Methodology for Weighting 2010-2011 Modified-BRFSS Data Collected for the CPPW Communities

Authors

Ismael Flores Cervantes
Jing Kang
Richard Sigman
Klaus Teuter

August 2011

Prepared for:
Centers for Disease Control and Prevention
Atlanta, Georgia

Prepared by:
Westat
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Methodology for Weighting 2010-2011 Modified-BRFSS Data Collected for the CPPW Communities

The state-level Behavioral Risk Surveillance System (BRFSS) conducts telephone interviews to collect health and demographic data from samples of adults in each state and territory. A modified version of the state-level BRFSS data collection methodology was used to collect data in 2010 and 2011 in individual communities participating in the Communities Putting Prevention to Work (CPPW) program. This document describes the methodology used to create the analytical weights for the BRFSS data collected from the CPPW communities in 2010 and 2011. The first section describes the sample designs, and the second section describes the creation of the initial weights and the weighting adjustments used to create the final weight. The third section describes the imputation procedures used to impute the missing values for the variables used in weighting. The fourth section describes the control totals used in raking. The last section describes how to use the developed weighting software.

Sample Designs 1

The modified-BRFSS surveys for the CPPW communities are stratified landline telephone surveys of each community's civilian, non-institutionalized adult population. The design of the CPPW-community BRFSS builds on the design of the state-level BRFSS and consists of a random digit dialing (RDD) landline telephone sample for each community. The CPPW communities are listed in Table 1. The geographic definitions of the communities were specified in terms of counties, cities, ZIP codes, telephone exchanges, or Census tracts.

The selected landline telephone samples used a list-assisted method for sampling telephone numbers in each CPPW community. This single-stage, unclustered sampling method selects a probability sample from all telephone numbers that are in 100 banks containing at least one residential listed telephone number (referred to as *1+ banks*). The sample of telephone numbers for each community was selected using disproportionate stratified random sampling. (See Behavioral Risk Factor Surveillance System Operational and User's Guide at ftp://ftp.cdc.gov/pub/Data/Brfss/userguide.pdf for additional details).

Some of the CPPW communities requested that their telephone samples be stratified by geography or demographic characteristics. For these communities, the sample was selected in two phases. First, geographic or demographic strata were created by classifying each of the community's telephone exchanges to a particular stratum, and then a sample of telephone numbers was drawn from all the 1+ banks in the exchanges assigned to each stratum. The sampled numbers were purged and matched to lists of telephone numbers to determine if they were listed residential telephone numbers. Using this information, three substrata were created. The high density substratum contained all working telephone numbers found to be listed. The medium density substratum contained all working telephone numbers found not to be listed. The third substratum contained all non-working numbers. In the second phase, a subsample is selected from the first and second substrata oversampling the high density stratum relative to the medium density stratum by a factor of 1.5. For communities that did not request sample stratification by geography or demographic characteristics, an unstratified sample of phone numbers was first selected from the 1+ banks of the community's telephone exchanges, the listed-residential-telephone status of each sampled telephone number was determined, and then density-stratum subsampling was performed.

Table 1. CPPW Communities

Community code		Description	Type of geography
1	AL073	Jefferson County	County
2	AL097	Mobile County	County
3	AR063	Independence County	County
4	AR119	North Little Rock-Pulaski County	Zip Codes
5	AZ019	Part of Pima County	Census tracts
6	CA037	Los Angeles County	County
7	CA073	San Diego County	County
8	CA085	Santa Clara County	County
9	CO999	Adams, Arapahoe and Douglas Counties	County
10	DC000	District of Columbia	District
11	FL086	Miami-Dade County	County
12	FL095	Orange County	County
13	FL103	Pinellas County	County
14	GA089	Dekalb County	County
15	HI007	Kauai County	County
16	HI009	Maui County	County
17	IA113	Linn County	County
18	IA159	Ringgold County	County
19	IL031	Cook County	County
20	IL1600	Chicago	City
21	IN003	Bartholomew County	County
22	IN082	Vanderburgh County	County
23	KY111	Lousville	County
24	MA025	Boston	Census tracts
25	ME998	Healthy city of Portland	ZIP codes
26	ME999	Healthy Lakes	ZIP codes
27	MN053	Minneapolis	Census tracts
28	MN109	Rochester	County
29	MO999	St. Louis County	Census tracts
30	NC147	Pitt County Health District in Pitt County	County
31	NC999	Appalachian Health in Alleghany County	Counties
32	NE999	Douglas County	County
33	NV003	Clark County	County
34	NY999	New York City	Counties
35	OH061	Hamilton County	County
36	OK999	Cherokee Nation	Telephone Exchanges
37	OR051	Multnomah County	County
38	PA101	Philadelphia County	County
39	RI999	City of Providence	Census tracts
40	SC041	Florence County	County
41	SC051	Horry County	County
42	TN037	Davison County	County
43	TX453	Austin inTravis County	County
44	TX999	San Antonio in Bexar County	County
45	WA033	King County	County
46	WI063	Lacross County	County
47	WI141	Wood County	County
48	WI999	Great Lakes Inter-Tribal Council in Menominee, Barron, Bayfield, Burnett, Sawyer, Shawano, Polk, Oconto, Langlade, Washburn Counties	Counties
49	WV999	Mid-Ohio Valley	Counties

Weighting Procedures 2

2.1 Weighting Approach

We developed a set of weights—consisting of a base weight, a person weight, a raked weight, and a trimmed weight—for each adult who completed an extended interview. We used the same weighting procedures across the different CPPW communities, taking into account each community's sample design. To the extent possible, the weighting procedure accomplished the following objectives:

- Compensated for differential probabilities of selection;
- Reduced biases due to nonresponse;
- Adjusted for undercoverage due to households without landline telephones; and
- Made the estimates consistent with population totals from other sources while simultaneously reducing the variance of the estimates.

2.2 Recoded Variables

Recoded variables were created from the collected survey data or from information associated with sample selection. Only the recoded variables were used in the weighting calculations, and if the recoded variables contained missing data they were imputed. Table 2 lists the names of the recoded variables and their associated imputation-flag variables.

Table 2. Recoded variables

Recoded variable	Source variable	Imputation-flag variable	Description
RNUMPHON	NUMPHON2 and NUMHHOL2	IMP_RNUMPHON	Recoded number of telephone numbers
RNUMADULT	NUMADULT	IMP_RNUMADULT	Recoded number of adults
RRACE	MRACE	IMP_RRACE	Recode of respondent's race. The variable MRACE includes all races that apply. The variable RRACE includes only the following levels: 1 = White alone 2 = Black or African American alone 3 = Asian alone 4 = Native Hawaiian or Other Pacific Islander alone 5 = American Indian, Alaska Native alone 6 = Other alone 7 = Two or more races
RSEX	SEX	IMP_SEX	Recoded respondent's sex 1 = Male 2 = Female
RHISP	HISPANC2	IMP_RHISP	Recoded respondent's ethnicity 1 = Hispanic 2 = Non-Hispanic
REDU	EDUCA	IMP_REDU	Respondent's education level 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
RMAR	MARITAL	IMP_RMAR	Respondent's marital education 1 = Married 2 = Divorced 3 = Widowed 4 = Separated 5 = Never married 6 = A member of an unmarried couple 9 = Refused
RSTR	_GEOSTR		Community sampling strata
RAGE	AGE	IMP_RAGE	Respondent's age

2.3 Base Weights

The first step of weighting was to compute the household base weight, defined as the inverse of the probability of selection. The base weight depends on how the sample was selected. As described above, the samples were selected using disproportionate stratified random sampling. In addition to

the way the sample was selected, the value of the base weight reflects whether additional samples were selected at subsequent times during the data collection period. Samples selected at later times may have been drawn from an updated frame different from the one used in the first selection. Since duplicate sampled telephone numbers were removed from the subsequent samples, the base weight was created as if the samples were drawn at the same time.

The household base weight $BSWGT$ was computed as

$$BSWGT = \frac{\overline{N}_h}{\sum_t n_{th}},$$

where $\overline{N}_h$ is the average frame size in all selections t and n_{th} is the number of telephone numbers sampled in selection t in stratum h.

2.4 Multiple Telephone Adjustment

During the telephone interviews, information about the existence of additional telephone numbers and their use in the household was collected. If the additional telephone number was used for residential voice communications (not solely for business, fax or computer use, etc.), the household had a greater probability of selection because it could have been selected through any of the additional telephone numbers in the household. In this case, the household weight was adjusted to reflect the increased probability of selection. The multiple telephone adjusted household weight, $HHAWGT$, is computed as

$$HHAWGT = \frac{BSWGT}{RNUMPHON},$$

where *RNUMPHON* is the variable for the number of residential telephone numbers in the household. When *RNUMPHON* was missing, it was assumed that there was only one telephone number in the household. In other words, the variable *RNUMPHON* was imputed with a value of 1.

2.5 Person Weights

The initial person weight was computed using the adjusted household weight and the inverse of the probability of selecting the sampled person within household. The initial person weight, $PWGT$, was computed as

$$PWGT = HHAWGT * RNUMADULT ,$$

where *RNUMADULT* was the number of eligible adults in the household. When this variable was missing, the modal value within the sampling stratum was used.

2.6 Raked weights

The last step in weighting was to rake the person weights to population control totals. Raking is a commonly used estimation procedure in which estimates are controlled to known marginal population totals. It can be thought of as a multidimensional poststratification procedure because the weights are poststratified to one set (a dimension) of control totals, and then these adjusted weights are poststratified to another dimension. The procedure continues until all dimensions are adjusted. The process is then iterated until the control totals for all dimensions are simultaneously satisfied (at least within a specified tolerance).

An important advantage of raking over other simpler adjustment methods such as poststratification is that it permits the use of information with multiple characteristics (e.g., race, ethnicity, sex, geographic area). Raking also allows us to use information at different levels of geography, so that adjustments to population totals at the community level and also at smaller areas can be made simultaneously.

The goal of raking is to mitigate sources of survey error, such as under-coverage and nonresponse. Nonresponse biases arise in survey estimates of means and proportions when the characteristics of respondents differ from those of nonrespondents. Under-coverage also biases survey estimates when the characteristics of individuals in households that do not have a chance to be selected differ from those in households that do have a chance to be selected.

The raked weight, $RAKEDW_i$, for person i can be expressed as

$$RAKEDW_i = PWGT_i \cdot \prod_{k=1}^{K} RAKEDF_{k_l} ,$$

where $RAKEDF_{k_l}$ is the raking factor for dimension *k* and level *l* (which contains person *i*). For example, if the 4th dimension ($k=4$) is sex with two levels ($l=1$ for male and $l=2$ for female), then the raking factor for this dimension is $RAKEDF_{4_1}$ for the males. The raking factors are derived so that the following relationship holds for each raking dimension *k* and level *l*:

$$CNT_{k_l} = \sum_j \delta(k_l)_j \cdot RAKEDW_j ,$$

where CNT_{k_l} is the control total, and $\delta(k_l)_j = 1$ if the person is in level *l* of dimension *k* and equals zero, otherwise.

2.7 Raking Dimensions

Raking has many potential benefits, but obtaining these full benefits depends on the choice of the dimensions and their levels. The raking dimensions that we used were based on those used in the state-level BRFSS weighting. For communities defined by counties or Census tracts, we used the raking dimensions described in Table 3.

Table 3. Raking dimensions

Dimension	Description	Levels	Description
1	Age group by gender Variables RAGE and RSEX	11 12 21 22 31 32 41 42 51 52 61 62 71 72	18-24 years old, male 18-24 years old, female 35-34 years old, male 35-34 years old, female 35-44 years old, male 35-44 years old, female 45-54 years old, male 45-54 years old, female 55-64 years old, male 55-64 years old, female 65-74 years old, male 65-74 years old, female 75 years old or older, male 75 years old or older, female
2	Race/ethnicity Variables RHISP and RRACE	1 2 3 4	White non-Hispanic Black non-Hispanic Hispanic Other
3	Education Variable REDU	1 2 3 4	Less high school High school graduate Some college College graduate
4	Marital Status Variable RMAR	1 2 3	Married Never married or part of an unmarried couple Divorced, widowed, or separated
5	Sex by race/ethnicity Variables RSEX, RHISP and RRACE	11 12 13 14 21 22 23 24	Male, White non-Hispanic Male, Black non-Hispanic Male, Hispanic Male, Other Female, White non-Hispanic Female, Black non-Hispanic Female, Hispanic Female, Other
6	Age by race/ethnicity Variables RAGE, RHISP and RRACE	11 12 13 14 21 22 23 24 31 32 33 34	18-34 years old, White non-Hispanic 18-34 years old, Black non-Hispanic 18-34 years old, Hispanic 18-34 years old, Other non-Hispanic 35-54 years old, White non-Hispanic 35-54 years old, Black non-Hispanic 35-54 years old, Hispanic 35-54 years old, Other non-Hispanic 55 years old or older, White non-Hispanic 55 years old or older, Black non-Hispanic 55 years old or older, Hispanic 55 years old or older, Other non-Hispanic
7	Sampling strata Variable RSTR		Sampling strata when defined as counties or a set of Census tracts

The last dimension, Dimension 7, was only used when a stratum was defined as one or more whole counties or a set of census tracts. This dimension was not used in communities where strata were defined by ZIP codes or by demographic characteristics of the community's telephone exchanges.

For communities defined in terms of ZIP codes or telephone exchanges, there was no detailed information to create the same raking dimensions. In these cases, the communities were raked using one dimension defined by age group and sampling strata. More details related to the control totals are provided in Section 4.

Raking with so many dimensions can produce aberrant results if care is not taken during the process. Small cell sizes cause problems in the convergence of the estimates to the control totals. The minimum number of respondents in a raking cell was 50. If small sample sizes were found, the cells in the dimensions were combined or collapsed according to a set of rules. The collapsing rules were similar to those used in the state-level BRFSS procedure.

2.8 Collapsing Rules for Raking Cells

Cells that caused a failure of convergence in raking, had a large adjustment factor, or contained less than 50 respondents were collapsed with one or more similar or adjacent cells. Only cells that needed to be collapsed were collapsed. The collapsing rules for each dimension are described below:

- **Dimension 1: Age and Sex**. Sex was a hard boundary, and it was never collapsed. Adjacent age groups were collapsed, but no collapsed cell was created that crossed the ages ranges 18 to 44 and 45 or older. For example, if any of the cells 18 to 24 years old, 25 to 34 years old, 35 to 44 years old were deficient, they were collapsed to an adjacent cell. If any of the cells 45 to 54 years old, 55 to 64 years old, 65 to 74 years old, 75 and up were deficient, they were collapsed to an adjacent cell.

- **Dimension 2: Race/Ethnicity.** Minority groups were kept as separate as possible. Preferred collapsed cells included combining Black non-Hispanic with Other or combining Black non-Hispanic with Hispanic and Other if these collapsed cells yielded the minimum number of respondents in the cell. In a few communities, all minorities were grouped into a single cell.

- **Dimension 3: Education.** When collapsing was needed, we created the two collapsed cells: (1) high school or less and (2) more than high school.

- **Dimension 4: Marital status.** This dimension was rarely collapsed. In few cases, never married or part of an unmarried couple was collapsed with divorced, widowed, or separated.

- **Dimension 5: Sex by Race/Ethnicity.** Sex was a hard boundary, and it was never collapsed. Race/Ethnicity was collapsed following the rule for Dimension 2.
- **Dimension 6: Age (3 levels) by Race/Ethnicity.** Age groups 18 to 34 years old and 55 years old or older were never collapsed. Within these groups, race/ethnicity was collapsed following the rule for Dimension 2. In the younger groups with very small samples, all race/ethnicity groups were collapsed within age the 18 to 34 group.
- **Dimension 7. Sampling stratum:** This dimension was never collapsed.

2.9 Trimming

Raking to multiple dimensions can sometimes yield very large weights that have a large impact on estimated totals or their variances. After raking the person weights to the known control totals, the distribution of the weights were examined to determine the presence of very large weights. If observations with large weights were found, the weights for these cases were reduced in a process called trimming.

We examined the distribution of the 15 largest weights to identify weights that were candidates for trimming. A cut-off weight was determined that was the lower bound of a large gap in the distribution of the 15 largest weights. The weights greater than the cut-off weight were trimmed. The trimmed weight, $TRMW_i$, was computed as

$$TRMW_i = TFACT_i * PWT_i,$$

where $TFACT_i$ is the trimming factor for the sampled adult *i* given by

$$TFACT_i = \begin{cases} 1 & \text{if the weight } i \text{ is not trimmed} \\ \dfrac{CUT\_OFF\_WGT}{RAKEDW_i} & \text{otherwise} \end{cases}.$$

where CUT_OFF_WGT is the cut-off weight for sampled adult *i*.

The trimming process consisted of several steps. First, the person weight was raked to produce the raked adjusted weight. Using this weight, the trimming factor was computed and applied to the person weight before raking. The trimmed person weight was then raked again to produce the raked weight. In this way, the new raked weight incorporated the trimming factor. The new raked weights were examined again to identify extreme weights. If this was the case, the process was repeated,

applying the new trimming factor to the person weight. This process was repeated until there are no extreme weights left in the file or no further reductions in large weights was possible,

Imputation 3

As in most surveys, the responses to some data items were not obtained for all interviews. The items that were needed for raking but were missing were imputed. We used a procedure similar to the one used in the state-level BRFSS. The imputation of variables needed for raking was sequential, and imputed values were used to create imputation cells for later imputations. We used three imputation procedures: modal, mean and hot deck imputation.

3.1 Modal Imputations

In the first part of the imputation process, modal values were imputed for race (RRACE), ethnicity (RHISP), and sex (RSEX) within cells based on which of these three variables were missing. See Table 4, which indicates the sequence of the modal impuations. The values with the highest frequency (i.e., modal value) within the cell were used to impute missing values. For example, in Cell 1, all respondents that have missing values of race (RRACE) and ethnicity (RHISP) were imputed with the most common values of RRACE and RHISP within sampling strata (RSTR). In Cell 3, respondents with missing value of RSEX were imputed with the modal value of RSEX in the cells created by the cross tabulation of RSTR, RRACE, and RHISP that matched the respondent's RSTR, RRACE, and RHISP. In Cell 4, there were no missing values of race or ethnicity because if these variables had contained missing values they would have been imputed in the previous processing for Cells 1, 2, or 3.

Table 4. Sequence of modal imputation for RSEX, RHISP, and RRACE

Cell	Cell condition	Procedure
1	RRACE = missing, RHISP = missing	Modal value of RRACE and RHISP by RSTR (geography)
2	RRACE = missing, RHISP ≠ missing	Modal value of RRACE by RSTR *RHISP
3	RRACE ≠ missing, RHISP = missing	Modal value of RHISP by RSTR *RRACE
4	RSEX = missing	Modal value of RSEX by RSTR *RRACE *RHISP

3.2 Mean Imputation

In the second part of the imputation process, the mean age rounded to the nearest whole year computed from respondent data in cells defined by sex, race, and ethnicity was used to impute missing values of age. In this case, there were no missing values for the variables used to define the cells because they were imputed in the modal-imputation step.

3.3 Hot Deck Imputations

Hot deck imputation was used to impute education (REDU) and marital status (RMAI). In this procedure, the response from a unit that answered the question was imputed to the missing case. The case that was imputed was called the recipient, and the case that was used to complete the missing value was called the donor. The donor was randomly selected among all respondents within imputation cells created by cross tabulating the variables for geography, sex, race- ethnicity, and age group. Donors were used only once. In cases where there were insufficient donors in the imputation cell, the cells were collapsed until there a sufficient number of donors.

4 Control Totals

The American Community Survey (ACS) was the main source for the control totals for those communities that were defined in terms of counties or Census tracts. These control totals were derived from the 2005-2009 ACS summary file (see http://www.census.gov/acs/www/). The summary file contains tables with totals of population for the following groups:

- **Tables B01001A-G**: Race (White alone, Black or African American alone, American Indian and Alaska, Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone, some other race alone, two or more races) by age group and sex.
- **Table B01001H**: White alone non-Hispanic by age group and sex.
- **Table B01001I**: Hispanic by age group and sex.
- **Table B03002**: Hispanic by race.
- **Table B12001**: Sex by marital status for the population 15 years and over.
- **Table B15001**: Sex by age by educational attainment for the population 18 years and over.

Although most of the control totals used in raking were available from the ACS Summary file, some control totals were estimated. For example, the information about marital status was available for the population 15 years old or older but the eligible population was 18 years old or older. The totals by age group and gender for groups such as Black alone non-Hispanic and other race non-Hispanic were also estimated.

As part of the development of the control totals, a single file containing detailed totals was created for the combination of the variables used in raking. This file was created in such a way so that if it was summarized for any of the raking dimensions the control totals from the ACS summary table could be reproduced. Deriving the control totals from a single file of detailed totals ensured that the control totals were consistent. The file of detailed totals was created using raking to the totals obtained from the ACS summary file.

For communities that were defined in terms of ZIP codes or telephone exchanges, we used the information provided by the Marketing Systems Group (MSG), the sampling vendor in charge of

selecting and processing the telephone samples. MSG maintains demographic information for telephone exchanges, which is derived from annual demographic estimates produced by Claritas for Census geographies. The MSG-provided demographic information was limited to totals by age groups. Consequently, communities defined in terms of ZIP codes and telephone exchanges were raked using only one dimension defined by age group and sampling strata. Table 5 lists the communities and the source of the control totals.

Table 5. Source of control totals

Community code		Type of geography	Source of control totals
1	AL073	County	ACS
2	AL097	County	ACS
3	AR063	ZIP codes	ACS
4	AR119	Census tracts	Claritas
5	AZ019	Census tracts	ACS
6	CA037	County	ACS
7	CA073	County	ACS
8	CA085	County	ACS
9	CO999	County	ACS
10	DC000	Disctrict	ACS
11	FL086	County	ACS
12	FL095	County	ACS
13	FL103	County	ACS
14	GA089	County	ACS
15	HI007	County	ACS
16	HI009	County	ACS
17	IA113	County	ACS
18	IA159	County	ACS
19	IL031	Census tracts	ACS
20	IL1600	Census tracts	ACS
21	IN003	County	ACS
22	IN082	County	ACS
23	KY111	County	ACS
24	MA025	Census tracts	ACS
25	ME998	ZIP codes	Claritas
26	ME999	ZIP codes	Claritas
27	MN053	Census tracts	ACS
28	MN109	County	ACS
29	MO999	Census tracts	ACS
30	NC147	County	ACS
31	NC999	County	ACS
32	NE999	County	ACS
33	NV003	County	ACS
34	NY999	Counties	ACS
35	OH061	County	ACS
36	OK999	Telephone Exchanges	Claritas
37	OR051	County	ACS
38	PA101	County	ACS
39	RI999	Census tracts	ACS
40	SC041	County	ACS
41	SC051	County	ACS
42	TN037	County	ACS
43	TX453	County	ACS
44	TX999	County	ACS
45	WA033	County	ACS
46	WI063	County	ACS
47	WI141	County	ACS
48	WI999	Counties	ACS
49	WV999	Counties	ACS