# UnderStandingAmericaStudy

## WEIGHTING PROCEDURE, JUNE 2022 – FEBRUARY 2023

USC Dornsife Center for Economic and Social Research

# Contents

## INTRODUCTION

This document provides details of the weighting procedures and benchmark distributions used to create final sample weights for data sets collected by the Center for Economic and Social Research's Understanding America Study internet panel.[1] The weighting procedure described in this document was used starting June 2022 until February 2023.

## 1. SAMPLING

In this section, we provide a summary of UAS's sampling procedures as background for our weighting protocol. For a full description of the UAS sampling and recruitment procedures, please check the UAS website at uasdata.usc.edu.

The UAS is a nationally representative panel of U.S. households recruited through Address Based Sampling (ABS). Eligible individuals are all adults in the contacted household aged 18 and older.

Sampling in the UAS is carried out in batches. There are currently 25 batches, targeting either the U.S. population at large, or specific subsets of it, such as the population of Native Americans, California residents, and Los Angeles County residents. Table 1 below lists all the UAS recruitment batches as of June 2022 and their corresponding reference populations.

Batch 1 is a simple random sample (SRS) from the list of all individuals in the ASDE Survey Sampler database. Batches 22-23 are simple random samples of all U.S. addresses provided by the vendor Marketing Systems Group (MSG). Batch 4 is a simple random sample of all addresses listed on birth certificates issued in Los Angeles County in the years 2009-2012 in a limited set of zip codes.

Batches 2-21 are based on a two-stage sample design, in which zip codes are drawn first, and then households are randomly drawn from the sampled zip codes. More precisely, these batches are selected using an adaptive sampling algorithm, which we refer to as Adaptive1. This algorithm allows to refresh the panel in such a way that its demographic composition moves closer to the population composition. Specifically, before sampling an additional batch, the algorithm computes the unweighted distributions of specific demographic characteristics (e.g., sex, age, marital status

[1] Mick Couper and Jon Krosnick have provided insightful and valuable comments throughout the development of the UAS weighting procedure.

and education) in the UAS at that point in time. It then assigns to each zip code a non-zero probability of being drawn, which is an increasing function of the degree of "desirability" of the zip code. The degree of desirability is a measure of how much, given its population characteristics, a zip code is expected to move the current distributions of demographics in the UAS towards those of the U.S. population. For example, if at a particular point in time the UAS panel underrepresents females with high school degree, zip codes with a relatively high proportion of females with high school degree receive a higher probability of being sampled. This sampling algorithm is implemented iteratively. That is, after selecting a zip code, the distributions of demographics in the UAS are updated according to the expected contribution of this zip code towards the panel's representativeness, updated measures of desirability are computed and new sampling probabilities for all other zip codes are defined. Such procedure provides a list of zip codes to be sampled. The implementation of this algorithm implies that the marginal probability of drawing each zip code depends on the composition of the UAS panel at a particular point in time, but also on the unknown response probabilities of selected households in that zip code. Hence, the marginal probability of drawing each zip code is not known ex ante and cannot be used to construct design weights. The weighting procedure corrects for the unequal sampling probabilities generated by this adaptive sampling algorithm.

Table 1: UAS Recruitment Batches

| Batch | Reference Population | PSU | Method |
|---|---|---|---|
| 1 | U.S. | Address | SRS |
| 2-3 | Native American | Zip Code | Adaptive1 |
| 4 | Los Angeles County (birth certificate list sample) | Address | SRS |
| 5-12 | U.S. | Zip Code | Adaptive1 |
| 13-14 | Los Angeles County | Zip Code | Adaptive1 |
| 15-16 | California | Zip Code | Adaptive1 |
| 17 | U.S. | Zip Code | Adaptive1 |
| 18-19 | Los Angeles County | Zip Code | Adaptive1 |
| 20-21 | U.S. | Zip Code | Adaptive1 |
| 22-23 | U.S. | Address | SRS |
| 24-25 | U.S. | Address | Adaptive2 |

Batches 24-25 are random samples of all U.S. addresses provided by MSG, with selection probabilities generated by a second adaptive sampling algorithm, which we refer to as Adaptive2. This second sampling approach involves the following two steps. First, we combine auxiliary information attached to the post office delivery sequence files provided by MSG and information about population characteristics at the Census tract level to compute the probability of belonging to a certain population stratum for each potential invitees. Second, we assign sample inclusion probabilities so that, in expectation, the final sample of selected invitees has a desired distribution of demographic characteristics. Such desired distribution is defined in order to increase the representation of under-represented segments of the population in the UAS (e.g., individuals with high school or less) and/or to obtain sufficient sample sizes to detect meaningful differences between groups (e.g., differences in socio-economic status and health outcomes between racial/ethnic groups). The inverse of the inclusion probabilities generated by this second adaptive sampling algorithm represent the individual-level base weights that account for differential selection probabilities across sampled individuals.

### 1.1. Respondents with a weight of zero

Recruitment batches 2 and 3 target the population of Native Americans. Even though non-Native Americans contacted within these two batches were not eligible to become panel members, some were accidentally invited to join the UAS. Because we are unable to attach a probability to this happening, these panel members receive a weight of zero.

As mentioned above, recruitment batch 4 was a simple random sample of addresses listed on birth certificates issued in Los Angeles County in the years 2009-2012 in a limited set of zip codes. Because of the highly specific nature of this subsample, we do not provide weights for members recruited within this batch and assign to all of them a weight of zero.

Thus, **we provide weights for respondents in all batches listed in Table 1, except for non-Native American respondents in batches 2 and 3 and all respondents in batch 4**.

## 2. WEIGHTING

In the UAS, sample weights are survey-specific. They are provided with each UAS survey and, unless otherwise indicated, are meant to make each survey data set representative of the U.S. population with respect to a pre-defined set of socio-demographic variables. Sample weights are constructed in two steps. In a first step, a *base weight* is created to account for unequal probabilities of sampling UAS members generated by the adaptive sampling algorithms described above. In a second step, *final post-stratification weights* are generated to correct for differential non-response rates and to bring the final survey sample in line with the reference population as far as the distribution of key variables of interest is concerned.

## 2.1. Step 1: Base Weights

*Base weights for members sampled via Adaptive1*

When computing base weights for batches using Adaptive1, the unit of analysis is a zip code. We estimate a logit model for the probability that a zip code is sampled as a function of its characteristics, namely Census region, urbanicity, population size, as well as sex, race, age, marital status and education composition. Estimation is carried out on an American Community Survey (ACS) file that contains 5-year average characteristics at the zip code level.[2] The outcome of this logit model is an estimate of the marginal probability of a zip code being sampled, which, given the implementation of the adaptive sampling algorithm Adaptive1 described above, is not known *ex ante*.

We indicate by $\pi_k$ the logit estimated probability of sampling zip code $k$. The probability of sampling household $h$ after drawing zip code $k$ is the ratio of the number of households sampled divided by the number of households in the zip code. We indicate this by $\pi_{h|k}$. Hence, the marginal probability that household $h$ from zip code $k$ is sampled is $\pi_{hk} = \pi_{h|k} \times \pi_k$.

The base weight is a zip code level weight defined as:

$$w_{hk}^{base} = \Lambda \times \frac{1}{\pi_{hk}}$$

where the constant $\Lambda$ is chosen such that the sum of the base weights is equal to the number of sampled households. A comprehensive discussion of how base weights for UAS members sampled via Adaptive 1 are computed is provided in Angrisani et al. (2020), available here.

[2] Strictly speaking, all files from the U.S. Census Bureau use "zip code tabulation area" (zcta), which is based on, but not identical to, USPS's definition of zip codes. We ignore the distinction between the two.

UAS members sampled via Adaptive 1 are assigned a base weight, computed as described above, depending on the zip code where they reside at the time of recruitment.

*Base weights for members sampled via Adaptive2*

The adaptive sampling algorithm Adpative2 assigns specific inclusion probabilities to all invitees depending on their demographic characteristics and on the desired distribution of demographic characteristics in the entire UAS that should be achieved with each refresher batch. The inverse of these inclusion probabilities constitute the individual-level base weights that account for differential selection probabilities across UAS members sampled via Adaptive2.

## 2.2. Step 2: Poststratification Weights

The execution of the sampling process for a survey is typically less than perfect. Even if the sample of panel members invited to take a survey is representative of the population along a series of dimensions, the sample of actual respondents may exhibit discrepancies because of differences in response rates across groups and/or other issues related to the fielding time and content of the survey. A second layer of weighting is therefore needed to align the final survey sample to the reference population as far as the distribution of key variables is concerned.

In this second step, we perform **raking weighting** (also known as iterative marginal weighting), starting from the base weights, $w_{hk}^{base}$, described in the previous section. With this, we assign poststratification weights to survey respondents such that the weighted distributions of specific socio-demographic variables in the survey sample match their population counterparts (benchmark or target distributions).

The benchmark distributions against which UAS surveys are weighted are derived from the Basic Monthly Current Population Survey (CPS).[3] We use the 6 most recent available monthly CPS at the time a UAS survey is completed. This ensures a minimum gap between the period of survey completion and the period benchmark distributions refer to.

[3] We rely on IPUMS-CPS, University of Minnesota, https://ipums.org/.

Unless otherwise required by the aims of the survey and specified in the sample selection process, the reference population for UAS surveys is the U.S. population of adults, age 18 or older, excluding institutionalized individuals and military personnel.

## 2.3. Categorization and Imputation of Variables

For post-stratification weighting purposes, we use demographic information taken from the most recent *My Household* survey, which is answered by all active UAS members every quarter. All socio-demographic variables in the *My Household* survey are categorical, but some, such as age, education, and income, take values in a relatively large set. We recode all socio-demographic variables considered for poststratification into new categorical variables with no more than 5 categories. The aim of limiting the number of categories is to prevent these variables from forming strata containing a very small fraction of the sample (less than 5%), which may cause sample weights to exhibit considerable variability.

The list of all recoded categorical variables considered for poststratification is reported in Table 2.

Before implementing the poststratification weighting procedure, we employ the following imputation scheme to replace missing values of recoded socio-demographic variables.

- Gender is obtained from administrative records.
- When age is missing, the age range available in the My Household survey is used to impute age categories. If the age range is also missing, age categories are imputed using gender-specific sample mode.
- Once age categories have been imputed (if missing), the variable with the fewest missing values is the first one to be imputed by means of a regression featuring gender and the age categories as regressors. This newly imputed variable is then added to the set of regressors to impute the variable with the second smallest number of missing values. The procedure continues in this fashion until the variable with the most missing values is imputed using information on all other available socio-demographic variables.

Table 2: List of Recoded Categorical Variables for Poststratification

| Recoded Variable | Categories |
|---|---|
| *Gender* | 1. Male; 2. Female |
| *Age* | 1. 18-39; 2. 40-49; 3. 50-59; 4. 60+ |
| *Born in the US* | 0. No; 1. Yes |
| *US citizen* | 0. No; 1. Yes |
| *Education* | 1. High School or Less; 2. Some College; 3. Bachelor or More |
| *Native American* | 0. No; 1. Yes |
| *Race-ethnicity* | 1. White; 2. Black; 3. Others; 4. Hispanic; 5. Native American |
| *Census region (augmented)** | 1. Northeast; 2. Midwest; 3. South; 4. West, excl CA; 5. CA, excl LAC; 6. LAC |
| *Marital status* | 1. Married; 2. Separated/Divorced/Widowed; 3. Never Married |
| *Work status* | 1. Working; 2. Unemployed; 3. Retired; 4. On leave, Disabled, Other |
| *Household composition* | 1. 1 Member; 2. 2 Members; 3. 3 or 4 Members; 4. 5 or More Members |
| *Household income* | 1. <$30,000; 2. $30,000-$59,999; 3. $60,000-$99,999; 4. $100,000+ |

* These are obtained from the respondent's zip code of residence.

For binary indicators, such as born in the US and US citizen, missing values are imputed using a logistic regression. For ordered categorical variables, such as education, household composition, and household income, missing values are imputed using an ordered logistic regression. For unordered categorical variables, such as marital status, race-ethnicity, and work status, missing values are imputed using a multinomial logistic regression. Census region are never missing, as they are obtained from respondents' zip codes of residence.

Each UAS survey data set including sample weights also contains a binary variable (imputation flag) indicating whether any of the recoded socio-economic variables used for poststratification has been imputed or taken from UAS administrative records not available to data users. The mean of the imputation flag across UAS surveys is typically between 0.01 and 0.015.

## 2.4. Raking/Trimming Algorithm

We adopt a **raking algorithm** to generate poststratification weights. This procedure involves the comparison of target population relative frequencies and actually achieved sample relative frequencies on a number of socio-demographic variables independently and sequentially. More precisely, starting from the base weights, at each iteration of the algorithm weights are proportionally adjusted so that the distance between survey and population marginal distributions of each selected socio-demographic variable (or raking factor) decreases. The algorithm stops when survey and population distributions are perfectly aligned. A maximum of 50 iterations is allowed for perfect alignment of survey and population distributions to be achieved. If the process does not converge within 50 iterations, no sample weights are returned and attempts using different raking factors are made.

Our raking algorithm trims extreme weights in order to limit variability and improve efficiency of estimators. We follow the general weight trimming and redistribution procedure described by Valliant, Dever and Kreuter (2013).[4]

Specifically, we define $N = N_w + N_{nw}$ the total sample size, where $N_w$ is the number of respondents who receive a weight, and $N_{nw}$ is the number of respondents with a pre-assigned weight of zero (non-Native Americans in batches 2 and 3; all respondents in batch 4). Indicating with $w_i^{rak}$ the raking weight for respondent $i = 1, \dots, N_w$, and with $\bar{w}^{rak} = \frac{1}{N_w}\sum_{i=1}^{N_w} w_i^{rak}$ the sample average of raked weights,

I. We set the lower ($L$) and upper ($U$) bounds on weights equal to the 10th and 90th percentile of the $w_i^{rak}$ distribution, respectively. While there is no consensus on which threshold should be used for trimming, these are among those often mentioned in the literature and adopted by other surveys (Battaglia et al., 2009).[5]

[4] Valliant, R., Dever, J. A., and Kreuter F., (2013) *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

[5] Battaglia, M. P, Izrael, D., Hoaglin, D. C., and Frankel M. R., (2009) "Practical Considerations in Raking Survey Data." *Survey Practice*, 2009 (June). http://surveypractice.org/2009/06/29/raking-survey-data/.

II. We reset any weights smaller than the lower bound to L and any weights greater than the upper bound to U::

$$w_i^{trim} = \begin{cases} L & w_i^{rak} \leq L \\ w_i^{rak} & L < w_i^{rak} < U \\ U & w_i^{rak} \geq U \end{cases}$$

III. We compute the amount of weight lost by trimming as $w^{lost} = \sum_{i=1}^{N_c}(w_i^{rak} - w_i^{trim})$ and distribute it equally among the respondents whose weights are not trimmed.

IV. If these new weights are all within the interval [L,U], no further adjustment is performed. If any of these new weights are outside the interval [L,U], the trimming procedure is repeated iteratively until all weights are within the interval [L,U] or until the maximum number of 50 iterations is reached.

While raking weights can match population distributions of selected variables, trimmed weights typically do not. We therefore iterate the raking algorithm and the trimming procedure until post-stratification weights are obtained that respect the weight bounds and align sample and population distributions of selected variables. This procedure stops after 50 iterations if an exact alignment respecting the weight bounds cannot be achieved. In this case, the raked weights will ensure an exact match of (weighted) survey relative frequencies to their population counterparts, but the weights will not be within the pre-determined bounds.

## 2.5. Final Poststratification Weights

Indicate by $w_i^{post}$ the final poststratification weight for respondent $i$, obtained by applying the raking algorithm to the base weights and after iterating the raking/trimming procedure as described above. Each weighted UAS survey data set includes final poststratification weights relative to their sample mean. Formally, this is

$$w_i^{final} = \frac{w_i^{post}}{\left(\frac{1}{N_w} \sum_{j=1}^{N_w} w_j^{post}\right)}$$

For the $N_w$ respondents who receive a weight, and $w_i^{final} = 0$ for the $N_{nw}$ respondents who do not. Hence, relative final poststratification weights sum to the size of the sample of respondents who receive a weight ($N_w$) and average to 1 within that sample.

---

## Default Weights

Poststratification weights for UAS surveys including all batches are generated using the following set of raking factors (as defined in Table 2):

- *Gender*
- *Race-ethnicity*
- *Age*
- *Education*
- *Census region (augmented)*

For UAS surveys including only batches targeting the U.S. population (batches 1, 5-12, 17, and 20-25), the set of raking factors includes gender, race-ethnicity (excluding Native Americans), age, education, and Census region (Northeast, Midwest, South, West).

We have carried out extensive testing and concluded that raking weights produced by this combination of factors perform well across different dimensions. In particular, they exhibit moderate variability, thereby leading to better precision of weighted estimates, and allow matching the distributions of variables not used as raking factors in a satisfactory manner, thereby improving overall representativeness. Our Monte Carlo studies have shown that these desirable properties are robust to sample sizes ranging from 500 to 2,000 respondents, an interval that includes most of the UAS surveys.

For UAS surveys currently in the field, default weights can be obtained by sending a request to uas-weights-l@mymaillists.usc.edu.

For completed surveys, the data set with default weights is available for download on the UAS webpage.

Table 3: Unweighted and Weighted Distributions of Raking Factors (%)

| Variable | Unweighted | Base weights | Final weights* |
|---|---|---|---|
| Male | 40.43 | 41.26 | 48.54 |
| *Race-ethnicity* | | | |
| White | 64.56 | 73.34 | 62.05 |
| Black | 8.35 | 9.64 | 12.04 |
| Other | 6.91 | 4.87 | 6.99 |
| Hispanic | 13.64 | 7.53 | 16.45 |
| Native American | 6.54 | 4.62 | 2.47 |
| *Age (years)* | | | |
| 18–39 | 28.68 | 25.63 | 37.32 |
| 40–49 | 19.55 | 19.43 | 15.81 |
| 50–59 | 18.21 | 18.96 | 16.35 |
| 60+ | 33.56 | 35.98 | 30.52 |
| *Education* | | | |
| High school or less | 20.97 | 18.32 | 38.71 |
| Some college | 35.58 | 34.07 | 26.40 |
| Bachelor or more | 43.45 | 47.61 | 34.90 |
| *Census region (augmented)* | | | |
| Northeast | 11.32 | 14.44 | 17.41 |
| Midwest | 22.51 | 27.01 | 20.66 |
| South | 28.19 | 36.41 | 38.20 |
| West, excl CA | 8.66 | 12.46 | 11.91 |
| CA, excl LAC | 12.57 | 6.90 | 8.80 |
| LAC | 16.74 | 2.79 | 3.02 |

*Weighted percentages using final weights match their population counterparts by construction, given the convergence of the raking/trimming algorithm.

Table 3 shows the distributions of the default raking factors in the entire UAS sample as of June 2022. The demographic information used in this exercise is obtained from the My Household survey as of June 1st, 2022 (each UAS member is asked to update their My Household survey every quarter). In column 1, we report unweighted percentages, in column 2 the percentages after applying base weights, and in column 3 the percentages after applying final weights. By construction, these match population benchmarks, which are taken from the Basic Monthly Current Population Survey (months from October 2021 to April 2022). The unweighted sample size (excluding respondents from batch 4 and non-Native American households recruited through

Native American batches) is 9,281 and the effective sample sizes using base weights and final poststratification weights are 5,692 (61% of 9,281) and 5,379 (58% of 9,281), respectively.

Table 4 shows the distributions of some of the demographics included in the My Household survey (transformed as described in Table 2) that are not used as raking factors. In column 1, we report unweighted percentages, in column 2 the percentages after applying the final weights, and in column 3 the percentages in the population (taken from the CPS for all variables, except for the urban indicator, whose benchmark is taken from the 2013–2017 5-year ACS aggregate file).

Table 4: Unweighted and Weighted Distributions of Selected Variables Not Used as Raking Factors (%)

| Variable | Unweighted | Final weights | Population |
|---|---|---|---|
| *Household size* | | | |
| 1 member | 17.06 | 15.90 | 19.11 |
| 2 members | 40.78 | 40.82 | 34.53 |
| 3 or 4 members | 30.80 | 31.28 | 33.05 |
| 5 or more members | 11.36 | 12.00 | 13.32 |
| *Household income* | | | |
| <$30,000 | 23.39 | 27.90 | 18.61 |
| $30,000–$59,999 | 23.55 | 24.33 | 24.81 |
| $60,000–$99,999 | 24.10 | 23.01 | 23.57 |
| ≥$100,000 | 28.95 | 24.76 | 33.01 |
| *Work status* | | | |
| Working | 59.15 | 58.54 | 58.89 |
| Retired | 22.43 | 19.82 | 20.09 |
| Other | 18.38 | 21.64 | 21.02 |
| *Employment type* | | | |
| Government | 18.34 | 17.08 | 13.51 |
| Private | 69.02 | 71.67 | 76.14 |
| Self-employed | 12.65 | 11.25 | 10.34 |

## Custom Weights

Data users can customize the weighting procedure and obtain weights that better suit the goals of their research and data analysis. Custom weights can be obtained by choosing which socio-demographic variables should be used by the raking algorithm to generate post-stratification weights. Raking can be performed on one-way marginals, by matching population distributions of single socio-demographic variables, such as gender or education, as well as on two-way marginals, by matching the distributions of interaction variables, such as gender × education. The preferred set of raking factors may feature both single and interaction variables, such as, for instance, race-ethnicity and gender × education. The use of two-way marginals corrects for discrepancies between distributions referring to specific subgroups that would not be accounted for by using one-way marginals alone.

Custom weights requests should be sent to uas-weights-l@mymaillists.usc.edu, alongside with the preferred set of raking factors. This set is **limited to the variables listed in Table 2 and to a maximum of 6 variables**. (single variables, interaction variables or a combination of both). Restrictions on the number and type of raking factors are imposed to ensure convergence of the algorithm and to reduce weight variability.

## Weighting Output

Each weighted UAS survey data set includes the following variables:

- *base_weight*

  Base weights correcting for unequal sampling probabilities.

- *imputation_flag*

  A binary variable indicating whether any of the variables used for poststratification has been imputed.

- *cps*

  A variable indicating the CPS monthly surveys used to obtain benchmark distributions for poststratification.

- *final_weight*

  Relative final poststratification weights ensuring representativeness of the survey sample with respect to key pre-selected demographic variables.

NOTE: base_weight and final_weight are both zero for non-Native American respondents in recruitment batches 2 and 3, and all respondents in recruitment batch 4.