Development of Interviewer Performance Standards for Six National Surveys

Chandra Erdman
Center for Statistical and Methodology
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233

September 2012

Prepared for the 23rd International Workshop on Household Survey Nonresponse
Statistics Canada, Ottawa, Ontario, Canada

1 Objective

Determine model-driven expectations for response rates in the U.S. Census Bureau's six largest surveys in order to establish scientifically defendable interviewer performance standards.

2 Background

The U.S. Census Bureau's national surveys include the American Community Survey (ACS), the Current Population Survey (CPS), the Consumer Expenditure Survey (Diary: CED and Quarterly: CEQ), the Survey of Income and Program Participation (SIPP), the National Health Interview Survey (NHIS), and the National Crime and Victimization Survey (NCVS). It is evident from their names that these surveys are diverse in content and level of privacy, and this leads to variation in response rates. They are also diverse in sample size, sampling design, and geographic coverage, among other attributes.

One thing that these surveys have in common is that they are conducted (in at least one phase of sampling) via Computer Assisted Personal Interviewing (CAPI), and interviewers are expected to achieve response rates that vary with the level of "urbanicity" of the county in which the majority of their cases resides. More specifically, each county is designated as either "MA," for metropolitan areas which are considered to be very difficult, "A," for urban, moderately difficult areas, "B," for low-density urban to suburban areas that are less difficult than higher-density urban areas, or "C, D, or E," for rural areas that are thought to be relatively easy.

The original stratification system was based solely on population density.[1] Following Census 2000, a discriminant analysis was performed to determine whether areas needed to be reclassified, and to suggest strata designations for areas that were not stratified in 1990. The following variables were used in the discriminant analysis.

1. Percent of vacant housing units.
2. Percent of units in structures containing 2 or more housing units.
3. Percent of units in structures containing 10 or more housing units.
4. Percent of occupied units with more than 1 person per room.
5. Percent of occupied units where the householder moved into the unit in 1999-2000.
6. Median property value.
7. Percent of households that are not husband/wife families.
8. Percent of households with public assistance income.
9. Percent of linguistically isolated households.
10. Percent of households with 4 or more members.
11. Percent of households with journey to work times greater than 60 minutes.
12. Percent of households with householder less than 25 years old.
13. Percent of one person, single male households.
14. Percent of single person households.
15. Percent of population below poverty level.
16. Percent of population who are minorities.
17. Percent of foreign born persons.
18. Percent of population greater than or equal to 65 years.
19. Percent of population who are college students.
20. Percent of the population who are Hispanic or Latino.
21. Percent of group quarters population.
22. Census 2000 mail response rate.

Once all counties were assigned strata, management of the twelve regional offices (ROs; Atlanta, Boston, Charlotte, Chicago, Dallas, Denver, Detroit, Kansas City, Los Angeles, New York, Philadelphia, and Seattle) determined, for each survey and strata, ranges of response rates corresponding to five performance levels.

[1] As far as I can tell!

Interviewers who obtain level 1 response rates are given Performance Improvement Plans (and may be terminated if response rates do not improve), while interviewers achieving level 4 and 5 response rates are awarded bonuses.

An example of the variability in performance standards for the Current Population Survey is given in Table 1. For this survey, the Boston regional office combines the two most rural strata, the Kansas City office combines the suburban and the least rural of the rural strata, but only varies the cutoffs for ratings of 4 and 5, and the Los Angeles office combines all strata. Interviewers working on CPS in all areas served by the Los Angeles RO must essentially achieve perfect response rates in order to receive the highest performance rating, while interviewers working in metropolitan areas served by the Boston RO need response rates of just ninety-two percent to be rated at the highest level.

Table 1: CPS Performance Standards for Three Regional Offices

	Performance Rating				
Stratum	**1**	**2**	**3**	**4**	**5**
Boston Regional Office					
MA	0 - 83.00	83.00 - 83.90	84.00 - 88.90	89.00 - 91.90	92.00 - 100
A	0 - 84.00	84.00 - 85.90	86.00 - 89.90	90.00 - 92.90	93.00 - 100
B	0 - 85.00	85.00 - 86.90	87.00 - 90.99	91.00 - 92.90	93.00 - 100
C	0 - 86.00	86.00 - 87.90	88.00 - 91.90	92.00 - 93.90	94.00 - 100
D,E	0 - 87.00	87.00 - 88.90	89.00 - 92.90	93.00 - 94.90	95.00 - 100
Kansas City Regional Office					
MA	0 - 91.00	91.00 - 93.24	93.25 - 95.99	96.00 - 98.49	98.50 - 100
A	0 - 91.00	91.00 - 93.24	93.25 - 96.99	97.00 - 98.49	98.50 - 100
B,C	0 - 91.00	91.00 - 93.24	93.25 - 97.49	97.50 - 98.99	99.00 - 100
D	0 - 91.00	91.00 - 93.24	93.25 - 97.99	98.00 - 98.99	99.00 - 100
E	0 - 91.00	91.00 - 93.24	93.25 - 98.49	98.50 - 98.99	99.00 - 100
Los Angeles Regional Office					
ALL	0 - 88.51	88.51 - 90.50	90.51 - 94.50	94.51 - 99.50	99.51 - 100

Table 1 shows the variability in performance standards for just one survey and three regional offices. There is a great deal of variability in the standards for all surveys and managing offices, and even more variability in respective performance. Table 2 displays the average difference between actual response rates from the first half of 2012 and minimum acceptable response rates (maximums of level 1 intervals) by survey and RO. In all regions, CEQ response rates are at least 10% higher than the minimum acceptable rates, on average. These rates receive a level 3 rating in the Kansas City region where the minimum is 75%, and a level 5 rating in the New York region where the minimum is 55%. In contrast, the observed SIPP response rates are less than acceptable in all regions. The SIPP minimum acceptable rates range from 67% in the Seattle region to 84% in the Charlotte region.

Regional performances across surveys (shown in the "Average" column of Table 2) are above the minimums for most regions. However, there are still issues with the current performance standards, the most glaring of which is that they are not consistent across the nation. The performance standards of a highly urban county on the east coast are likely to be different from the performance standards of a highly urban area on the west coast, even if it is equally difficult to obtain interviews in each county. A less apparent issue is that within many counties there is significant variability in demographics and other characteristics that are known to be related to survey response propensities. The population of Los Angeles county, for example, is greater than the population of forty-two states! There are areas of Los Angeles county that are very urban, and others that are very suburban, yet all areas are treated equally under the current system. Lastly, the "current" performance standards were developed nearly a decade ago, using decade-old data. In Section 4 we discuss alternative methods for developing new performance standards that address these issues, and in the following section we describe the data used in our analyses.

Table 2: Weighted Performance Rates by RO and Survey (January-June, 2012)

	ACS	CED	CEQ	CPS	NCVS	NHIS	SIPP	Average
Atlanta	6.03	4.26	14.92	3.64	8.62	1.21	-2.01	4.73
Boston	3.15	11.31	20.47	2.24	-2.45	-5.66	-7.98	1.30
Charlotte	3.24	7.87	11.81	-1.67	-0.72	11.28	-11.80	-0.15
Chicago	4.14	6.12	14.12	-0.32	5.73	1.53	-11.33	0.88
Dallas	4.97	17.33	25.44	2.74	2.75	-1.13	-3.96	3.74
Denver	1.19	13.18	24.08	0.75	-1.91	-5.28	-13.80	0.01
Detroit	5.66	10.68	22.29	2.58	1.89	-3.67	-2.47	3.38
Kansas City	5.45	16.54	10.59	1.65	-5.25	-2.90	-9.87	1.16
Los Angeles	0.49	15.66	24.11	0.69	2.80	-1.36	-15.85	0.15
New York	5.50	13.72	32.80	3.71	0.82	-1.24	-4.27	4.03
Philadelphia	3.60	3.86	14.31	1.67	-1.95	-0.20	-12.74	0.68
Seattle	-0.82	20.18	25.46	-0.30	-2.03	3.07	-5.37	-0.11

3 Data

The Planning Database (PDB) was developed to identify reasons that people are missed in decennial censuses. The first PDB was compiled from 1990 Census data and contained tract-level characteristics of households and persons that are associated with response rates (Bruce and Robinson, 2003, 2007). The 1990 PDB was used in the planning, implementation and evaluation of the 2000 Census and was eventually updated with Census 2000 long-form data. The 'Tract-Level Planning Database with Census 2000 Data' has been used in a number of initiatives including segmenting the nation for the 2010 Census advertising campaign (Bates and Mulry, 2011), and creating a 'hard-to-count' (HTC) score which identifies areas that are difficult to enumerate (Bruce et al., 2001).

On June 8, 2012, the U.S. Census Bureau released a beta version of the 2012 Planning Database. The 2012 PDB contains block-group-level estimates of many person and household characteristics compiled from the 2006-2010 American Community Survey and the 2010 Census, as well as several 'operational' variables that describe 2010 Census mail-back behavior. For households that did not respond to the 2010 Census via mail, we have block-level summaries of the nonresponse follow-up (NRFU) operation that include the number of contact attempts made, the number of interviews completed, and the number of interviews obtained by proxy.

4 Methods

4.1 Regression

With the wealth of information in the 2012 PDB and the supplemental NRFU data, we would like to develop a model that estimates small-area response rates for each survey. The problem is that for the vast majority of block-groups (and even many tracts) response "rates" are binary observations because there is often just a single case, and more often no cases, in each block group. One way to address the sparseness of survey cases is outlined in the following algorithm.

1. Impute synthetic cases into each block-group, with response rates equal to the expected response rate, chosen as the average midpoint of the current performance level 3 interval across regions, within survey and current strata.

2. Estimate the response rate for each interviewer, i, as the average of the model-estimated propensities of his or her cases.

3. Let $\bar{p}_i$ be the observed response rate of interviewer i, $\hat{p}_i$ be his or her estimated response rate, and n_i be the number of cases worked by him or her in the given month. Compute the performance score, S_i,

for each interviewer as

$$\frac{\bar{p}_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}}}.$$

4. Evaluate performance according to the following scale.

Performance Rating	1	2	3	4	5
Performance Score	< -1.5	$[-1.5, 0.5)$	$[-0.5, 0.5]$	$(0.5, 1.5]$	> 1.5
Theoretical Distribution	6.7%	24.3%	38%	24.3%	6.7%

5. Update the model each month with real case outcomes and repeat.

Another way to address, or circumvent, the sparseness of survey cases is to model 2010 Census response rates in lieu of survey response rates, and adjust the predictions for each survey based on historic differences in cooperation. To determine the combination of available variables that are most predictive of block-group-level 2010 Census response rates, we perform an exhaustive search for the best subset of up to twenty predictors using the method implemented in Lumley and Miller (2009). A summary of the resulting model is shown in Table 3.

Table 3: Predicting Census 2010 Block-Group Response Rates

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	72.7544	0.1971	369.18	0.0000
Percent <5	0.1716	0.0146	11.76	0.0000
Percent 5-17	0.0375	0.0023	16.30	0.0000
Percent 18-24	-0.1496	0.0135	-11.12	0.0000
Percent 65+	0.8552	0.0136	62.78	0.0000
Percent Hispanic	-0.6394	0.0119	-53.61	0.0000
Percent White	0.0338	0.0007	49.17	0.0000
Percent Black	-0.5316	0.0093	-57.33	0.0000
Percent American Indian, Alaskan	-0.4655	0.0132	-35.21	0.0000
Percent College Graduate	0.3080	0.0082	37.61	0.0000
Percent Below Poverty	-0.1155	0.0097	-11.92	0.0000
Percent Married	0.0347	0.0012	29.12	0.0000
Persons/Household	-1.3234	0.0398	-33.28	0.0000
Percent Moved in 2006-2009	-0.0157	0.0010	-15.10	0.0000
Percent Vacant	-3.0041	0.0289	-103.80	0.0000
Percent Renter	-0.0466	0.0010	-45.66	0.0000
Percent Single Family Home	0.0495	0.0008	65.86	0.0000
Percent 2-9 Unit Structures	-0.0541	0.0010	-56.29	0.0000
Percent No Phone	-0.0313	0.0025	-12.64	0.0000
Percent No Plumbing	-0.1140	0.0031	-36.49	0.0000
Population Density	0.1953	0.0087	22.44	0.0000
Percent NRFU Interviews by Proxy	0.1440	0.0014	103.81	0.0000
Residual standard error: 5.593 on 214,569 degrees of freedom				
Multiple R-squared: 0.5653, Adjusted R-squared: 0.5653				
F-statistic: 1.329e+04 on 21 and 214,569 DF, p-value: < 2.2e-16				

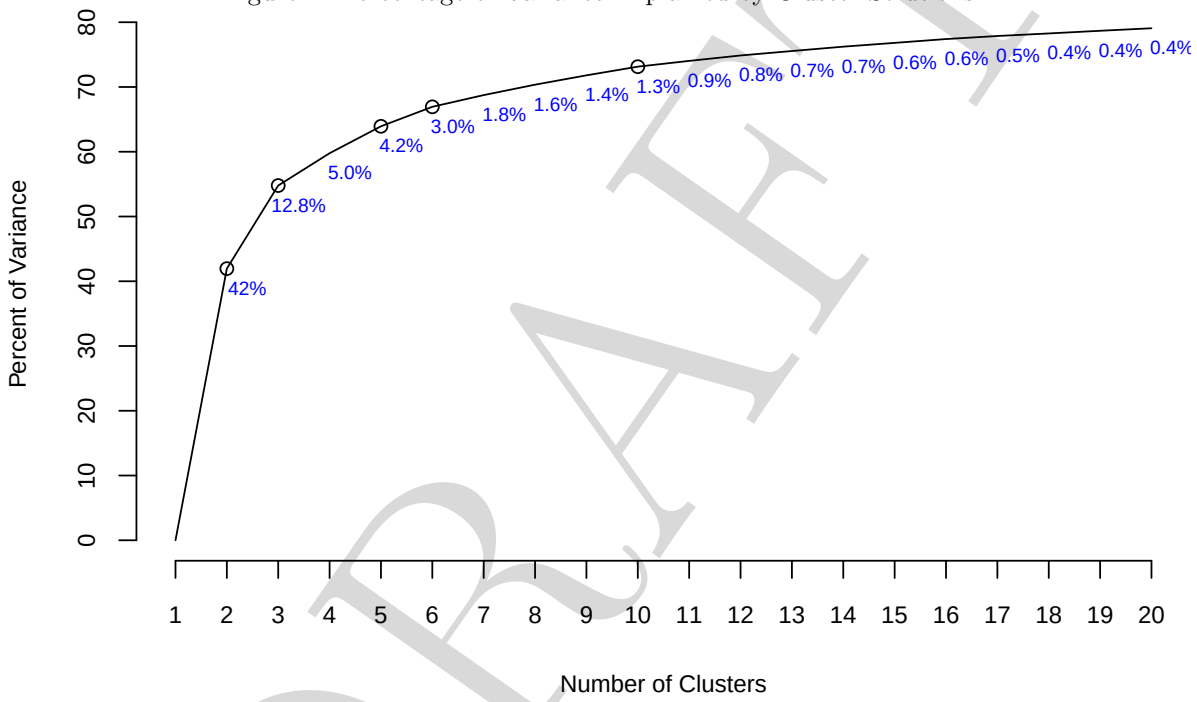
Note: Most predictors are either log or square root transformed.

4.2 Cluster Analysis

As an alternative to regression-based models, we segment the country on the predictors from Table 3 using the k-means algorithm of Hartigan and Wong (1979). There are several methods for determining the optimal number of clusters, k, in a k-means analysis. In this particular problem, the number of block-groups to be

clustered is greater than 200,000, and methods that depend on the number of observations such as Hartigan's rule of thumb (Hartigan, 1975) and choosing $k \approx \sqrt{n/2}$ (Mardia et al., 1979) suggest a number of clusters that is too large ($k > 100$) for our purposes. As such, we employ the "elbow method," which examines a plot of a range of k versus the percentage of variability explained by each cluster solution to look for a point at which the marginal gain drops and creates an elbow (Thorndike, 1953). If W_k is the within cluster sums of squares for k clusters and B_k is the between cluster sums of squares, the percentage of variability explained by the cluster solution is $W_k/(B_k + W_k)$. This statistic is plotted in Figure 1, for cluster solutions with $k = 1, 2, \ldots, 20$, with the increase in variability explained by each additional cluster shown in blue. There are obvious elbows at $k = 2$ and 3, and less visible elbows at $k = 5, 6$ and 10.

Figure 1: Percentage of Variance Explained by Cluster Solutions.



5 Points for Discussion

1. In a cluster-based solution, what is the best way to set expected response rates for each cluster by survey by performance level?
2. Weaknesses of each proposed method and potential improvements.
3. Alternative priors for expected response rates in the regression-based solution.
4. In the regression models, should we include an indicator for region? There is a strong desire to move toward one standard for the nation. However, the demographic and other variables that we have may not reflect regional differences in attitudes toward government, for example.
5. In the performance score, S, how "bad" is the approximate variance, given that the proportions are averages over distributions with varying rates of success, rather than means of *i.i.d.* binomial observations.

References

Nancy Bates and Mary H. Mulry. Using a geographic segmentation to understand, predict, and plan for census and survey mail nonresponse. *Journal of Official Statistics*, 27(4):601–618, 2011.

Antonio Bruce and J. Gregory Robinson. *The Planning Database: Its Development and Use as an Effective Tool in Census 2000.* Paper presented at the Annual Meetings of the Southern Demographic Association, Arlington, VA, October 24, 2003.

Antonio Bruce and J. Gregory Robinson. *Tract Level Planning Database with Census 2000 Data.* U.S. Government Printing Office, Washington, DC, 2007.

Antonio Bruce, J. Gregory Robinson, and Monique V. Sanders. Hard-to-count scores and broad demographic groups associated with patterns of response rates in census 2000. *Proceedings of the Social Statistics Section, American Statistical Association*, 2001.

John A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

John A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Thomas Lumley and Alan Miller. *leaps: regression subset selection*, 2009. URL `http://CRAN.R-project.org/package=leaps`. R package version 2.9.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Academic Press, 1979.

Robert L. Thorndike. Who belongs in the family. *Psychometrika*, pages 267–276, 1953.