OVERVIEW OF THE U.S. CENSUS 2000 LONG FORM WEIGHTING

Steven P. Hefter and Philip M. Gbur, U.S. Bureau of the Census
Steven P. Hefter, U.S. Bureau of the Census, Washington, D.C. 20233-7600

Key Words: Estimation, Raking, Long Form

I. Introduction and Background

The U.S. Census Bureau conducted Census 2000 on April 1, 2000. Respondents were enumerated using one of two general types of census questionnaires: the long form or the short form. While the short form only required basic, minimal information such as name, age, and sex, the long form asked more detailed questions regarding such items as employment status, disability, income, etc. These answers, when weighted, allow for a variety of socio-economic and demographic estimates to be made at many levels of geography.

A systematic sample of addresses on the Decennial Master Address File and of housing units in the field received a long form questionnaire. After the data was collected from these questionnaires it was weighted using the iterative proportional fitting methodology, commonly referred to as raking.

Variances were estimated for a subset of resulting long form estimates using a successive difference replication methodology and generalized for use with all estimates.

In the following sections we present an overview and a description of the long form sampling and weighting procedures for Census 2000. We present selected results from both the long form sampling and weighting and describe the components of each which were changed from 1990. We also discuss methodologies that were developed for 2000 but not utilized.

In general, the Census 2000 design was similar to 1990, but revisions were introduced to improve selected aspects of the 1990 process. In addition we discuss possible implications for data users resulting from implementation of the weighting methodology.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

II. Sample Design

A. 2000 Methodology and Results

The addresses that were to receive the long form questionnaire were chosen either by taking a systematic, variable rate sample of addresses from the Decennial Master Address File (computer based sampling) or were chosen in subsequent field sampling operations. The computer based sampling occurred on a flow basis beginning in July of 1999 and concluding in April of 2000.

The ultimate goal was to sample roughly one out of every six addresses in the U.S. and Puerto Rico. This was achieved through appropriate application of the selected sampling rates to each Long Form Sampling Entity (LFSE) - such as a city, county, or school district - or census tract. The sample design included four sampling strata using the rates of 1-in-2, 1-in-4, 1-in-6 and 1-in-8.

Application of the long form sampling rates was based on a measure of size for each LFSE and census tract. An estimate of the number of occupied housing units was used as the measure of size. An interim census tract delineation was used, as updated census tracts were not yet available [1].

The sampling strata cutoff points were chosen based on an analysis of the range of coefficients of variation (CVs) obtained from simulation research. The sampling rates were applied at the collection block level. For blocks that fell into more than one sampling stratum, we applied the higher sampling rate.

The sampling strata and their cutoff points were:

- 1-in-2 for LFSEs < 800 housing units;
- 1-in-4 for LFSEs between 800 and 1200 housing units; and if not 1-in-2 or 1-in-4; then
- 1-in-6 for census tracts < 2000 housing units; and
- 1-in-8 for census tracts $\geq$ 2000 housing units.

The following sampling rates were used for certain enumeration areas and special populations irrespective of the size of their associated LFSE:

a. Update/Leave adds and List/Enumerate areas were sampled according to the sampling rate of the blocks in the assignment area (AA). When an AA included more than one sampling stratum, the higher of the rates was used for the entire AA.

b. Group Quarters (such as nursing homes and college dormitories) and Service Sites (such as shelters and soup kitchens) were sampled at a 1-in-6 rate.

For further details of the long form sampling methodology see [2].

The results of the computer based sampling for Census 2000, including Puerto Rico, are given in Table 1. Table 1 gives the count of addresses found on the Decennial Master Address File which was the sampling frame, the number of addresses selected in sample, and the percent of addresses in sample [3].

Table 1: Summary of Census 2000 Computer Based Designated Long Form Sample

Universe	Addresses in Sample	Percent in Sample
123,411,977	21,107,353	17.1

Table 1 shows that we met our goal in the computer based portion of the sampling with 17.1% of the addresses found on the Decennial Master Address File selected.

B. Changes from 1990

There are three major differences in the long form sample design between the 1990 Census and Census 2000. First, the sampling rate cutoffs for the long form were based solely on occupied housing unit estimates, not on a mix of population and housing unit counts as in 1990. Ideally, the cutoffs would have been based on population counts but reasonable counts were not available for all areas at the level of geography at which we sampled. Therefore, estimates of the number of occupied housing units were used for all areas to maintain consistency for all geographic areas.

In 1990 three sampling rates were used, 1-in-2, 1-in-6 and 1-in-8. In addition to these three rates, a 1-in-4 sampling rate was added for Census 2000 [1]. This rate was added to achieve more reliable estimates for LFSEs that would have been sampled at 1-in-6 using the 1990 rates, and to reduce respondent burden in the medium sized LFSEs that would have been sampled at 1-in-2.

For sampling purposes, school districts were treated as LFSEs [1]. In 1990, school districts were not considered in the sample design. Since school districts may receive funding as separate entities, this is expected to produce more reliable long form estimates for these areas.

C. Changes from Initial Design

The initial sample design included 10 percent oversampling which yielded sampling rates of 1-in-1.8, 1-in-3.6, 1-in-5.4, and 1-in-7.2. This was motivated by the 10 percent sample loss – or long form non-response rate – observed in the 1990 census [4]. Sample loss was expected to be at a similar rate in 2000. Long form sample loss occurs when a respondent completes a long form questionnaire with data only for the 100 percent questions. Subsequent to the computer based sampling performed using these rates, we felt that introducing the additional cost and operational complexities of oversampling did not justify the slight gain in reduced variance, therefore we resampled the initial address file using the original sampling rates [6].

III. Weighting

A. Overview

As in every census since 1940, when we introduced content sampling [7], the iterative proportional fitting methodology [8] was used in Census 2000 to estimate various detailed characteristics of the entire country based on the long form sample. We carried out the iterative proportional fitting methodology, also known as raking, within relatively small geographic areas called final weighting areas.

B. Design

Initial Weighting Area Formation

During the first step in the weighting, we partitioned the U.S. and Puerto Rico into geographic areas referred to as initial weighting areas. Initial weighting areas were defined to be all records within a tabulation block group and sampling stratum (rate) combination.

Augmentation

To ensure that we had basic minimal housing unit, group quarters person, and service based enumeration person sample sizes, augmentation of the long form sample, using a set of predetermined rules, occurred in initial weighting areas that did not meet the pre-specified criteria. This was done to attain a minimum observed sampling rate within each initial weighting area, and to reduce weight variation within these groups. Records were chosen for augmentation through a systematic sample of long forms first and short forms if the number of records required exceeded the number of eligible long forms. Augmentation of sample counts used the smallest number of records needed to reach the desired minimum observed sampling rate. If necessary, after the weighting was completed, sample data was imputed for census records chosen in augmentation.

Final Weighting Area Formation

Subsequent to augmentation, final weighting areas, the geographic level within which we conducted the weighting, were formed within counties by combining

initial weighting areas until each had a minimum of 400 sample persons. Final weighting areas were generally in close agreement with census tabulation areas. If necessary, we allowed small counties with fewer than 400 sample persons to stand alone as final weighting areas. For Census 2000 a total of 65,343 final weighting areas were formed.

Pass 2 Augmentation
Upon formation of the final weighting areas, we checked the number of occupied housing units, vacant housing units, and group quarters and service based enumeration persons in sample against the corresponding number of 100 percent cases within each final weighting area. In a small number of final weighting areas we found that, although there was at least one 100 percent case of a given type, there were no cases of the same type in sample. To ensure that there was at least one sample record in the weighting area to carry the weight of the associated 100 percent count we implemented a second augmentation procedure at the final weighting area level. Again, sample data was imputed for cases selected in this procedure we called pass 2 augmentation.

Initial Weight Calculation
For each sample unit within each initial weighting area, we set an initial weight equal to the inverse of the observed sampling proportion. This was done separately for persons in housing units, persons in group quarters, persons enumerated at service sites, occupied housing units, and vacant housing units. After augmentation and initial weight calculation, the weighting proceeded separately for persons, occupied housing units, and vacant housing units.

Person Weighting Matrix Formation
Within each final weighting area, we formed a four-dimensional person weighting matrix using the following characteristics: 21 levels of household type (such as family with own children and family without own children) by household size; three levels of sampling rate (1-in-6 and 1-in-8 were combined); whether or not the person is the householder; and Hispanic origin by six levels of race by 26 levels of age and sex. Thus, every person weighting matrix contained 39,312 cells.

Occupied Housing Unit Weighting Matrix Formation
For occupied housing unit weighting, we created a three dimensional matrix using the following variables: 19 levels of household type by household size; three levels of sampling rate; and tenure by Hispanic origin by six levels of race yielding 1,368 cells.

Vacant Housing Unit Weighting Vector Formation
Vacant housing units were weighted in a three cell vector using a one step proportional adjustment. The cells of the vector were defined as: vacant for sale; vacant for rent; other vacant.

Three types of counts or totals were associated with each cell of the two matrices and the vacant housing unit vector: the 100 percent count, the uninflated (unweighted) sample count, and the initially inflated (initially weighted) sample count. These were summed within each cross-classification to produce three marginal totals for each cross-classification within the weighting structures.

Collapsing
Before raking, we tested the marginal totals against predefined collapsing criteria. If the uninflated sample marginal totals were not "large" enough, or the ratio of the 100 percent marginal total to the initially inflated sample marginal total failed a collapsing test, then we combined failing classifications (or cross-classifications) with other classifications (or cross-classifications) within the same category (e.g. race).

Raking
Raking is an iterative proportional adjustment of the cross-classified initially inflated sample cell counts and was used in the person and occupied housing unit weighting. We raked the initially weighted sample count to the 100 percent count in several stages.

The interior cell counts within a classification were multiplied by the ratio of the control or 100 percent total (for that classification) to the initially inflated sample total (for that classification). An iteration of the raking consisted of one stage of adjustment for each dimension. Each stage adjusted all interior cell counts of a dimension by the appropriate ratio. In Census 2000, the raking continued until all weighted sample marginals were within 0.1 percent of the corresponding control marginal or after a total of five iterations, whichever was reached first. We also designed the raking so that the dimension that included race and Hispanic origin was adjusted last. This allowed for complete agreement between the weighted sample count and the 100 percent count for each collapsed classification in this dimension.

Weight Integerization
After raking was complete, a controlled rounding algorithm was implemented. This allowed for weights to be integer valued, which traditionally long form weights are, and also maintained the weighted sample totals across final weighting areas.

Further details of the long form weighting procedure can be found in [9].

C. Accounting for Coverage Error
In the event that the results of the Accuracy and Coverage Evaluation survey were incorporated into census data products, specifically the PL 94-171 "redistricting" data, we developed a methodology to incorporate coverage error

correction into the long form weights [10,11]. The methodology called for the long form person and housing unit weights to be multiplied by the respective coverage correction factor prior to integerizing the weights. Upon the Census Bureau's decision not to correct the redistricting data for coverage error, this component of the weighting was dropped.

D. Changes from 1990

In Census 2000, we had to accommodate multi-race responses in long form weighting for the first time. Since each person could only be represented in one and only one cell of the weighting matrix, a multi-race classification method was developed. Respondents affiliating with more than one race group were placed into the largest nonwhite race group they checked. Note that this was done only for weighting and that all multi-race respondents were tabulated in the appropriate multi-race category.

One of the major methodology changes from 1990 was in the structure of the weighting matrices. In 1990, the final dimension that was raked contained race by Hispanic origin by sex/age. Due primarily to concerns with the 1990 estimates of Hispanic/nonHispanic, and to a lesser degree the multi-race classification methodology, we changed the ordering of the dimension, placing Hispanic origin over race. This allowed the Hispanic/nonHispanic distinction to remain separate – or survive collapsing – with greater frequency then it would have under the 1990 design. The trade off for this was that race was collapsed more often within Hispanic/nonHispanic.

Due to changes in the content of the short form from 1990 to 2000, some items used to define the weighting matrices in 1990 were unavailable. Specifically, questions asking the number of units in the structure and the value of the house or amount of rent were dropped from the short form, precluding their use as weighting controls in the occupied housing unit weighting matrix.

Another change from 1990 was in the final weighting area formation. In 1990, where possible, final weighting areas respected place boundaries, whereas in 2000 initial weighting areas were combined within tabulation block group, then census tract, and then county.

Also, block code assignment methodology was changed from 1990. Thus, while in 1990 the weighting began with collection geography and the mapping to tabulation geography occurred during the weighting, in 2000 the process was carried out solely with tabulation geography.

In addition, we made several revisions to the collapsing criteria for the raking matrices and modified the raking stopping criteria. In 1990, the raking was stopped after two iterations. For Census 2000 we used two criteria: 1) weighted sample marginal totals being within a specific distance from the 100 percent marginal totals; and 2) five iterations of the raking. We ended the raking when one of the criteria was met. It was expected that allowing a maximum of five iterations of the raking would result in more consistent long form estimates.

IV. Selected Results

A. Race

Table 2 gives the unweighted long form sample count, final weighted long form estimate, and the 100 percent count for persons by selected race groups. The last column of the table shows the absolute and percentage differences between the 100 percent total and the final weighted long form sample estimate.

Table 2: National Summary of Long Form Weighting Results by Selected Race Groups

Race	Unweighted Sample[1]	Final Weighted Sample	100% Count	Difference (% diff)
White Only	33,750,956	211,297,184	211,460,626	-163,442 (-0.08)
Black Only	4,651,385	34,371,190	34,658,190	-287,000 (-0.83)
AIAN[2] Only	467,668	2,440,586	2,475,956	-35,370 (-1.43)
Asian Only	1,345,770	10,207,328	10,242,998	-35,670 (-0.35)
NHPI[3] Only	54,308	378,524	398,835	-20,311 (-5.09)
Other Only	2,144,115	15,433,825	15,359,073	74,752 (0.49)
Two or More	1,045,247	7,293,269	6,826,228	467,041 (6.84)

[1]Unweighted sample count includes augmented cases
[2]AIAN is the American Indian and Alaska Native race group
[3]NHPI is the Native Hawaiian and Pacific Islander race group

In Table 2, the percent difference between the 100 percent total and the weighted sample total for persons marking only one race ranges from -5.09% for the NHPI race group to 0.49% for the "other only" group. Every "single" race group except "other only" is under represented in the sample. Persons marking two or more races are over represented in the census sample by 6.84%.

Table 2 shows the effects of several factors. The tabulation cells shown in table 2 do not correspond to the weighting

cells, therefore it is not unreasonable that we observe differences. These results are also likely influenced by the frequent collapsing of the "other only" group and to placing multi-race respondents into the largest nonwhite race group in the weighting area that they marked. The single race groups consistently had their weights reduced by this method, while the multi-race respondents had their weights increased. This interacted with the collapsing methodology which followed the same basic reasoning and collapsed failing race groups into the largest nonwhite, nonother race group in the weighting area.

Note that when both the Hispanic and nonHispanic classifications survived, race collapsing was done independently within each. If the Hispanic origin classification failed, all persons in that classification were placed into the largest nonwhite, nonother, nonHispanic race group in the weighting area. If the nonHispanic classification failed, each nonHispanic race was placed into the corresponding Hispanic race group, maintaining the sex/age classification. We may also see differences due to differential response rates between the long form and the short form by race.

B. Hispanic Origin Totals

The ordering of the variables used in the person weighting matrix appears to have affected the Hispanic/nonHispanic estimates in a positive way. Table 3 shows the national Hispanic and nonHispanic unweighted sample count, final weighted sample count, 100 percent count, and absolute and percentage difference between the 100 percent count and the final weighted sample count.

Table 3: National Summary of Long Form Weighting Results by Hispanic/NonHispanic Origin

Origin	Unweighted Sample[1]	Final Weighted Sample	100% Count	Difference (% diff)
Hispanic	4,789,118	35,261,281	35,305,818	-44,537 (-0.13)
NonHispanic	38,670,331	246,160,625	246,116,088	44,537 (0.02)

[1] Unweighted sample count includes augmented cases

Table 3 shows that the estimates of Hispanics underestimate the total by a relatively small 0.13% as compared to the 2.03% seen in 1990 [4]. Similarly, the nonHispanic estimate was only 0.02% above the 100 percent count.

V. Summary

Overall the long form weighting methodology and implementation worked well. We were able to accurately assign weights to over 43 million people and over 18 million housing units within the budgeted time and with few surprising results.

Differences observed between the long form estimates and the 100 percent count for population sub-groups were expected. Due to the final weighting area formation methodology, long form estimates of total population may differ from the 100 percent total for places smaller than county.

It appears that the changes implemented had the desired effects. Of particular significance is the consistency of the Hispanic/nonHispanic estimates in comparison to the 1990 results.

Acknowledgments

The authors wish to thank the following people:

Michael Clark, from the Decennial Systems and Contracts Management Office who, with the utmost dedication and perseverance, programmed and implemented the computer software for the Census 2000 long form weighting.

Rajendra P. Singh, who's substantial input and direction aided in writing and refining this paper.

Scot Alan Dahl, whose thorough review and attention to detail is greatly appreciated.

References

[1] U.S. Census Bureau, "Long Form Sampling Decisions for Census 2000 Dress Rehearsal and Census 2000", internal memorandum for Thompson from Killion, Census 2000 Decision Memorandum #40, January 28, 1998.

[2] U.S. Census Bureau, "Long Form Sampling Specifications for Census 2000", internal memorandum for Longini from Hogan, DSSD Census 2000 Procedures and Operations Memorandum Series #LL-5, November 11, 1999.

[3] U.S. Census Bureau, "Census 2000 Long Form Computer Based Sampling: Approval and Summary of Results", internal memorandum for Lynch from Gbur, Census 2000 Procedures and Operations Memorandum Series #LL-9, August 28, 2002.

[4] Schindler, E., Griffin, R., Swan, C., "Weighting the 1990 Census Sample," Proceedings of the Government Statistics Section of the American Statistical Association, pp. 726-731, 1992.

[5] U.S. Census Bureau, "Oversampling for Long Form Sample Loss in Census 2000", internal memorandum for Thompson from Hogan, Census 2000 Decision Memorandum #101, April 17, 2000.

[6] Stephan, F.F., Deming, W.E., Hansen, M.H., "The Sampling procedure of the 1940 Population Census", *Journal of the American Statistical Association*, December, 1940, Vol. 35, pp. 615-630.

[7] Stephan, F.F., Deming, "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known", *The Annals of Mathematical Statistics*, 11, pp.427-444.

[8] U.S. Census Bureau, "Long Form Weighting Specifications for Census 2000", internal memorandum for Longini from Singh, DSSD Census 2000 Procedures and Operations Memorandum Series #LL-10, September 3, 2002.

[9] U.S. Census Bureau, "Reducing Coverage Error in Long Form Person Estimates for the Census 2000 Dress Rehearsal and Census 2000", internal memorandum for Thompson from Hogan, Census 2000 Decision Memorandum #55, June 10, 1998.

[10] U.S. Census Bureau, "Reducing Coverage Error in Long Form Housing Unit Estimates for the Census 2000 Dress Rehearsal and Census 2000", internal memorandum for Thompson from Hogan, Census 2000 Decision Memorandum #59, July 27, 1998.