Table of Contents

FAB Analytic Guidance Document

1. Background

In 2005, a draft version of the Food Attitudes and Behaviors (FAB) Survey was developed. Several rounds of cognitive interviewing were conducted, modifications were made, and psychometric testing was conducted in a pilot study to identify distinct correlates of fruit and vegetable intake. Based on the pilot, extensive changes were made (e.g. modification of survey items) and a final version of the FAB Survey was implemented between September and December 2007. The purpose of the study was to assess the strongest correlates of fruit and vegetable intake among adults in the US. A total of 5,705 questionnaires were mailed out for this survey, of which 3,397 were returned with useable data for analysis.

2. Survey Design and Sample Selection

2.1 Introduction

The sample was drawn from Synovate's Consumer Opinion Panel (COP). The Synovate panel includes 450,000 households, which represent over one million U.S. respondents. Respondents are invited to join the Synovate panel through the direct mailing of recruitment surveys. Approximately 25% of the panel is replaced annually and the average length of tenure is 5.9 years. On average, panel households are contacted about once a month. Demographic variables available in the panel include age and gender (all household members); race, Hispanic origin, education level, employment status, and occupation (head of household); and geographic region, state, county code, MSA code, population density, household income, household size, dwelling type, and home ownership (household). For the current study, inclusion criteria required that the participants be at least 18 year of age and English-speaking.

2.2 Sample Size and Stratification and Response Rate

The target sample size was 3,600 overall, with an allocation of 25 percent African-Americans, with the remaining 75 percent being White and other races. Other factors listed above were represented in proportion to the general US household population.

Table 1 shows the stratification for the survey, along with target sample sizes. The stratification was initially based on race, age, education, and gender, with specific cells defined using a CHAID analysis with data from a pilot data. This analysis derived a set of cells with maximum differences in nonresponse rates, as measured by chi-square statistics.

The target sample sizes were 900 for African Americans and 2,700 for White/other race. Based on the pilot data, we assumed an initial response rate of 60%, with some variations as shown in Table 1. However, a postcard prompt was utilized to increase the response rate to between 63% and 65%; adjustments made to the pilot response rates to account for the postcard prompt are also shown in Table 1. Adjusting for expected response rates, the overall initial sample size was 5,705, consisting of 1,645 African Americans and 4,060 of White/other race, respectively. A total of 6,003 surveys were mailed (5,803 initial mailing and 200 additional mailing) and 3,418 were returned for a response rate of 57 percent.

2.3 Sampling Methods

The sample selection process, carried out by Synovate, consisted of several steps. Initially, Synovate drew a sample from their COP using five "balancing" factors: Region, Income, Age, Household Size, and Population Density. These factors are defined as follows:

- **Region** is based on the nine standard Census divisions: New England; Middle Atlantic; East North Central; West North Central; South Atlantic; East South Central; West South Central; Mountain; Pacific.

- **Household Income** is divided as Under $20,000; $20,000- $39,999; $40,000 - $59,999; $60,000 - $99,999; and $100,000 or more, per year.

- **Population Density** is classified as Non-Metropolitan Statistical Area (MSA); MSA of population up to 0.5 million; MSA of 0.5 to 2.0 million; and MSA with 2.0 million or more.

- **Age**: Under 30; 30-39; 40-49; 50-59; and 60 or older.

- **Household Size**: 1 member; 2 members; 3 members; 4 members; 5 or more members.

This sample was selected so that households in the COP sample have the same proportion in each region as the US household population, the same proportion of households by income as the US household population, and so forth.

The size of this initial sample was based on the population prevalence of the smallest stratum and the expected response rate for respondents in this cell. For example, African Americans aged 18 to 34 years with less than high school education have the lowest prevalence (i.e., lowest proportion of the population) and the lowest response rate; these are 0.8% and 18% respectively. Thus, the initial sample was large enough to contain at least 63 expected respondents in this group.

The second step was to select samples for each of the strata shown in Table 1. These samples were selected randomly across the five balancing factors, so that each of the samples was representative of the initial sample with respect to the five factors. For example, the sample drawn for FAB of 18-34 year old African Americans with less than a high school education had the relative frequency by income as the initial sample, which (as noted above) had the same relative frequency by overall income as the US household population.

In all, a sample of 5,705 cases was drawn from the initial balanced sample using the stratification and sample sizes specified in Table 1. This initial release was followed by a postcard reminder sent to all persons in the initial sample.

3.0 Calculation of the Weights

The stratification shown in Table 1 was combined with income to create weighting cells. Groups with approximately 200-400 respondents were further post-stratified into two subgroups by whether respondents were above or below the median US income. Two groups outside this range (one with 191 respondents and one with 405) were also divided into two subgroups. If there were more than 400

respondents, the group was divided into three subgroups (below 33rd percentile, between 33rd and 67th percentile, above 67th percentile) or four subgroups (using quartiles).

Because of uncertainty in the true sample counts in the original cells used for sampling[1], the usual step of creating nonresponse adjustments was by-passed[2]. Instead, post-stratification adjustments were created using 2007 Current Population Survey (CPS) data. Using this method, sampling weights were calculated by comparing sample counts for each weighting cell to the Census population counts for the cell, basing the sample counts on the survey responses rather than the original stratification assignment, except in cases of missing data, where the original sampling assignment was used.

Table 2 shows the weighting cells and the resulting weights. There are 21 cells, each with a corresponding weight. For example, the weight for African Americans aged 18-34 with less than high school education was calculated as:

$$W = \frac{\text{Census count for Afr. Am., age } 18-34, < \ HS\ educ.}{\text{Sample count for Afr. Am., age } 18-34, < \ HS\ educ.}$$

$$= \frac{1{,}633{,}460}{64} = 25{,}522.8.$$

As noted above, the counts for the denominator were based on questionnaire responses. These weights were then assigned to sample respondents based on their questionnaire responses. That is, all respondents who reported being African Americans aged 18-34 with less than a high school education were assigned the sampling weight 25,522.8.

Table 2 gives the strata used to form the weights, the population for each stratum, the sample count for each stratum, and the resulting weight. In addition, the table shows the mean, standard deviation, and coefficient of variation (CV) for the weights. The CV can be used to estimate the design effect for unequal weighting, through the formula:

$$\text{design effect} = 1 + CV^2.$$

The design effect is 1.21, meaning that the effective sample size is reduced by about 21% by weighting to adjust for disproportionate sampling and nonresponse.

4.0 Using the Weights

The weights shown in Table 2 can be used in SAS, SUDAAN, Stata, or other programs for analyzing survey data. Weights are typically used when analyzing survey data to calculate frequency counts, means, percents, and other statistics. The weights should be used with software designed for analysis of survey data, such as SUDAAN, Stata, or the survey procedures in SAS v9.2.

[1] While questionnaires were mailed to selected panel members within households, there were instances where another person in the household actually completed the questionnaire and returned it. This practice was generally accepted in the Synovate panel.

[2] For more detail on weighting for nonresponse and poststratification, please refer to: Bethlehem, J.G. (2002), "Weighting Nonresponse Adjustments Based on Auxiliary Information" in *Survey Nonresponse*, ed. R.M. Groves, D.A. Dillman, J.L., Eltinge, and R.J.A. Little , John Wiley and Sons, New York, pp. 275-287.

Examples from each of these software packages are given below for a single-stage, stratified, with replacement design. The examples estimate proportions and totals with standard errors and confidence intervals for two-way tables using the Taylor series linearization method of variance estimation. The final weight (WEIGHT) and stratification variable (SAMPLEGROUP) have been placed on the delivery files.

SAS v9.2

SAS v9.2 features a number of procedures for survey data analysis, including SURVEYFREQ, SURVEYMEANS, SURVEYREG, and SURVEYLOGISTIC. These procedures implement data analysis for frequencies and cross-tabulations, means and proportions, linear regression, and logistic regression. The example below shows how to use SURVEYFREQ for the FAB design to obtain frequencies of age (q57), sex (q56), and race:

```
proc surveyfreq;
  weight weight; /* final person weight */
  strata samplegroup; /* stratum variable */
  tables q56 q57 race; /* fill in variable names as necessary */
run;
```

The critical parts of this code are the WEIGHT and STRATA statements. These statements are used in all the SAS survey analysis procedures named earlier.

SUDAAN v10

SUDAAN provides a wide range of procedures for analyzing survey data, including CROSSTAB (frequencies and cross-tabulations), DESCRIPT (for means and totals), RATIO (for estimates of ratios), REGRESS (linear regression), and LOGISTIC (logistic regression). The example below shows how to use CROSSTAB:

```
proc crosstab design=strwr;
 subpopn <subset criteria>; /* to produce subpopulation estimates */
  weight weight;
  nest samplegroup/missunit;
  subgroup q56 q57 race;
  levels 2 3 4;
  tables q56 q57 race; /* fill in variable names as necessary */
run;
```

The critical components of this code are the WEIGHT and NEST statements, along with the “DESIGN =“ specification in the PROC CROSSTAB statement. These statements are used in all the SUDAAN analysis procedures indicated earlier. **Note:** All SUDAAN variables must have numerical (not character) values and the dataset must first be sorted by the stratum variable (SAMPLEGROUP in this case).

FAB Analytic Guidance Document

Stata

Like the other software discussed here, Stata can be used to do means, frequencies, and both linear and logistic regression. The following statements will result in estimated proportions and totals, with standard errors and confidence intervals.

```
svyset [pweight=weight], strata(samplegroup) vce(linearized)
svy: proportion q56 q57 race
svy: tabulate q57 race, count se ci
svy: tabulate q57 q56, count se ci
```

The svyset command specifies the weight, sample design, and variance estimation (vce) method. This code works with Stata v9 through v11.

SPSS Complex Samples

SPSS Complex Samples supports the analysis of stratified or cluster samples for basic frequencies, general linear model, logistic regression, and other statistical methods. SPSS can be run using syntax but also includes a graphical user interface (GUI) that allows analysis to be menu driven (i.e., point-and-click).

In order to run frequencies on age, race, and sex, the SPSS user must first create an analysis plan file (CPSLAN). In this step, the SAMPLEGROUP variable will be designated as the stratum variable and WEIGHT will be designated as the weight variable. No further specifications in CSPLAN are necessary. The frequencies are then generated using the Complex Samples Frequencies procedure (CSTABULATE). The syntax below shows how to use CSPLAN and CSTABULATE:

```
* Analysis Preparation.
CSPLAN ANALYSIS
 /PLAN FILE='Directory and Name of the csplan file'
 /PLANVARS ANALYSISWEIGHT=weight
 /PRINT PLAN
 /DESIGN STRATA= samplegroup
 /ESTIMATOR TYPE=WR.

* Complex Samples Frequencies.
CSTABULATE
 /PLAN FILE = 'Directory and Name of the csplan file'
 /TABLES VARIABLES = q56 q57 race
 /CELLS POPSIZE TABLEPCT
 /STATISTICS SE CIN(95) COUNT
 /MISSING SCOPE = TABLE CLASSMISSING = EXCLUDE.
```

This code works with SPSS 12.0 and later versions.

5.0 Benefits and Limitations in Using and Interpreting the FAB Weights

5.1 The Uses of Sampling Weights

In a probability-based sample, sampling weights serve several purposes. First, they provide a method for adjusting for unequal selection probabilities so that the combined sample (or a subgroup of the sample) accurately represents the target population. Thus, if Hispanics are "oversampled" – meaning that the number of Hispanics in the sample is large relative to the proportion of Hispanics in the general population – the sampling weights will adjust the data so that Hispanics do not represent more than their share of the population. Second, sampling weights are needed to make unbiased estimates of population totals. They weight each sampled unit so that it contributes to a total in proportion to the number of units it represents in the sample (e.g. if a unit was sampled at a rate of 1 in 100, it receives a weight of 100 so that a weighted count adds up to the population). Third, sample weights form a basis for adjusting for survey nonresponse. For example, if younger persons are less likely to participate in the survey, then the sampling weights for this group can be adjusted upwards to offset this loss in response.

However, the sample drawn for the FAB study was not a probability-based sample. It was selected from a panel of enrolled volunteers. In a probability-based sample, participants are selected at random and the probability with which they are selected is known. The weights calculated for FAB are based on *ad hoc* adjustment factors, not selection probabilities, so that weighted estimates from FAB represent a "balanced" sample from the Synovate panel, not the US household population. This complicates the interpretation of the weights, as well as any weighted data analysis, including standard errors based on the weights. See below for further discussion.

5.2 Benefits of Using Weights with FAB Survey Data

While the sample for the FAB study was not selected using probability sampling, using the weights provided here would serve the same purposes as with probability-based samples: individuals and subgroups would be given their appropriate population representation.

Furthermore, the use of the weights in variance calculations would probably result in more realistic standard errors, confidence intervals, and significance tests. Though drawn from a panel, the FAB sample is stratified by race, age, and other characteristics, with a disproportionate sample allocation designed to provide adequate sample sizes for important analytical subgroups. If variance calculations are based on the assumption of "simple random sampling" (SRS) – i.e., without using weights as described above in Section 4.0 – then estimates of standard errors will not reflect the sample design and will probably be somewhat smaller than if they reflected the effects of stratification and differential nonresponse. In other words, the standard errors based on SRS will be smaller and less accurate in FAB than those based on stratified sampling, since the former would not reflect the effects of disproportionate stratified sampling and adjustments for nonresponse, both of which tend to inflate standard errors.

5.3 Disadvantages of Using Weights with FAB Survey Data

The primary disadvantage of weights in this situation is that many statisticians and other researchers would question the validity of using them for a non-probability sample. Some researchers might reject the advantages outlined in the previous section because of the lack of scientific or mathematical justification. From this point of view, the weights do not necessarily provide unbiased estimates because the FAB sample was not selected using probability-based methods from the target population. The FAB

sample can only represent the individuals in the sample, not the larger population from which it was drawn. Thus, researchers should note this limitation when interpreting results and be cognizant to include this caveat when generalizing the results using FAB data to the US population.

Table 1: Stratification and sample sizes for main FAB survey

African American			Target completes	Expected Resp. Rate	Adjustment for postcard	Initial sample	Proportion of US population
Age 18-34	Less than HS education		63	0.17	0.18	360	0.8%
	At least HS education		250	0.56	0.58	435	3.6%
Age 35 or older	Less than HS education		117	0.58	0.60	200	1.6%
	At least HS education	Male	205	0.75	0.77	265	2.6%
	At least HS education	Female	265	0.67	0.69	385	3.4%
		Subtotal	900			1,645	
White and other							
Age 18-34	Male		408	0.52	0.54	760	13.4%
	Female		393	0.64	0.66	600	12.9%
Age 35 or older	Less than HS education		285	0.58	0.60	480	8.8%
	At least HS education	Male	772	0.75	0.77	1,000	25.3%
	At least HS education	Female	842	0.67	0.69	1,220	27.6%
		Subtotal	2,700			4,060	100.0%
		Overall total	3,600			5,705	

Table 2: Weights for analyzing FAB data

Stratum	Race	Age	Education	Sex	Income	Population count	Sample count	Weight
1	African American	18-34	less than HS			1,633,460	64	25,522.8
2a			at least HS		< $40,000	5,019,194	113	44,417.6
2b					> $40,000	2,979,944	92	32,390.7
3		35+	less than HS			3,317,244	71	46,721.7
4a			at least HS	Male	< $40,000	3,372,755	103	32,745.2
4b					> $40,000	2,547,833	88	28,952.6
5a			at least HS	Female	< $35,000	4,519,375	177	25,533.2
5b					> $35,000	3,202,463	172	18,619.0
6a	White and other	18-34		Male	< $60,000	19,267,771	154	125,115.4
6b					> $60,000	10,425,665	135	77,227.1
7a				Female	< $60,000	18,728,268	230	81,427.3
7b					> $60,000	9,779,875	175	55,885.0
8a		35+	less than HS ed.		< $30,000	12,416,613	131	94,783.3
8b					> $30,000	6,778,678	83	81,670.8
9a			at least HS ed.	Male	< $35,000	19,980,947	148	135,006.4
9b					$35k to $60k	11,466,520	141	81,322.8
9c					$60k to $100k	12,702,896	173	73,427.1
9d					> $100,000	11,821,225	154	76,761.2
10a			at least HS ed.	Female	< $40,000	28,288,759	341	82,958.2
10b					$40k to $75k	14,989,842	285	52,595.9
10c					> $75,000	18,048,634	367	49,178.8
Total						221,287,962	3397	

Mean of weights	65,142
Standard deviation of weights	29,578
CV of weights	0.45
Design effect	1.21

Appendix A: Creation of the FAB Derived Variables

Three types of variables where derived from questionnaire responses:

1) Cup Equivalents for fruits and vegetable consumption (Q37-Q41)

2) Hours of physical activities in a typical week in the past month (Q33)

3) Race/ethnicity combination (Q58 and Q59)

Cup Equivalents

The derivation of Cup Equivalents variables is based on a formula published elsewhere: http://riskfactor.cancer.gov/diet/screeners/fruitveg/scoring/allday.html.

Step 1) A frequency multiplier is derived based on the answer to Q34A through Q41A

If the answer is…	then the multiplier is…
0 or 1	0
2	.067
3	.214
4	.5
5	.786
6	1
7	2
8	3
9	4
10	5
Missing	missing

Step 2) A specific amount multiplier is derived based on the answer to Q34B through Q41B

For Q34B (Juice)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.5
3	1
4	1.625
5	2.5

For Q35B (Fruit) and Q36B (LSALAD)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.25
3	.5
4	1
5	1.5

For Q37B (FRFRY)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.2
3	.5
4	.75
5	1.3

For Q38B (WHPOT) and Q39B (DRBEAN)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.25
3	.75
4	1.2
5	2

For Q40B (OTHVEG)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.25
3	.75
4	1.5
5	2.25

For Q41B (TOMSAUCE)

If the answer is…	then the amount multiplier is…
1 or the frequency multiplier is 0	0
2	.25
3	.5
4	1
5	1.5

Step 3) The Amount multiplier and frequency multiplier are multiplied together to produce the Cup Equivalent derived variable.

Example: Q34A=8, Q34B=3 the multipliers are 3 and 1 for JUICE_CUP=3

Step 4) Combined Cup Equivalent variables are totaled:

FRTVEG_CUP=JUICE_CUP+FRUIT_CUP+LSALAD_CUP+FRFRY_CUP+WHPOT_CUP+DRBEAN_CUP+OTHVEG_CUP+TOMSAUCE_CUP;

FRUIT_ONLY_CUP= JUICE _CUP+ FRUIT _CUP;

VEG_WFRYPTO_CUP=LSALAD_CUP+FRFRY_CUP+WHPOT_CUP+DRBEAN_CUP+OTHVEG_CUP+TOMSAUCE_CUP;

VEG_NFRYPTO_CUP=LSALAD_CUP+WHPOT_CUP+DRBEAN_CUP+OTHVEG_CUP+TOMSAUCE_CUP;

FRTVEG_NOFRYPTO_CUP=JUICE_CUP+FRUIT_CUP+LSALAD_CUP+WHPOT_CUP+DRBEAN_CUP+OTHVEG_CUP+TOMSAUCE_CUP;

Hours of Physical Activities in a Typical Week in the Past Month (PHY_WK_HRS)

For non-missing Q33B_HRS, Q33B_MIN, and Q33A (days):

PHY_WK_HRS is calculated as Q33A-1 * (Q33B_HRS + Q33B_MIN/60) rounded to the hour.

Race/Ethnicity Combination (RACE)

If…	then RACE is…
Q58 is 1	3 (Hispanic)
more than one of Q59A, Q59B, Q59C, Q59D, or Q59E is 1	7 (mix non-Hispanic)
Q59A is 1	1 (non-Hispanic white)
Q59B is 1	2 (non-Hispanic black)
Q59C is 1	4 (non-Hispanic Asian)
Q59D is 1	5 (non-Hispanic A. Indian)
Q59E is 1	6 (non-Hispanic Native Hawaiian)