Evaluation of Differential Sampling in the Knowledge Networks Panel: What is the Effect on Coverage, Variance and Bias for Key Statistics?

Vicki J. Huggins, Xiuli Tang, and Mikyoung Park
Knowledge Networks, 1360 Willow Road, Menlo Park, CA 94025

KEY WORDS: Differential Sampling, Internet Surveys, Coverage, Design Effects

I. Introduction

Sample selection for the Knowledge Networks Panel is accomplished using standard RDD methods, which in principle generates equal probability sample designs. To improve the efficiency of some estimates and to reduce cost, oversampling of certain groups is applied. The following groups are differentially sampled in the Knowledge Networks Panel: households with Black and Hispanic members, households for which addresses can be obtained for the randomly generated telephone numbers, households in central region states, and households in areas of the U.S. not serviced by the MSN®TV as an Internet Service Provider. Results of an evaluation of the impact of the differential sampling on cost, coverage, sampling variance and bias for key statistics will be discussed.

II. Overview of Knowledge Networks Panel Design and Sample Weighting for Individual Surveys

Successfully targeting a nationally representative panel sample over the Internet has been intractable, primarily because a large proportion of U.S. households do not have Internet access. An innovative approach implemented at Knowledge Networks overcomes this inherent shortcoming. The methodology begins with selection of a representative sample of households using RDD telephone methods. By phone, the sampled households are asked to participate in the Knowledge Networks research. Once recruited, households are shipped an MSN TV device to connect to the Internet and their T.V. Surveys are sent to them over the Internet to complete. This approach ensures that both households with and without Internet access at the time of recruitment are included in the Knowledge Networks panel sample.

The following differential sampling features are included in the RDD sample selection and telephone recruitment methods for the Knowledge Networks panel:

1. Once the RDD telephone numbers have been purged and screened, we address match as many of these numbers as possible. The success rate so far has been in the 60-70% range. The telephone numbers with addresses are sent an advance letter informing the household of the opportunity to join the panel and the fact that they will be telephoned. The remaining, unmatched numbers are subsampled in order to reduce costs. The reduced field costs resulting from this allocation strategy more than offsets increases in the design effect for most key characteristics.

2. As part of the field data collection operation, we collect information on the number of separate phone lines in the selected households. We correspondingly down-weight the households with multiple phone lines.

3. Two pilot surveys carried out in Chicago and Los Angeles increased the relative size of the sample from these two cities. The impact of this feature is disappearing as the panel grows, but we still include it as part of our correction process.

4. Since we anticipated additional surveying in the four largest states, we double-sampled these states during January-October 2000. Similarly, the Central region states were oversampled for a brief period.

5. Certain areas of the U.S. are not serviced by MSN®. We select a smaller sample of phone numbers in those areas and use other Internet Service Providers for Internet access of recruited households in those areas.

6. As of October 2001, we began oversampling minority households (Black and Hispanic) to increase panel capacity for those subgroups.

We will focus the analyses in this paper on four of the differential sampling features listed above (1, 4, 5 and 6) since they have the biggest impact on the sample composition, coverage, variance, cost and bias issues.

B. Preparation of Final Weights for Individual Internet Studies

Once the samples are drawn, assigned, and the data returned from the field, we subject the final respondent data to a poststratification process to adjust for variable nonresponse and noncoverage. Demographic distributions from the most recent

Current Population Survey data are used as benchmarks in this adjustment. A separate nonresponse adjustment to reduce the effects of differential nonresponse for the individual survey is applied on a survey by survey basis prior to poststratification to independent benchmarks.

The primary purpose of a poststratification adjustment to CPS data is to reduce the sampling variance for characteristics highly correlated with known demographic and geographic totals – called population benchmarks. Benchmark distributions for variables such as race, ethnicity, education, age, and region using the latest CPS data. Comparable tables are prepared using the completed cases from the individual Study. Since the sample sizes are typically too small to accommodate a complete crosstabulation of these variables, we apply iterative proportional fitting. Iterative proportional fitting ratio adjusts the sample data back to the benchmarks by iteratively fitting the sample data to the marginal distributions of the benchmark data until the sample distributions converge to the benchmark distributions. [Deming, 1943]

The purpose of implementing a separate nonresponse adjustment on sample weights of completed cases is to reduce bias associated with the fact that nonresponders to the survey may have different characteristics than responders to the survey. Nonresponse adjustment is implemented using data known about those initially selected to receive the survey. Within each of this cross-classified cells, the sample weight for each case is multiplied by the ratio of assigned cases to completes cases.

The final steps to produce the weights include an examination of the distribution of the final weights to identify outliers, truncation of outliers at the tails, and a ratio adjustment of the weights back to the completed sample size.

III. Address listed households

Telephone numbers we are able to find addresses for are sent an advance letter informing the household of the opportunity to join the panel and the fact that they will be telephoned. The remaining, unmatched numbers are subsampled in order to reduce costs. The cost reduction comes from the fact that recruiting is more successful when we are able to send an advance letter. Thus, we have to sample and call fewer households. By oversampling the addressed listed phone numbers we increase our recruiting rates for the same sample size, thus reducing recruiting costs.

The sample yield from recruiting is 3 times higher for address listed phone numbers as compared to phone numbers without an address. Also, the response rate is 60% higher for address listed phone numbers as compared to phone numbers with no known address.

Early RDD samples of non-address listed phone numbers were subsampled at a rate of 1 in 2. More recent RDD samples of non-address listed phone numbers were subsampled at a rate of 1 in 3.

Table 1 presents the differences in selected demographic characteristics between address listed and non address listed groups. Race/ethnicity, income and persons age 65+ show the largest differences.

Table 1. Characteristics of Address Listed and Non Address Listed Households

Characteristic	Address Listed (60%)	Non Address Listed (40%)
Age 18-24	11%	17%
Age 35-44	20%	24%
Age 65+	15%	6%
Midwest Region	23%	20%
Income < 25K	15%	18%
Income > 75K	20%	13%
White	82%	68%
Black	11%	21%
Hispanic	10%	17%

Table 2 presents the effect of this oversampling on the design effect of six characteristics. Clearly, the variable less than high school is the most severely affected.

Table 2. Change in DEFF due to Oversampling Address Listed Households

Characteristic	Percent Change in DEFF
Black	+0.7%
Hispanic	+0.9%
Age 18-29	+1.0%
Age 45=59	+1.2%
Less than High School	+10.7%
Income < $25 K	+0.8%

IV. Areas of U.S. not serviced by MSN TV

Certain areas of the U.S. are not serviced by MSN® -- mostly rural areas. We select a smaller sample of phone numbers in those areas and use other Internet Service Providers for Internet access of recruited households. The subsampling provides the necessary coverage for those areas in the U.S. that otherwise would be missed. Paying for out of Network ISP service is more expensive than in MSN covered areas, so the subsampling helps reduce the overall cost of including the out of network areas.

Table 3 summarizes the change in the design effect (DEFF) for selected characteristics and the mean square error (MSE) due to subsampling Non-MSN Covered Areas. Variances were calculated using Jackknife variance estimation on the sample before and after the subsample of non-MSN covered areas were added.

Table 3. Change in Design Effects and MSE due to Undersampling Non-MSN Areas

Characteristic	Percent Change in DEFF	MSE
Black	+1.0%	Smaller
Hispanic	+1.2%	Smaller
Age 18-29	+1.2%	No Change
Age 45-59	+1.1%	No change
Less than High School	+0.2%	No Change
Income< $25K	+1.4%	No Change

The percent change in the design effect is very small and the MSE was statistically lower (at the 90% confidence level) for Black and Hispanic subgroups.

We will continue to include a subsample of households in these areas to ensure panel coverage of the U.S.

V. Households in Central region and 4 other states

In anticipation of several large projects, we increased the sample size for the four largest states, double-sampling these states during January-October 2000. Similarly, the Central region states were oversampled for a brief period. Table 4 below summarizes the proportion of the sample in those areas as compared to CPS estimates.

Table 4. Households in Midwest and 4 Largest States

Area	KN Unweighted	CPS Estimates
Midwest States (11)	27.0%	18.5%
Florida	6.4%	5.6%
New York	6.7%	6.6%
California	13.9%	12.5%
Texas	7.5%	7.5%
All other States	38.5%	49.3%

Table 5 summarizes the percent change in the design effect for selected characteristics as a result of oversampling these states. Design effects increase on the order of 5% for most characteristics, with one decrease of 4.9% for the less than high school group. Overall, this results in more variance increase than we would like to have in the panel going forward.

Table 5. Change in DEFF due to Oversampling Central Region and 4 States

Characteristic	Percent Change in DEFF
Black	+5.2%
Hispanic	+6.9%
Age 18-29	+4.7%
Age 45-59	+4.9%
Less than High School	-5.9%
Income < 25K	+4.7%

Currently, we are subsampling Midwest region states to reverse the disproportionate sampling.

VI. Black and Hispanic households

Many surveys conducted by Knowledge Networks for clients require either a proportionate number of Black and Hispanic relative to the U.S. population or an oversample of these minority groups. In order to boost the panel size for these subgroups, we began to oversample households with Black and Hispanic members. The methodology is as follows:

1. For each of the 9 census regions, stratify all exchanges in the region into two strata: one strata that has a higher than average proportion of Blacks and Hispancis and one strata that has a lower proportion of Blacks and Hispanics. The exchange information comes from Marketing Systems Group (RDD vendor) which uses aggregate Block level data from the Census to estimate the proportion of the population in an exchange that is Black and Hispanic. The ratio of the proportion of Blacks and Hispanics in the high minority strata to the proportion of Blacks and Hispanics in the low minority strata ranges from 2 to 8 depending on the Census region.
2. Differential sampling rates are applied to the strata to select the RDD sample, with the ratio of the sampling rates of the high minority strata to the low minority strata equaling 2.0
3. The panel weights are adjusted to reflect the differential sampling rates in the RDD sample generation.

We began the oversampling in September of 2001. Figure 1 below summarizes the success in sampling and recruiting an oversample of Black and Hispanic members. The bar groups include panel sizes for Black, Hispanic, Income < $20K, and < High School for different recruit periods: As of Sept 2001, Jan 2002, August 2002 , a projection for Dec. 2002 and only the replicates that include the oversample of Black and Hispanic households.

It is clear the oversampling is working. We recruited approximately 29% more Blacks (14.3% compared to

11.1%) and 43% more Hispanics (13.3% compared to 9.3%) than in the previous replicates of 2001. We also recruited more members with a less than High School Diploma and/or with household income less than $20,000 (23% and 37% more respectively).

One concern we had was whether the oversampling would exacerbate the oversample in the KN panel for the Midwest region. This does not appear to be a problem. The oversample replicates tend to result in an undersample of the Midwest region, which will reduce the Midwest oversample in the panel over time.

Table 6 presents the effect thus far in the variance of selected estimates.

Table 6. Change in Variance Estimates due to Black and Hispanic Oversampling

Characteristic	Percent Change in Variance
Black	**+1.0%**
Hispanic	**+7.5%**
18-29	**+0.6%**
45-59	**+3.9%**
Less than High School	**+0.4%**
Income < $25K	**+1.4%**

There are overall increases in the variance for Hispanics that is higher than expected. We need to look at estimates and variances of subpopulations within the Hispanic population to assess how well it is working or not working. We also want to investigate splitting out the oversample strata of Black and Hispanic households into independent strata of Blacks and Hispanics and sampling them at different rates. The current approach doesn't appear to be optimal for the Hispanic population.

It will take approximately 6 months more of oversampling at the current rate to show up significantly in the KN panel distributions. We plan to continue the oversampling of Black and Hispanic households and undersample in the Midwest region.

VII. Conclusions and Recommendations

Having a panel sample with proportionate allocation of sample across major demographics is extremely important for the majority of Knowledge Networks clients. To the extent that cost is not a factor or certain groups are not subject to under or over coverage, we will continue to strive to maintain a proportionately allocated panel. Thus, we are reversing the oversample in the Midwest region. Clearly the variance increases from this differential sampling of the panel are the most severe.

To contain costs, we will continue to undersample non-MSN covered areas and subsample phone numbers without address matches.

To improve the proportion of two important subgroups, Blacks and Hispanics, we will continue oversampling. However, we need to re-assess the stratification scheme currently in use and calculate sampling rates that are optimum under the new stratification scheme.

Figure 1. Increases in Panel Sample Size due to Black and Hispanic Oversampling