable according to our model.

Table 1. Odds Ratios for Variables Predicting Adequate Race/Ethnicity List Coverage

Variable or Predictor	Odds Ratio	95% Confidence Interval	
DSF-to-Census Ratio < 0.9 (vs. 0.9 to 1.1)	0.103***	0.081	0.131
DSF-to-Census Ratio > 1.1 (vs. 0.9 to 1.1)	1.381***	1.173	1.625
Density of Priority Group	5.595***	2.043	15.323
% Pop. Below Poverty	0.594	0.346	1.02
% Pop. Foreign Born	1.109	0.678	1.816
% HUs Occupied	5.038**	1.644	15.438
Rural TEA Block	1.171	0.702	1.954
HU Density > 6,000 HUs/sq. mile	0.525***	0.43	0.641
% HUs Rented (not Owner Occupied)	0.013***	0.009	0.019
Non-Hispanic white density	5.501**	1.927	15.706
Census region - Northeast (vs. Midwest)	2.623***	1.745	3.943
Census region - South (vs. Midwest)	0.335***	0.164	0.685
Census region - West (vs. Midwest)	0.529**	0.374	0.748
African American	0.447***	0.342	0.585
Hispanic/Latino	1.365	0.767	2.431
Asian/Pacific Islander	2.884***	2.327	3.573

* $p \leq .05$, ** $p \leq 0.01$, *** $p \leq 0.001$

Table 2. Model Success Outcome

Accuracy	Obs	Pct
Target Ratio = 1, Model predicts 1	2,279	39.1%
Target Ratio = 0, Model predicts 0	2,215	38.0%
Target Ratio = 1, Model predicts 0	685	11.8%
Target Ratio = 0, Model predicts 1	645	11.1%
Total	5,824	100.0%

5. Discussion and Conclusions

The purpose of our paper was to evaluate the coverage properties of race/ethnicity targeted lists, due to their utility for enhancing address-based surveys. In so doing our results hint at how targeted lists are constructed, based on the kinds of block groups that tend to have higher or lower coverage. Block groups with the highest ratios of targeted households to Census 2010 tended to have a concentrated priority group, high levels of home-ownership, and/or groups with distinct surnames. Targeted lists were the least successful in situations with PO Box delivery, low priority group density, renters, very high population density, or an African-American target population. Such areas are under-

represented on vendor-provided lists due to the dependence on commercial transactions (renters would not have property-transfer information, for example), and surname lists (African Americans do not have distinct surnames). Users should be aware of the varying effectiveness of specialized targeted lists, depending on specific groups and areas of interest. Area-stratification based on Census or American Community Survey controls may be preferable in some situations, either alone or in concert with race/ethnicity targeted lists.

In developing a regression model to predict which block groups would experience "better" or "worse" coverage we found that different communities can be challenging to treat in a single model. Overall, it was reasonably successful (77%), with some groups experiencing higher accuracy than others.

We know from the literature that coverage rates may relate to coverage bias (Groves, 2004). However, there may not be a direct relationship between the coverage rates and any substantive bias in *REACH U.S.* Therefore, one of our next steps is to analyze whether the key statistics obtained from the targeted lists versus DSF-only lists impact coverage bias for targeted sub-populations (Bilgen et al., 2012). Moving forward, it would be beneficial to consider spatial regression models to correct for spatial correlation within communities, and to investigate "islands" or "hot-spots" of predictability.

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.* Forthcoming.

Amaya, Ashley and Christopher Ward. 2011. *Cost Efficiency: Which Design is Cheapest?* Presented at the American Statistical Association Annual Meeting, Miami, FL.

Bilgen, Ipek, Ned English, and Lee Fiorio. "Coverage and Data Quality Association in Enhanced Address-Based Sample Frames". *Proceedings of the Joint Statistical Meetings.*2012

Brick, J. Michael, Douglas Williams, and Jill M. Montaquila. 2011. Address-based Sampling for Subpopulation Surveys. *Public Opinion Quarterly* 75(3), 409-428.

English, Ned, Colm O'Muircheartaigh, Katie Dekker, Michael Latterner, Stephanie Eckman. "Coverage Rates and Coverage Bias in Housing Unit Frames." *Proceedings of the Joint Statistical Meetings.*2009.

Groves, R. 2004. *Survey Errors and Survey Costs*. John Wiley and Sons.

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

Infogroup. 2012. *Cultural Coding: Ethnicity, Ethnic Group, Language, Country of Origin, Religion*. Infogroup Database Content Group Business & Consumer Databases www.infogroup.com

Kennel, Timothy L., and Mei Li. "Content and Coverage Quality of a Commercial Address List as a National Sampling Frame for Household Surveys." 3 *Proceedings of the Joint Statistical Meetings.* 2009.

Link, Michael W., Gail Daily, Charles D. Shuttles, Tracie L. Yancey, and H. Christine

Bourquin. 2009. "Building a New Foundation: Transitioning to Address-Based Sampling After Nearly 30 Years of RDD". *Proceedings of the American Statistical Association, AAPOR* [CD ROM], Alexandria, VA: American Statistical Association.

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008b. "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) For General Population Surveys. *Public Opinion Quarterly* 72(1), 6-27.

O'Muircheartaigh, Colm, Ned English, Michael Latterner, Stephanie Eckman, and Katie Dekker. "Modeling the Need for Traditional vs. Commercially-Available Address Listings for In-Person Surveys: Results from a National Validation of Addresses." *Proceeding of the Joint Statistical Meetings.*2009.

O'Muircheartaigh, Colm, Edward English, and Stephanie Eckman. "Predicting the Relative Quality of Alternative Sampling Frames." *Proceedings of the Joint Statistical Meetings.*2007.

O'Muircheartaigh, C. A., S. A. Eckman, and C. Weiss. 2003. Traditional and Enhanced Field Listing for Probability Sampling. *Proceedings of the American Statistical Association, AAPOR* [CD ROM], Alexandria, VA: American Statistical Association.