POPULATION AGING RESEARCH CENTER

UNIVERSITY of PENNSYLVANIA

Using Successive Censuses to Reconstruct the African-American Population, 1930-1990

Samuel H. Preston, Irma T. Elo and Lynn Gale

PARC Working Paper Series

WPS 96-06

" The authors acknowledge the support of the National Institutes of Health - National Institute on Aging, Grant number P30 AG12836, B.J. Soldo, P.I."

3718 Locust Walk • Philadelphia, PA 19104-6298 • Tel 215-898-6441 • Fax 215-898-2124

Population Aging Research Center, University of Pennsylvania
Working Paper Series No. 96-06

December 1996

Using Successive Censuses to Reconstruct the African-American Population, 1930-1990

Samuel H. Preston*
Irma Elo*
Lynn Gale**

*Population Studies Center, University of Pennsylvania
**Center for Advanced Study in the Behavioral Sciences, Stanford, Cal.

Revised version of this paper was presented at the Annual Meeting of the Population Association of America, San Francisco, California, April 6-8, 1995.

Much of this paper was prepared while the senior author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. We are extremely grateful to Gregory Robinson of the U.S. Bureau of the Census for supplying certain of the data used in this analysis and for keeping us abreast of Census procedures. Herbert Smith, Lincoln Moses, and Andrew Foster provided valuable advice. Ira Rosenwaike, Mark Hill, Laura Shrestha, and Julie Goldsmith made many useful contributions to the project on which this paper is based. The research was supported by a grant from the National Institute of Aging, AG10168.

The Census Bureau's program to estimate the completeness of decennial census counts for age, sex, and race groups relies principally upon what it terms "demographic analysis." The essence of this approach is to introduce extraneous information on the number of births, deaths, and migrations, derived from non-census sources, to estimate the true size of each birth cohort at the time of a census (Robinson et al., 1993; Himes and Clogg, 1992). Comparison of this alternative estimate to the census count provides an estimate of the degree of under- or over-enumeration in the census, often termed the census undercount. Acceptance of the estimated undercount implies that the census itself is irrelevant to estimating the true size of the population; whatever deficiencies it contained would be accurately and completely revealed by comparison to the estimate based on demographic analysis.

Unfortunately, as the Census Bureau frankly acknowledges, the data employed in demographic analysis are also subject to error. The most important source of error applies to the birth series. The Birth Registration Area of the United States was not completed until 1933 and tests of birth registration completeness that were conducted in conjunction with censuses of 1940 and 1950 revealed a substantial degree of underregistration, especially for African-Americans (U.S. Bureau of the Census, 1943, 1953). The degree of registration completeness is uncertain, and the Census Bureau has recently modified its estimates of registration completeness based upon the 1940 tests for African-Americans (Robinson et al., 1993). The uncertainty not only affects historical estimates of undercounts but contemporary estimates as well, since persons born in 1940 reached age 50 in 1990.

This paper explores an alternative approach to estimating census undercounts for African-Americans, who have persistently shown the highest undercount rates in demographic analysis. Rather than ignoring census counts in estimating the true size of cohorts at particular census dates, it makes census counts themselves the basis of estimation. In particular, by examining census counts for various cohorts in successive censuses, it identifies systematic errors associated with age and with census date and develops a single preferred estimate of cohort size at each census date.

Our estimates begin with the census of 1930. We cannot extend the analysis to earlier dates because the Death Registration Area (DRA) was not completed until 1933 and we require death counts in order to relate expected cohort size in one census to that in another. (The death series is completed back to 1930 by adding deaths for Texas, the one state that was missing from the DRA between 1930 and 1933.) We make no use of data from the birth registration system, whether corrected or uncorrected. Instead, we use corrected birth registration data beyond 1950, when these data have relatively low uncertainty, primarily as a useful test of our procedures. However, for cohorts born before 1930, who contribute most of the observations used in this paper, national birth registration data are not available; for those born between 1930 and 1950, birth registration completeness is uncertain. It is these cohorts - and especially those born between 1905 and 1950, for each of which we have 5-7 observations on cohort size in censuses from 1930 to 1990 - where the present set of estimates is expected to prove most useful. Estimates are made separately for each sex, in five-year wide age intervals.

Data

For each of the census years 1930-1990, we obtained population counts for African-Americans by age and sex from both published sources and unpublished data provided to us by the Bureau of the Census.[1] Each of these censuses occurred on April 1. We use U.S. borders as defined in 1960, so that estimates of the African-American population by sex and age in Alaska and Hawaii are added for the census years 1930, 1940 and 1950. For the 1980 and 1990 Censuses, the Bureau of the Census has released two different population counts by race because a large percentage of the Hispanic-origin population wrote in a response to the census question on race that identified ethnic origin rather than race. For the 1980 and 1990 census population counts, we use the Census Bureau's unmodified race series, because we believe the unmodified

[1] The 1970 Census data used in this study are based on unpublished data provided to us by the Bureau of the Census that correct for errors in the population counts of local areas discovered after the initial Census tabulations were published.

series to be more comparable than the modified series to previous censuses and to death registration data used in this paper. No similar reassignments of the Hispanic-origin population, for example, were carried out in previous censuses, although the problem seems to have been of smaller magnitude.

The annual series of vital statistics are our primary source of data on deaths by age and sex. These data were obtained from the published volumes of Vital Statistics of the United States for the period 1930-1967 and from the annual National Center for Health Statistics (NCHS) mortality data tapes containing information on each death from 1968 to 1989. Data for the first three months of 1990 were obtained from the NCHS monthly and final vital statistics reports (U.S. National Center for Health Statistics, 1990).[2] These data were then adjusted to correct for the exclusion of Texas from the Death Registration Area prior to 1933, for the omission of Alaska and Hawaii from the U.S. statistics prior to 1959 and 1960, respectively, and for the lack of racial detail on deaths for New Jersey residents in 1962 and 1963. When vital statistics data were available only by five-year age groups, Sprague multipliers were used to allocate

[2] We have assumed that infant death registration completeness equals birth registration completeness. To estimate completeness, we used birth series adjusted for birth underregistration provided to us by the Bureau of the Census for the period 1940-1990. These series are based on birth registration test results from 1950 and 1964-68 for the period 1950 through 1990 and on Passel's (1992a) estimates for the period 1940-1950. In these series, the race of the child is assigned according to the race of the father. Birth registration completeness is then determined by dividing the number of registered births by the adjusted birth series. For the 1930-40 period, we assumed a steady pace of improvement in completeness. The implied birth registration completeness for the 1930-40 period, assumed equal for males and females, is given below. Details are available from the authors.

Period	Completeness
1930-31	.833
1931-32	.836
1932-33	.839
1933-34	.843
1934-35	.846
1935-36	.849
1936-37	.853
1937-38	.856
1938-39	.859
1939-40	.863

deaths into single years of age (Shryock, Siegel, and Associates 1976).[3]

Because censuses during this period do not occur at the beginning of the year, the calendar year data on deaths by single years of age had to be separated into single-year groups defined at the April 1 census dates. To compute the required separation factors, we assumed deaths within a one-year block of age for a particular calendar year to be evenly distributed by time of occurrence and age of the decedent, i.e., we assumed that the lexis surface is flat in both dimensions (time and age). Beginning in 1968, NCHS data are available by month of death and thus the assumption of a lexis surface that is flat over time can be avoided.

For estimates of intercensal migration, we rely primarily on data provided to us by the Bureau of the Census for the period 1940-1990. For each of the intercensal decades these data are available by sex and race for five-year birth cohorts defined at census dates.[4] We have made one modification in the Census Bureau's estimates. Our estimates, which are based on the African American Puerto-Rican born population, are designed to take account of both net migration between Puerto Rico and the United States

---

[3] From 1930 to 1967, deaths by single years of age for African Americans are published for children under five and by five year age groups thereafter. However, beginning in 1951, NCHS published nonwhite deaths by single years of age at ages 85 and above. Deaths among African-Americans make up the great majority of nonwhite deaths at these ages; thus we used the single-year age distribution of nonwhite deaths to allocate African-American deaths ages 85 and above between 1951 and 1967. Unknown ages at death were allocated into five-year age groups based on the age distribution of deaths of known age. From 1968 to 1989 deaths by single years of age are available from the NCHS mortality data tapes. For the first three months of 1990, deaths in five-year age groups were allocated into single years of age based on the age distribution obtained from the NCHS mortality data tapes for 1988 (for further detail, see Elo and Preston, 1994).

[4] The Census Bureau's estimates of intercensal migration are based on separate estimates of legal alien migration, refugees and parolees, not civilian citizen migration, net Puerto-Rican migration, net flows of foreign students, net movements of US armed forces oversees, legal emigration and net movements of illegal aliens, although the detail on components varies by decade (for a summary of the Bureau's procedures, see Himes and Clogg, 1992). We included the Bureau's estimates of illegal migration in our analyses. Because the number of estimated migrants at oldest ages is very small relative to the volume of deaths, we allocated open-ended age intervals (variously 65+ and 75+) into five-year age groups based on the simple assumption that net migration rates were constant by age within open-ended categories. Sprague multipliers were then used to allocate data by five-year age groups into single years of age.

and changes in racial classification of Puerto Rican-born individuals in the various censuses. This approach was taken to minimize the effects of changes in the racial classification of Puerto-Rican born individuals in censuses over time.

Our demographic accounting identities assume that each individual identified as African American will remain a member of the African American population, and be so identified in all statistical systems up to and including the point of out-migration or death. Thus, the integrity of our accounting identities depends not only on the comparability of race classification from one census to the next, but also on the comparability of race classification systems between censuses, death and migration statistics. Fortunately, available evidence suggests that African-Americans are highly consistent in their reporting of race. A record linkage study of death certificates with the 1960 Census records, for example, found that 98.2% of African-Americans had the same race reported on the death certificate as in the Census record; the net difference was only 0.3% (U.S. National Center for Health Statistics, 1969). A more recent record linkage study of records from 12 Current Population Surveys (CPS) with the National Death Index for years 1979-85 found a similarly high correspondence in the reporting of race among African-Americans; 98.2% had the same race reported on the death certificate as in the CPS record with a net difference of 0.4% (Sorlie, Rogot, and Johnson, 1992). Census and CPS record-linkage studies have further shown that the reporting of race among African-Americans is highly consistent (see U.S. Bureau of the Census, 1964, 1975).

Methods

Designate the number of people enumerated in cohort i in the census taken at time t as C_{it}. We are considering censuses from 1930 to 1990 and age groups from 0-4 to 80-84, so that there are 119 (7x17) observations on C_{it} in the original population data matrix for each sex. These pertain to 29 different cohorts: 17 alive in 1930 and 12 five-year wide birth cohorts born in the period between the 1930 census and the 1990 census. The number of observations available on a particular cohort ranges from one (cohorts aged 75-79 and 80-84 in 1930 and cohorts born during 1980-85 and 1985-90) to seven (cohorts aged 0-4 to 20-24 in 1930).

Since the true size of a cohort changes between censuses as a result of death and migration, the observations on C_{it} in the original population data matrix are not directly comparable. In order to make them commensurate, it is necessary to add or subtract the deaths and migrations between the censuses. Comparability could be assessed at any time in the life of the cohort; we have chosen to make the comparisons at the first appearance of the cohort in the population data matrix, i.e., 1930 or at ages 0-4 or 5-9 if born subsequent to 1930. Cohorts are numbered from 1 to 29, with 1 referring to the last born (aged 0-4 in 1990), and censuses are numbered from 1 to 7, with 1 referring to the 1930 census. Thus, $C_{6,5}$ refers to the census count of the number of persons aged 5-9 in the census of 1970. Three estimates of true size of this cohort in 1970 ($X_{6,5}$) are available: $C_{6,5}$ itself; $C_{6,6} + D_{6,6}$; and $C_{6,7} + D_{6,7}$ where D_{it} refers to the cumulative deaths and net migrations in the cohort between the time of its first appearance and the census taken at time t.

Designate an estimate of the size of cohort i at its first appearance as X_i and the estimate of X_i based on the census at time t as X_{it}. One strategy to construct a final estimate of X_i, X_i*, is simply to take the mean of all available estimators. However, this strategy would ignore the evidence that is generated by the estimation strategy itself that some censuses are more complete than others (e.g., estimators based on that census tend to be higher than estimators based on other censuses) and that some age groups tend to yield estimates that differ systematically from estimates based on other age groups. We have chosen instead to model errors in census counts through a multiplicative model containing an age effect and a period (or census-specific) effect: $\hat{C}_{it} = \alpha_a \tau_t \; C_{it}$, where $\hat{C}_{it}$ is an estimate of the census count that should have been observed for cohort i in census t; C_{it} is the original census count; α_a is a multiplier for age group a (the age group occupied by cohort i at census t); and τ_t is a multiplier for the census taken at time t. Thus, $\hat{C}_{it}$ adjusts for the typical pattern of error by age and census observed over the period 1930-90.

Each estimator of X_i now has the form

$$X_{it} = \hat{C}_{it} + D_{it}, \text{ or}$$
$$X_{it} = \alpha_a \tau_t C_{it} + D_{it} \tag{1}$$

Our most important objective in the estimation is to develop a sensible estimate of X_i, X^*_i. With this estimate, we can derive all other estimates of the true size of the cohort at its multiple appearances in subsequent censuses by subtraction of cumulative deaths and migrations. We choose to solve for the set of α_a's and τ_t's that minimizes the sum of squared distances of estimators of X_i from their mean, and use the mean of the resulting estimators as the final estimate of X_i, X^*_i. This minimization is done simultaneously across 25 cohorts (all but the four cohorts with single observations); in the process, each original census count C_{it} is used once and only once.

In particular, we minimize

$$f(\alpha_1,...\alpha_{17}, \tau_1,... \tau_7) = \sum_i \sum_t (X_{it} - \overline{X}_i)^2, \quad i = 3, 4,...27, \quad t = 1...7, \tag{2}$$

where $\overline{X}_i$ is the mean of all available estimates of X_i (numbering from two to seven, depending on the cohort as noted above), and X_{it} has the construction identified in equation (1). We use an iterative approach to estimating the values of α_a and τ_t, the parameters in the estimation process. Initial values of the age multipliers, α_a, are set at 1.000. Initial values of the census multipliers, τ_t, are set at values estimated by the U.S. Bureau of the Census and summarized by Himes and Clogg (1992).[5]

[5] The initial values of τ_t are drawn from the summary of the Census Bureau's demographic analyses by Himes and Clogg (1992):

Year	Males	Females
1990	1.093	1.031
1980	1.081	1.017
1970	1.100	1.042
1960	1.096	1.046
1950	1.107	1.057
1940	1.122	1.064
1930	1.110	1.071

The iteration proceeds as follows. Given these initial values of α_a and τ_t and observed values of C_{it} and D_{it}, we use (1) to calculate starting values of $\overline{X_i}$. Treating $\overline{X_i}$ (i = 1,2,...29) in expression (2) as fixed at these values, we minimize f(α_1, ..., α_{17}, τ_1, ..., τ_7) with respect to τ_1, ..., τ_7 by taking partial derivatives, equating to zero, and solving for $\hat{\tau}_1^{(1)}$,..., $\hat{\tau}_7^{(1)}$, where the superscript refers to the iteration number. Then we use these new τ_t values to obtain new $\overline{X_i}$ values in (2), and proceed to minimize f(α_1, ..., α_{17}, τ_1 ..., τ_7) anew, this round with respect to α_1, ..., α_{17}, taking partial derivatives, equating to zero, and solving for $\hat{\alpha}_1^{(1)}$, ... , $\hat{\alpha}_{17}^{(1)}$. The iterations continue, alternating between re-estimating the set of τ's and then the set of α's, each time deriving a new set of $X_i^{*}=\overline{X_i}$ estimates.

In a paper related to our own, Passel (1992b) applies an age/period/cohort framework to Census Bureau estimates of undercounts from 1940 to 1980. While "cohort effects" in the Passel paper are a useful first approximation to errors in Census Bureau estimates of births, the specification of the model is flawed. Errors in Census Bureau estimates of the number of births would not have constant proportionate effects on undercount estimates for the cohort each time it appears, as specified in the age/period/cohort model. Instead, the proportionate error would get larger in each successive census as deaths diminish the true size of the cohort. To take an extreme example, the male cohort aged 40-44 in 1940 lost approximately 90% of its members between 1940 and 1980 (Table 3 below), so that an error in the number of births for this cohort would have 10 times the proportionate effect on the demographic estimate in 1980 as in 1940.

Results

Each successive iteration improves the fit of the model, i.e., it reduces the sum of squared

---

No value for 1930 has been estimated by the Census Bureau. The value for 1930 represents the Census Bureau's value for 1940 combined with the change in the factor between 1930 and 1940 estimated by Coale and Rives (1973). We did not use the Coale-Rives figure for 1930 directly because it appears too high and is incommensurate with the Census series (see text below and Elo and Preston, 1994).

differences between X_{it} and $\overline{X_i}$. But it does not necessarily improve the plausibility of estimates of X_i nor their consistency with other information. Table 1 shows the reduction in the sum of squared differences (SSD) after each round of iteration, i.e., after each set of iterations on both α_a and τ_t. For males, a large reduction of 78% occurs in SSD at the first iteration. Most of the improvement in fit is attributable to changes in the age-parameters, α_a. After one iteration on the τ's alone, the reduction in SSD is only 9.0%; the rest of the reduction is attributable to the α's. An "elbow" is apparent in the male pattern of SSD's, with changes in SSD becoming small after the third iteration.

At the outset, females show much less inconsistency than males; before any iterations, the female SSD is only 41% of the male value. However, the improvement in fit is also much smaller for females, so that the SSD after one iteration is nearly the same for males and females. Beyond the first iteration, it is lower for males. It requires 16 iterations before the female SSD is reduced to half of its initial value.

The improvement in fit is quite small for both males and females after the fifth iteration. Furthermore, the consistency of results with other information diminishes beyond that point, especially for females. Our interpretation of these results is that the early rounds of iteration are principally focussed on representing the large age-specific net omission rates from censuses, especially for African-American males at younger and middle ages. These modifications are necessary to make census counts consistent with one another. Later rounds are more attentive to the consistency between population counts at the older ages and deaths. Unfortunately, different patterns of age-misreporting in the basic data from censuses and deaths at older ages may make this latter effort fruitless (Elo and Preston, 1994). A large matching study of death certificates and census records in 1960, the middle of our estimation period, showed that only 44.7% of nonwhite males and 36.9% of nonwhite females had the same year of age reported in the death certificate and census form (U.S. National Center for Health Statistics, 1968). Above age 64, there were 15.4% more deaths for females and 7.1% more deaths for males using the age reported on the census than using the death certificate age.

That later rounds of iteration are primarily attentive to resolving this (basically unresolvable) inconsistency at older ages between age reporting in deaths and age reporting in censuses is suggested by the pattern of change in α_a, the age-effects. Between rounds 5 and 25 for males, the mean absolute change in α_a at ages 65-84 is .098, whereas it is only .006 at ages 0-39. For females, the equivalent figures are .092 and .008. For this reason, we present the basic results of our procedure after five iterations.

Table 2 presents the values of τ_t and α_i for males and females after five iterations. The values of α_a are graphed in Figure 1. There is substantial correspondence between the male and female age patterns of α_a, but with much more variability evident in the male series. For example, between ages 10-14 and 25-29, α_a rises by .117 for males but by only .029 for females. The results thus confirm the large relative underenumeration of African-American males in the age interval 20-39 that demographic analysis has previously suggested (e.g., Fay et al., 1988). Children in the age interval 0-4 are also relatively underenumerated. On the other hand, an overenumeration is clearly implied for persons aged 65+. This result probably reflects age overstatement of African-Americans reported at 65+ relative both to cohort size reported at earlier censuses and to death statistics. Because it is not clear whether the death statistics or the census counts are more accurate, the results at ages 65+ must be treated with caution.

Digit preference is also evident in the results. By age 45-49, a ratcheting pattern is established whereby age intervals ending with 5-9 have larger inflation factors than the two surrounding age intervals that include the digit, zero. The only exception occurs at age 65-69, where incentives to qualify for social security and medicare have probably affected the results. A large inflation of the 65-69 year old group was first evident in the 1940 census, after social security legislation was enacted (Elo and Preston, 1994; Coale and Rives, 1973).

The census-specific inflation factors shown in Table 2, and graphed in Figure 2, are always higher for males than for females. The gap between the sexes grows steadily from 2-3% in 1930 or 1940 to 7-8% in 1980 or 1990. Females show a clear trend towards improved census coverage, but no trend whatsoever is evident for males. The 1990 census appears to be less complete than the 1980 census, as representatives of the Census Bureau have recently concluded from demographic analysis (Robinson et al., 1993). However, no direct inference can be made from Table 2 about the completeness of any particular census, which also depends on the age structure of the population in combination with age-specific completeness factors.

The reconstructed population of African-Americans by age and sex from 1930 to 1990 is shown in Table 3. To reiterate our procedures, these estimates are not the original census count times age-specific and census-specific inflation factors. Instead, a matrix with these values ($C_{it} \cdot \alpha_a \cdot \tau_t$) is first created. Each value for a particular cohort is back-survived (by adding deaths and subtracting net immigrants) to the time of its first appearance in the matrix. The mean of all estimates of initial cohort size is then computed, with one to seven observations available for each cohort; this mean serves as the final estimate of cohort size at initial appearance. Finally, cohorts are survived forward in time from their initial size by subtracting intercensal deaths and adding intercensal net migration.[6]

We have demarcated in the central diagonal section of Table 3 those cohorts for which we have five or more observations on cohort size. These are the cohorts for which estimation is expected to be most reliable, both because of the larger number of observations available and because the values of α_a, estimated over the entire period, are most likely to be accurate in the middle of the period where observations for these cohorts are concentrated.

Table 4 converts the estimates in Table 3 into implied levels of census net undercount. The age-time pattern of undercounts is similar to that of the α_a's and τ_t's estimated earlier. The series of undercount estimates for all ages combined is not far from that estimated by the Census Bureau using demographic

[6] Note that the same estimates appearing in Table 3 would be produced regardless of the time at which the various estimates of cohort size are compared and synthesized into a mean. Referring to the cohort's size at its initial appearance in the age-time matrix is a heuristic convenience; we could also have referred to cohort size at its last appearance.

methods.[7] The Bureau has not attempted to make undercount estimates for the 1930 Census. Coale and Rives (1973) have used a complex procedure beginning with stable population assumptions and using assumed life tables to estimate the black population in 1930 and other years. As shown in Table 4, their estimated undercount in 1930 is much larger than our own, especially for females. Some of the discrepancy between the two sets of estimates is a result of the much larger estimates of the female population above age 40 in Coale and Rives. We have shown using extinct generation methods that their high estimated number of older females in 1930 is not confirmed by subsequent deaths recorded among these cohorts (Elo and Preston, 1994).

Tests of the quality of estimates

Three checks on the quality of these estimates are available. The first is a comparison of the estimated number of births in recent cohorts to the recorded number of births adjusted for underregistration. The second is a comparison of reconstructions for males with selective service registration during World War II. The third is a comparison of recent estimates of the 65+ population to adjusted enrollment in Medicare.

A. Birth registration

Although the estimated number of births in various cohorts is not presented in Table 3, it can be inferred by taking the estimated population aged 0-4 and 5-9 at various census dates and adding deaths (and subtracting net migration) between the time of birth and the date of the census. Results for males and females are shown in Table 5. For purposes of comparison, we use the corrected birth series prepared by Passel (1992a) under an agreement with the Census Bureau. This series corrects for a bias discovered in the 1940 Birth Registration Test, which affected earlier Census Bureau demographic analysis for cohorts

[7] The Census Bureau estimates also include ages 85+ but the population in this age interval is never more than 1% of the total African-American population.

born in 1935-40 and 1940-45. [8]

For males, the results of the birth registration test are encouraging. For the period 1945-1980, the number of births implied by the reconstruction differs from corrected registration figures by less than 1.5%. For the 1980-90 birth cohorts, the reconstructions are based on only one observation, from the 1990 census, and cannot be expected to have a high degree of reliability. The results for 1935-45 show fewer male births than adjusted birth registration figures. These are the birth cohorts affected by the 1940 Birth Registration Test, which has been the focus of a great deal of attention because of mounting evidence that earlier estimates drawn from the test were flawed (Passel, 1992a; Robinson, 1991a). The estimated number of births for this period have recently been adjusted downwards, and the present results suggest that a slightly larger downward adjustment would have produced greater consistency with male census counts for the affected cohorts.

Results for females are problematic. In the early years, the reconstructed birth series tracks corrected vital registration data reasonably well. For cohorts born after 1960, however, a discrepancy greater than 2% appears between the two series, and becomes successively larger. The reconstructions for cohorts born after 1970 are not credible. They are based upon two or fewer census observations; a larger number of observations is evidently required in order to achieve reliable results. What may have gone wrong in the female reconstructions for recent birth cohorts is suggested in Table 4. The estimation procedure assumes that age effects (α_a) are constant over the period, whereas there is evidently an improvement in the enumeration of females under 10 relative to that at other ages.

B. World War II Selective Service Registration

Price (1946) used selective service registration in 1940 to examine the completeness of the 1940

---

[8] For corrected registered births during 1980-90, we use unpublished worksheets supplied by the U.S. Bureau of the Census. These were prepared in conjunction with the Bureau's 1990 demographic analysis program.

U.S. Census for young adult males. The First Selective Service Registration occurred on October 16, 1940. Men aged 21 to 35 were required to register, with severe penalties for non-compliance of imprisonment for up to five years and/or a fine of up to $10,000. Price used registration figures back-dated to the April 1 census of 1940 for the age group 21-35. He concludes that the 1940 census omitted 13.0% of Negro males in this age range.

We are not able to examine the age range 21-35 in 1940 but can use our estimates at ages 20-34 to compare to census counts at those ages. Our age range represents a 93.3% overlap with that of Selective Service Registration. Our reconstructions have 1,800,382 males in this latter age group, compared to a census count of 1,547,743. The implied census undercount is 14.0%, quite close to Price's figure. The U.S. Census Bureau has also estimated the size of the 20-34 year old population in 1940. Their estimates are based in large measure on adjusted Medicare counts among the older population in 1980, back-survived to 1940. They estimate that there were 1,853,000 black males aged 20-34 in 1940 (Fay et al., 1988:106). This figure implies a net undercount for this group in the 1940 census of 16.5%. The Census Bureau's estimates are thus less consistent with selective service registration system figures than are the reconstructions presented here.

C. Comparison with adjusted Medicare data

Starting with the 1970 Census, the Census Bureau's undercount estimation efforts began to rely heavily on the Medicare system to estimate the size of the 65+ population. This system does not provide a definitive number of elderly persons, however, because coverage is incomplete, incentives exist to overstate one's age in order to qualify for benefits, age ascertainment is imperfect, and race is missing for some cases (about 5% in 1970; U.S. Bureau of the Census, 1973).

The Census Bureau has developed strategies to overcome these problems. In 1970, an assumption was made that white males were fully enrolled; expected sex ratios were applied to estimate the number of white females who should have been enrolled; an assumption was made that underenrollment was the same for black males as for white females; and expected sex ratios were applied to estimate the true number of black females (Siegel, 1974). Little justification is given for the assumptions, and expected sex ratios for blacks are unreliable because of incompleteness of death registration and censuses earlier in the century. In 1980 and 1990, a better procedure was used in which patterns of cohort-specific underenrollment by age are identified and, where necessary, assumed to apply to other cohorts; an arbitrary assumption about the proportion of each cohort who would never enroll is still required (Passel and Robinson, 1987; Robinson, 1991b).

Table 6 compares the Medicare-based estimates to our own estimates.[9] At ages 65-69, there is generally close agreement between the two, except for females in 1990, where our estimates exceed the Medicare-based estimates by 6.1%. Above age 70, the disparity between the series grows sharply, especially for males. At ages 75+, our reconstructed series is far below the Medicare-based estimates for both sexes in all years.

What accounts for the discrepancies above age 70 is probably the fact that, as noted above, African-American age reporting on death certificates is substantially "younger" than that on censuses; a matched record study of Medicare and 1970 census records for African-Americans showed that Medicare age reporting was even older than census age reporting (U.S. Bureau of the Census, 1973). If deaths are improperly inflated at younger ages, then too many deaths are being subtracted from a cohort's initial size as it ages, and our estimates of population at older ages are too low. Most African-Americans above 65 as late as 1990 do not have birth certificates, making age ascertainment an uncertain process. The fact that discrepancies between Medicare estimates and our reconstructions above age 70 are larger for males than for females probably reflects the fact that deaths are proportionally a more important source of change in cohort size for males than for females.

---

[9] Our estimates stop at age 84. To make the estimates comparable to the Medicare estimates, we have survived forward cohorts from the last time they appear in our series at ages 75-79 and 80-84 by subtracting cohort deaths. Extinct generation estimates are used for the cohort aged 100+ in 1990.

This discrepancy between age reporting in deaths and censuses affects all population reconstructions using standard demographic analysis as well. If the Medicare estimates are correct, then population estimates for the same cohorts earlier in their lives will be too large because too many deaths will be added back in because of net understatement of age at death by older cohorts. For example, the cohorts of males aged 20-34 in 1940, the subject of the previous section, have been reconstituted from the cohorts aged 60-74 in 1980. The Census Bureau's Medicare-based estimates for these cohorts exceed our own by 53,000 (combining the 65-74 year olds in 1980 with the cohort aged 60-64 in 1980 when it was aged 70-74 in 1990). This is exactly the discrepancy between the two series for 20-34 year olds in 1940, when it appears (judging from selective service figures) that the Census Bureau's estimates are too high. In other words, a correct figure for a cohort when it is older may be translating, via death statistics, into too high a figure when it is younger; or a correct figure for a cohort when it is younger may be translating into too low an estimate when it is older. The proportionate differences between the two alternatives become larger as the cohort ages, as this instance illustrates.

Because age reporting among older African-Americans is quite poor, any reconstructions for the older population are fraught with uncertainty. It appears that our reconstructed series performs well (judging from comparisons with Medicare-based estimates) up to age 65-69. Above that age, relative underreporting of age at death is likely to be producing too low an estimate of population size. Accordingly, the α_a estimates above 70 may be too low.

Summary

A new procedure to estimating census undercounts is developed and applied to African-Americans from 1930 to 1990. The method adds a minimization of squared error statistical criterion to the standard "demographic analysis" procedure used by the Census Bureau. Rather than assuming that the size of a cohort at birth is known from birth registration data, the size is inferred from census count for that cohort in each census where it appears. To make this inference, a model is proposed in which the true size of a cohort aged *a* at time *t* is the product of the census count, an age-specific inflation factor, and a census-specific inflation factor.

The procedure appears to work well for males, in the sense that it is consistent with Selective Service Registration data for World War II and with corrected birth registration data. Results for females are much less satisfactory. One reason may be that inconsistencies in age reporting between deaths and census counts are more serious for females. This conclusion was reached by the 1960 study matching death and census records (U.S. National Center for Health Statistics, 1968). Inconsistencies between ages reported on death certificates and in social security records are also greater for females than for males (Elo et al., 1995). A second reason why results are less satisfactory for females may be that the assumption of independence between age effects and census effects is less valid for females. Subsequent research will investigate the results by attempting to correct for age reporting inconsistencies and of adding age-period interactions in census errors to the basic model.

For cohorts born after 1950, when the bounds of uncertainty on birth registration completeness are low, there is no reason to expect results of this procedure to be as reliable as those derived from demographic analysis. However, the large majority of estimates derived in this paper for earlier cohorts; in fact a majority pertain to cohorts born before 1930, when the Birth Registration Area was incomplete or non-existent. For these cohorts, demographic-analysis estimates derived from birth registration are not available. The value of the present estimates should be greatest for cohorts born between 1885 and 1930, each of which is observed five or more times in subsequent censuses.

References

Coale, Ansley J. and Norfleet W. Rives. 1973. "A Statistical Reconstruction of the Black Population of the United States 1880- 1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility." Population Index. Vol 39(1):3-35.

Elo, Irma and Samuel Preston. 1994. "Estimating African American Mortality from Inaccurate Data." Demography. Vol 31(3):427-58.

Elo, Irma, Samuel Preston, Ira Rosenwaike, Mark Hill and Tim Cheney. 1995. "Consistency of Age Reporting on Death Certificates and Social Security Administration Records among Elderly African-Americans." Aging Research Center WPS 95-03. University of Pennsylvania. Philadelphia, PA.

Fay, Robert E., Jeffrey S. Passel and J. Gregory Robinson. 1988. The Coverage of Population in the 1980 Census. 1980 Census of Population and Housing Evaluation and Research Reports PHC80-E4. Bureau of the Census. Washington, D.C.: Government Printing Office.

Himes, Christine and Clifford C. Clogg. 1992. "An Overview of Demographic Analysis as a Method for Evaluating Census Coverage in the United States." Population Index. Vol 58(4):587-607.

Passel, Jeffrey S. 1992a. Alternative Estimates of Black Births Corrected for Underregistration: 1935-1980. Manuscript submitted to U.S. Bureau of the Census. Urban Institute. Washington, D.C.

Passel, Jeffrey S. 1992b. "Age-Period-Cohort Analysis of Census Undercount Rates for Race-Sex Groups, 1940-1980." The Urban Institute Discussion Paper Series PSC-DPS-UI-9. Washington, DC.

Passel, Jeffrey S., and J.G. Robinson. 1987. Methodology for Developing Estimates of Coverage in the 1980 Census Based on Demographic Analysis: Population Aged 65 and Over in 1980. 1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 74. U.S. Bureau of the Census. Washington, D.C.

Price, Daniel O. 1946. "A Check on Underenumeration in the 1940 Census." American Sociological Review. Vol :44-49.

Robinson J. Gregory. 1991a. Error in the Birth Registration Completeness Estimates. 1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 74. Manuscript of the U.S. Bureau of the Census. Washington, D.C.

Robinson, J. Gregory. 1991b. Robustness of the Estimates of the Population Aged 65 and Over. 1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 79. U.S. Census Bureau. Washington, D.C.

Robinson, J. Gregory, Bashir Ahmed, Prithwis Das Gupta, and Karen A. Woodrow. 1993. "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis." Journal of the American Statistical Association. Vol 88(423): 1061-1079.

Shryock, H.S. and J.S. Siegel. 1973. The Methods and Materials of Demography. Revised edition. Washington, D.C. Government Printing Office.

Siegel, Jacob S. 1974. Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis. 1970 Census of Population and Housing Evaluation and Research Program Report PHC(E)-4. Washington, D.C.: Government Printing Office.

Sorlie, P.D., E. Rogot, and N.J. Johnson. 1992. "Validity of Demographic Characteristics on the Death Certificate." Epidemiology. Vol 3(2):181-184.

U.S. Bureau of the Census. 1943. Studies in Completeness of Birth Registration. Part I. Completeness of Birth Registration. United States, December 1, 1939 to March 31, 1940. Vital Statistics - Special Reports. Vol 17, No. 18. Washington, D.C.: Government Printing Office.

U.S. Bureau of the Census. 1953. Infant Enumeration Study: 1950. Procedural Studies of the 1950 Censuses No. 1. Washington, D.C.: Government Printing Office.

U.S. Bureau of the Census. 1964. Accuracy of Data on Population Characteristics as Measured by CPS-Census Match. 1960 Census of Population and Housing Evaluation and Research Program. Series ER60, No.5. Washington, D.C.: Government Printing Office.

U.S. Bureau of the Census. 1973. The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census. 1970 Census of Population and Housing Evaluation and Research Program PHC(E)-7. Washington, D.C.: Government Printing Office.

U.S. Bureau of the Census. 1975. Accuracy of Data for Selected Population Characteristics as Measured by the 1970 CPS-Census Match. 1970 Census of Population and Housing Evaluation and Research Program. Washington, D.C.: Government Printing Office.

U.S. National Center for Health Statistics. 1968. Comparability of Age on the Death Certificate and Matching Census Record: United States-May-August 1960. Vital and Health Statistics. Series 2, No. 29. Washington, D.C.: Government Printing Office.

U.S. National Center for Health Statistics. 1969. Comparability of Marital Status, Race, Nativity, and Country of Origin on the Death Certificate and Matching Census Record: United States - May - August 1960. Vital and Health Statistics Series 2, No. 34. Washington, D.C.: Government Printing Office.

U.S. National Center for Health Statistics. 1990. Monthly Vital Statistics Report. Vol 39(1-3). Hyattsville, Md.: Government Printing Office.

Table 1. Sum of Squared Differences Among Estimators of Cohort Size for 29 Cohorts of African-Americans, 1930-1990

Iteration Number	Males	Females
0	13.31	5.40
1	2.96	2.97
2	2.37	2.91
3	2.29	2.89
4	2.28	2.87
5	2.27	2.86
10	2.23	2.78
15	2.19	2.71
20	2.16	2.64
25	2.13	2.58

The units are number of persons squared times 10^{10}. The sum of squared errors is computed for 115 observations on 25 cohorts.

Table 2. Estimated Multipliers of Census Counts by Age and Census for African-Americans

Age Interval	α_a		Census Year	τ_t	
(a)	Males	Females	(t)	Males	Females
0-4	1.014	1.029	1930	1.092	1.067
5-9	.982	1.002	1940	1.101	1.076
10-14	.950	.983	1950	1.082	1.060
15-19	.956	.982	1960	1.082	1.043
20-24	1.025	.996	1970	1.104	1.040
25-29	1.067	1.012	1980	1.087	1.014
30-34	1.065	1.004	1990	1.102	1.026
35-39	1.033	.983			
40-44	1.011	.986			
45-49	1.018	1.021			
50-54	.972	.997			
55-59	1.006	1.049			
60-64	.953	.995			
65-69	.883	.930			
70-74	.822	.907			
75-79	.862	.957			
80-84	.715	.874			

Table 3. Reconstructed Populations by Age and Sex, 1930-1990

Age	1930	1940	1950	1960	1970	1980	1990
				Males			
0-4	697280	693257	1034657	1478347	1379968	1352121	1573127
5-9	715024	663345	793326	1270458	1506040	1363526	1460802
10-14	656476	674371	679543	1022862	1478165	1396307	1376985
15-19	661036	698364	652210	781220	1264481	1521470	1391083
20-24	605739	628019	651541	655168	952983	1453382	1400050
25-29	581717	617369	663433	632800	767422	1255517	1528093
30-34	504510	554994	589041	633970	650688	1000740	1480883
35-39	463640	527600	576966	635626	612645	759048	1243068
40-44	374871	441259	508739	554773	600675	631477	967513
45-49	330907	394201	470627	530749	587052	577649	717800
50-54	247653	299589	371264	448595	490868	536561	576186
55-59	197439	252336	313670	391502	444344	496787	501085
60-64	132172	177481	220416	279052	346255	384340	431057
65-69	80652	131805	178512	216158	273351	315564	361978
70-74	45544	74444	108302	124457	151393	209726	240545
75-79	27537	37379	71190	97658	101269	140635	163675
80-84	12001	14924	32193	50995	40765	46537	78826
				Females			
0-4	730355	708632	1017834	1430527	1299327	1266954	1454374
5-9	740275	692032	798934	1238232	1434340	1271535	1357433
10-14	689296	710124	697344	1008625	1434125	1319355	1292462
15-19	700358	723553	685740	795536	1249129	1465575	1309513
20-24	632633	657458	697547	690649	1020377	1461091	1353670
25-29	605561	653120	698781	674099	806504	1270681	1494765
30-34	495809	581607	628661	680440	694401	1033232	1482174
35-39	441654	552850	618756	676199	668873	806626	1273898
40-44	355528	438664	540029	597961	661221	685062	1023055
45-49	313267	381937	499774	577188	641810	645770	786727
50-54	221418	292211	379321	487616	551603	620258	653864
55-59	171372	247713	318202	432708	516521	585760	598583
60-64	117000	163054	227374	305437	412466	482648	549279
65-69	75291	118859	185910	243430	336645	427553	[illegible]
70-74	46326	70924	106000	149463	202166	307665	

	371147						
75-79		30144	39618	71400	118241	150288	219090
	294872						
80-84		16836	20818	36516	56540	72887	103113
	174513						

Table 4. Estimated Census Net Omission Rate by Age and Sex, 1930-1990

Age	1930	1940	1950	1960	1970	1980	1990
				Males			
0-4	0.1223	0.1032	0.0879	0.0765	0.1155	0.0919	0.1047
5-9	0.0481	0.0295	0.0385	0.0577	0.0845	0.0794	0.0757
10-14	0.0494	0.0193	0.0005	0.0311	0.0472	0.0372	0.0454
15-19	0.0978	0.0977	0.0877	0.0499	0.0486	0.0213	0.0351
20-24	0.0849	0.1239	0.1286	0.1294	0.1172	0.1054	0.1010
25-29	0.1385	0.1421	0.1139	0.1326	0.1423	0.1363	0.1586
30-34	0.1727	0.1569	0.1213	0.1096	0.1262	0.1296	0.1555
35-39	0.0703	0.1233	0.0743	0.1031	0.1170	0.1274	0.1287
40-44	0.0936	0.0929	0.0771	0.0826	0.0940	0.1028	0.1053
45-49	0.0222	0.1165	0.1068	0.0948	0.1132	0.1080	0.1050
50-54	-0.1220	0.0549	0.0521	0.0912	0.0649	0.0597	0.0767
55-59	0.1157	0.1787	0.1556	0.0653	0.0883	0.0610	0.0881
60-64	-0.0100	0.1308	0.1169	0.0705	0.0331	-0.0020	0.0390
65-69	-0.0290	-0.1530	-0.0670	-0.0620	-0.0150	-0.0510	-0.0030
70-74	-0.1190	-0.1260	-0.0030	-0.2160	-0.2150	-0.1170	-0.0590
75-79	-0.0620	-0.0720	0.0841	0.0336	-0.0860	-0.0860	-0.0910
80-84	-0.2800	-0.2520	0.0401	0.2173	-0.4400	-0.6110	-0.2770
Ages 0-84	0.0773	0.0909	0.0797	0.0732	0.0803	0.0725	0.0898
All ages, U.S. Census Bureau	0.129[1]	0.109	0.097	0.088	0.091	0.075	0.085
				Females			
0-4	0.1515	0.1146	0.0757	0.0489	0.0657	0.0463	0.0529
5-9	0.0687	0.0596	0.0382	0.0338	0.0438	0.0284	0.0270
10-14	0.0874	0.0574	0.0285	0.0241	0.0205	-0.0070	0.0041
15-19	0.0639	0.0677	0.0780	0.0489	0.0211	-0.0210	-0.0050
20-24	-0.0280	0.0189	0.0428	0.0693	0.0443	0.0250	0.0246
25-29	0.0556	0.0573	0.0469	0.0634	0.0436	0.0266	0.0487
30-34	0.0961	0.0973	0.0542	0.0247	0.0129	0.0150	0.0344
35-39	-0.0440	0.0535	0.0164	0.0348	0.0195	0.0139	0.0158
40-44	0.0198	0.0543	0.0672	0.0319	0.0096	0.0008	0.0124
45-49	0.0192	0.0978	0.1143	0.0746	0.0600	0.0280	0.0297
50-54	-0.0270	0.0852	0.0695	0.0875	0.0365	-0.0070	0.0104
55-59	0.2112	0.2330	0.2099	0.0901	0.0915	0.0264	0.0380
60-64	0.0688	0.1312	0.1562	0.0490	0.0308	-0.0060	0.0035
65-69	0.0382	-0.2180	-0.1330	-0.0620	-0.0400	-0.0410	-0.0210
70-74	-0.0420	-0.1150	-0.0540	-0.1600	-0.1490	-0.0700	-0.0390
75-79	0.0205	-0.0560	0.0944	0.0783	0.0380	-0.0710	-0.0270
80-84	-0.0720	-0.0450	0.0665	0.1051	-0.1720	-0.2120	-0.1060
Ages 0-84	0.0568	0.0675	0.0595	0.0434	0.0313	0.0073	0.0192
All Ages, U.S. Census Bureau	0.121[1]	0.060	0.054	0.044	0.040	0.017	0.030

[1]Source: Coale and Rives (1973).

Table 5. Implied Numbers of Cohort Births Compared to Corrected Numbers of Registered Births, 1935-1990 (in 1,000s)

	Males			Females		
Period	Births Implied by Reconstructions (1)	Corrected Registered Births (2)	Ratio (1) ÷ (2)	Births Implied by Reconstructions (4)	Corrected Registered Births (5)	Ratio (4) ÷ (5)
1935-40	766.7	788.8	.971	767.0	768.4	.998
1940-45	867.6	879.1	.987	858.7	857.2	1.002
1945-50	1094.5	1094.6	1.000	1065.1	1069.7	.996
1950-55	1349.4	1347.7	1.001	1302.1	1323.9	.983
1955-60	1556.2	1548.6	1.005	1493.0	1513.8	.986
1960-65	1580.9	1563.5	1.011	1491.4	1530.4	.974
1965-70	1435.8	1417.1	1.013	1343.4	1385.2	.970
1970-75	1394.6	1378.9	1.011	1293.3	1341.0	.964
1975-80	1380.4	1399.6	.986	1288.1	1361.7	.946
1980-85	1477.0	1501.9[1]	.983	1368.6	1459.2[1]	.938
1985-90	1600.0	1653.3[1]	.968	1475.8	1603.9[1]	.920

[1]Source: Unpublished tabulations of the U.S. C ensus Bureau. Otherwise, corrected registered births are drawn from Passel (1992).

Table 6. Comparisons of Reconstructions to Medicare-Based Estimates, 1970-1990 (in 1,000's)

	Males		Females	
Age Group	Reconstructions	Medicare Adjusted	Reconstructions	Medicare Adjusted
	1970			
65-69	273	272	337	334
70-74	151	187	202	244
75+	172	205	276	316
	1980			
65-69	316	323	428	418
70-74	210	231	308	322
75+	231	283	431	498
	1990			
65-69	362	368	490	462
70-74	241	263	371	370
75-79	274	354	621	678

Source of adjusted Medicare estimates: Fay et al. (1988); Robinson (1991b:Table 8).

Figure 1: Estimated Multipliers of Census Counts by Age

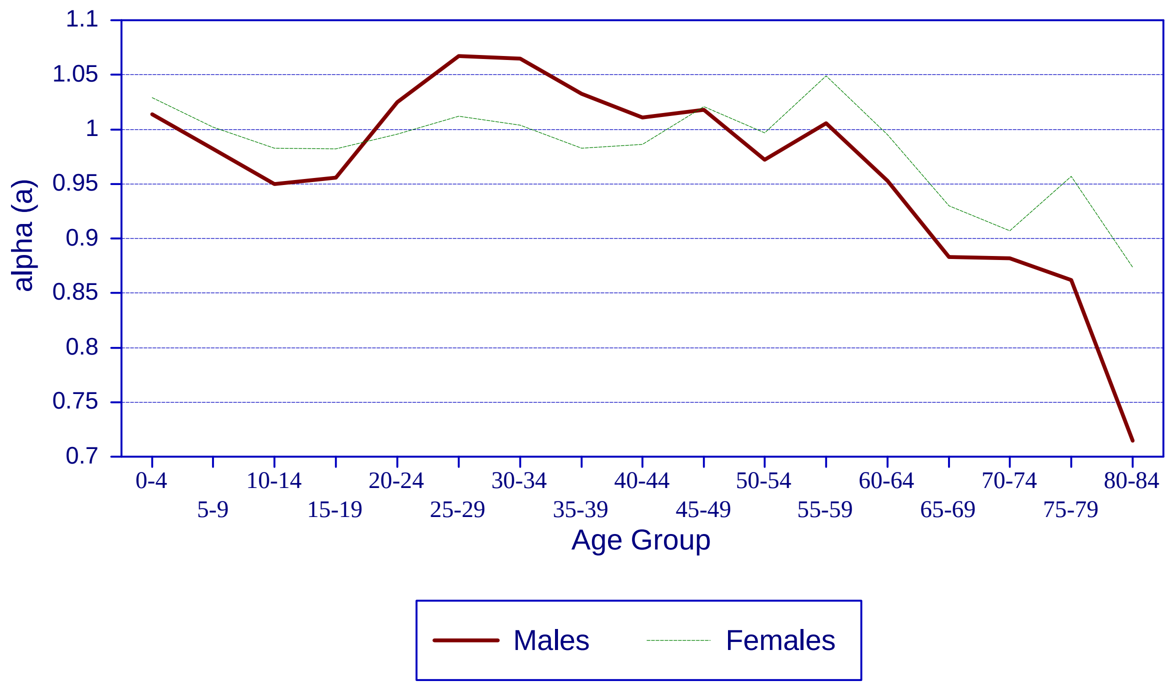

Figure 2: Estimated Multipliers of
of Census Counts by Census