Sample Design and Weighting

The American Panel Survey (TAPS)

Weidenbaum Center Washington University

The American Panel Survey (TAPS) is designed to be representative of the U.S. population of adults. Survey results, properly weighted, are generalizable within a known margin of error. The sample design and the recruitment methods are based on the GfK-Knowledge Networks (GfK-KN) experience using residential address samples for the mail recruitment of GfK-KN KnowledgePanel members. TAPS is designed to have approximately 2,000+ members by recruiting one person per household through a mail sample. The frame for the sample of addresses is the U.S. Postal Service's computerized delivery sequence file (CDSF). Marketing Systems Group (MSG), is a sample vendor licensed to work with this file and from whom this sample was purchased. The CDSF covers some 97% of the physical addresses in all fifty states including P.O. boxes and rural route addresses. Homes that are vacant or seasonal are identified as are other categories that help to refine the efficiency of the sample to be mailed. Using data from available U.S. Census files plus from a variety of commercial data bases, such as White Pages, Experian, Acxiom, etc., MSG adds names to these addresses, match with landline telephone numbers, and with some level of accuracy append information regarding race/ethnicity, age of householder, whether there are people of a certain age in the household, presence of children, home ownership status, etc. It should be pointed out that there is also some proportion of missing information, for example unknown age, in these databases. In order to have better control over the response of more difficult groups to recruit, the sample can be stratified using this appended or "ancillary" information. During 2010, GfK-KN experimented with this ancillary information to understand how well it predicts based on actual mail recruitment data used with KnowledgePanel. The GfK-KN stratification used for KnowledgePanel was used for the TAPS recruitment sample.

Stratification

The sampling strata are designed to specifically target young adults (ages 18-24) and also Hispanic persons in addition to the balance of the population. In this way, young adults and Hispanics have been modestly oversampled to offset their known tendency to under-respond to surveys. Because age and Hispanic ethnicity are not mutually exclusive groupings, the strata are classified as follows:

1. 18-24 year-old Hispanic adults
2. All other Hispanic adults ages 25+ or age unknown
3. 18-24 year-old non-Hispanic adults
4. All other adults that are non-Hispanic or ethnicity unknown and ages 25+ or age unknown

Mailing Size

Using estimated yield and profile rates, Table 1 shows the size of the mailing for each stratum (column A), the yields (column B) and the expected resultant sample sizes (column C). A modest oversample of the two young adult strata (#1 and #3) was fielded. This is shown in column D showing the sample distribution compared to the distribution of the sample frame (data provided by MSG).

Table 1. Recruitment Sample Design

	Stratum	A. Mailing		B. Yields		C. Profiled		D. Strata Distributions	
		count	*distribution*	*proportion*	*count*	*proportion*	*count*	*sample*	*frame*
1	Hispanic 18-24	333	1.0%	0.056	19	0.65	12	0.6%	0.2%
2	Hispanic 25+/unk	6,943	19.9%	0.064	445	0.65	289	14.4%	14.1%
3	Other 18-24	469	1.3%	0.144	68	0.65	44	2.2%	0.7%
4	All Else 25+/unk	27,071	77.8%	0.094	2,550	0.65	1,657	82.8%	85.0%
	Overall	**34,816**			**3,081**		**2,003**	100.0%	100.0%

Design Effect

Based on the strata population distributions in the CDSF frame tagged with ancillary information, the sample distribution to recruit exactly 2,003 panel members (see column D in Table 1 above) has a very low design effect of 1.04. This is due to the very mild oversampling of young adults. A perfect simple random sample has a design effect of 1.00. The actual design effect will depend on the reported demographic information from the profiled sample.

Stratum weight

Due to the stratified sample design, cases from each stratum will be adjusted in their base weight in order to return the actual mailed sample distribution (n=38,000, design effect 1.07) to the same distribution as existing in the frame. This corrects cases for the stratum-specific selection probability associated with the sample design. Table 2 shows the relevant stratum selection probabilities and weights necessary to make this adjustment. The stratum weight is called weight1 (⍰**1**).

Table 2. Recruitment Stratum Weight

	Column:	*a*	*b*	*c*	*d*	*e*
	Stratum	**Frame %**	**Mailed sample count**	**Mailed sample %**	**Selection probability** (*c/a*)	**Stratum weight** (1/*d*)
1	Hispanic 18-24	0.2%	363	1.0%	5.31	0.1884
2	Hispanic 25+/unk	14.1%	7,578	19.9%	1.41	0.7074
3	Other 18-24	0.7%	512	1.3%	1.94	0.5162
4	Other Else 25+/unk	85.0%	29,547	77.8%	0.91	1.0934
		100.0%	**38,000**	100.0%		

Landline telephone match weight

For all addresses in the sample, MSG searched available databases to match a landline telephone number to the exact address. The telephone numbers were used to do an interviewer administrated telephone recruitment among non-responding households. Thus, non-responders with a landline match had a higher chance of being recruited due to this telephone effort. To correct for this increased probability of recruitment due to this out-bound calling effort, the final cases, weighted by **⍰1** were then corrected to reflect the original sample's proportion of landline match within stratum and across strata (see Table 3). This now corrected weight is called weight2 (**⍰2**).

Table 3. Percent telephone match by Stratum

Stratum	No Match	Match
1 Hispanic 18-24	41.7	58.3
2 Hispanic 25+/unk	44.1	55.9
3 Other 18-24	47.2	52.8
4 Other Else 25+/unk	30.9	69.1
	32.3	67.8

Eligible adults within household weight

Each household had varying numbers of eligible adults from among which one was randomly selected to be recruited onto the TAPS. Persons from households with one or two adults are more likely to be represented in the sample, especially one-adult households. To correct for this increased selection probability from among one-adult households, a weight was calculated to adjust for selection from one-adult, two-adult and three or more-adult households. In the final sample, 34.4% were from one-adult households, 52.5% from two-adult households and 13.1% from households with three or more eligible adults. [Note: 34 cases had the number of eligible adults as missing or refused so a random imputation routine was used to assign an eligible adults number to these households.] Thus, multiplying ***⍰2*** by 1, 2, or 3 corresponding for the eligible adult number, a final or weight3 (**⍰3**) results for each case.

The base weight

This **⍰3** from the above step is then scaled to sum to the total number of cases recruited on the panel. The scaled **⍰3** is now the base weight (**basewt**) for each case and will be the starting weight for the post-stratification weighting procedure.

Summary of the study base weight components

Design Effect:

⍰1: 1.0247
⍰2: 1.2362
⍰3: 1.3689
basewt: 1.3689 [this is a scaled ⍰3 so the design effect is identical to ⍰3]

Range on Weights:

```
Variable        Mean        Minimum       Maximum        N        Sum        1st Pctl      99th Pctl

⍰1          1.0315235     0.1883239     1.0989011      2128    2195.08     0.5154639     1.0989011
⍰2          1.0315235     0.1431884     1.8582137      2128    2195.08     0.4654436     1.8582137
⍰3          1.8218228     0.1431884     5.5746412      2128    3876.84     0.4960175     5.5746412
basewt      1.0000000     0.0785962     3.0599250      2128    2128.00     0.2722644     3.0599250
```

Post-stratification weighting

When the full panel is assigned a survey, respondents completing the survey will undergo a post-stratification (PS) weighting that will use each respondent's base weight as their starting weight. The purpose of this PS weighting is to make survey respondents representative of the non-institutionalized U.S. adult population. The PS weighting adjusts for non-response by weighting all completed interviews to national benchmarks. Demographic and geographic distributions for the population ages 18+ from the most recent Current Population Surveys (CPS) are used as benchmarks for this adjustment. Some benchmark distributions come from the monthly CPS estimates and some come from special supplemental CPS estimates.

A description of the post-stratification process follows, using the January 2012 TAPS survey as an example. The same process is used for each monthly TAPS survey.

Data weighting in January 2012 are weighted on the following variables and use the benchmark sources as shown:

Benchmarks Source: December 2011 CPS

- Gender (Male/Female) by Age (18-29, 30-44, 45-59, and 60+)
- Race/Hispanic ethnicity (White/Non-Hispanic, Black/Non-Hispanic, Other/Non-Hispanic, 2+ Races/Non-Hispanic, Hispanic)
- Education (Less than High School, High School, Some College, Bachelor and higher)
- Census Region (Northeast, Midwest, South, West) by Metropolitan Area (Yes, No)

Benchmarks Source: March 2011 CPS Annual Social and Economic Supplement (ASEC)

- Household Income (Under $10,000, $10,000-$29,999, $30,000-$49,999, $50,000-$79,999, $80,000-$99,999, $100,000 or more)

Benchmarks Source: October 2010 CPS Supplement, Computer Use and Access Module

- Internet Access (Yes, No)

Comparable distributions are calculated using all January completed cases (n=1,609) from the TAPS using a SAS raking procedure. This procedure adjusts the completed sample data to the selected benchmark proportions through an iterative convergence process. The weighted sample data are optimally fitted to the marginal benchmark distributions.

Two post-stratification weights

Two sets of weights were produced. One set included all of the above mentioned weighting variables and the second set excluded the Internet Access adjustment. The purpose of excluding this adjustment was to lower the design effect and reduce the range of the weights.

The final resulting distribution of each of the calculated weights were examined to identify and trim outliers (Windsorized) at the extreme upper and lower tails of the weight distribution. The final trimmed weights for each of the two sets of weights align with the benchmark distributions within a tolerance of no more than 2 percentage points. The post-stratified and trimmed weights make up the final, single study weight. This weight in the data file is called **jan2012wt1** for the adjustments that include Internet Access and **jan2012wt2** when Internet Access is excluded.

Definition of final weights

jan2012wt1 INCLUDES Internet access adjustment

jan2012wt2 EXCLUDES Internet access adjustment

Given these final weights (W), a design effect (Deff_{est}) can be calculated for each set of weights that is the ratio of the average of the squared weights to the average of the weights. The formula for that estimation is:

$$\mathrm{Deff}_{est} = [(\Sigma\ W_i^2)/n] / [(\Sigma\ W_i)/n], \text{ where } n = \text{final sample size.}$$

When using weights that are scaled to sample size *n*, this formula gets simplified to the ratio of the sum of the squared weights to the sum of the weights:

$$\text{Deff}_{est} = \Sigma\, W_i^2 / \Sigma\, W_i\,.$$

The survey design effect is used to adjust standard errors to reflect the deviation of this weighted complex sample design from those of a simple random sample.

Summary of the final weights

Trimming (at low and high percentile)**:**

jan2012wt1: (0.99%, 99.01%)
jan2012wt2: (0.87%, 99.13%)

Design Effect:

jan2012wt1: 2.4780
jan2012wt2: 1.9734

Range on Weights:

```
Variable        Minimum        Maximum        N       Sum            1st Pctl         99th Pctl

jan2012wt1   0.0645682    7.4647148    1609    1609.00       0.0667383       7.4647148
jan2012wt2   0.1204608    6.1295313    1609    1609.00       0.1288430       5.7228047
```

Notes on Imputation

Introduction

Imputation is a group of methods used to substitute plausible values for values that are missing in a data set. Missing values are rarely missing completely at random (MCAR). Missing values lead to less efficient estimations as many statistical techniques drop incomplete cases. With cases dropped, the estimates are then based on a smaller sample of respondents and, given that these dropped cases are not MCAR, this smaller subsample is likely to produce skewed estimates. Since we want to use the information contributed by all the cases in a study, as a minimum we must have no missing data in the variables to be used for weighting. With missing data, these cases cannot be weighted and thus cannot be used. For the TAPS panel members, some portion of the cases were missing some of the data essential for weighting and thus an imputation method was employed to resolve this problem so that the maximum number of cases could be used.

Hot deck imputation

Imputation for The American Panel Survey (TAPS) was undertaken using the hot deck imputation method. The hot deck imputation method is a technique where respondents with missing values are matched to respondents who have identical values on other, correlated variables. If more than one match is found, the matching respondent is chosen at random. The missing value is then replaced with the value of the variable given by the matched respondent. This hot deck imputation was achieved using SOLAS for Missing Data Analysis software.

Panel data imputation

The following variables contain imputed values in the original dataset: Hispanic ethnicity, race, age categorization, labor force status, marital status, housing ownership, gender, education and income. Respondents who had missing information on all, or all but one, of the variables used for weighting were dropped from the dataset (6 cases). Table A indicates the number of missing values once these 6 cases were dropped (n=2,128). The variables identified with an asterisk are those used in weighting.

The variables used to match respondents for each of the imputed variables are given in Table B. The matching is done in the order in which variables are presented in the table. For example, the age categorization is matched first on parental status and then on being a student.

A series of variables are included in the dataset to inform the user whether the respondent has an imputed value on any particular dimension. These variable names begin with “ii”. Similarly, the “oo” series of variables give the original values of the variables.

Table A. Frequency of missing values in variables selected for imputation

Variables Imputed	# of Missing Values	% of Missing Values
Hispanic Ethnicity	17	0.8%
Race	35	1.6%
Age Categorization	53	2.5%
Labor Force Status	341	16.0%
Marital Status	554	26.0%
Home Ownership	25	1.2%
Gender	3	0.1%
Education	17	0.8%
Income	158	7.4%

Table B. Variables used in hot deck matching process

Variables Imputed	Variables Used to Match Respondents			
Hispanic Ethnicity*	Sampling Strata			
Race*	Hispanic Ethnicity			
Age Categorization*	Parental Status	Student Status		
Labor Force Status	Age Categorization			
Marital Status	Age Categorization	Parental Status		
Housing Ownership	Marital Status			
Gender*	Labor Force Status			
Education*	Labor Force Status			
Income*	Education	Home Ownership	Marital Status	Labor Force Status