Understanding the Modified Race File

The U.S. Census Bureau produces a modified set of age and race data[1] after each 10-year census, primarily because the race data collected in the census are not consistent with similar data collected by other federal agencies. Further, the revised data file fixes notable problems with the census age data, caused primarily by people other than household members guessing the ages of their neighbors or tenants (known as proxy responses).

Race

The Census Bureau produces a modified race file because the decennial census (which includes the American Community Survey, the ongoing part of the census) is the only federal survey to include a “Some Other Race” (SOR) category in the race question.[2] The official federal policy for collecting race data — Office of Management and Budget (OMB) Statistical Policy Directive 15 — does not permit such a category. Therefore, in order for the census to produce results that can be compared to other federal datasets, the Census Bureau re-codes individuals who select SOR into one or more of the OMB categories.[3] Stated another way, the goal of "modifying" census race results is to produce a dataset that is consistent not only with all other federal datasets, but also with the Census Bureau’s own annual population estimates and projections, which rely substantially on administrative records that do not offer a SOR category.

The process the Census Bureau uses to recode the SOR population has changed modestly each decade, but the general process remains the same:

1. If a respondent selects two or more races, and one of those choices is SOR, the bureau drops SOR and assigns the respondent only to the other race selections. For instance, if a person identified as Black and SOR in the census, they would be recoded as Black only in the modified race file. If they identified as White, Black, and SOR, they would be recoded as White and Black only.

2. If a respondent selects SOR only in the census race question, the Census Bureau needs to determine, using a statistical probability modeling technique, what other race would be an appropriate assignment. The bureau altered the methodology for the 2020 Census due to the increase in respondents who selected SOR. The Census Bureau anticipates that with the new Statistical Policy Directive 15 standards, the number of SOR responses will be reduced substantially in the 2030 Census.

Age

Changes are made to the census age data to reduce *age heaping* in the census results. Age heaping happens when there are more ages reported with digits ending in 0 or 5 as a result of proxy reporters estimating an individual’s age than experts expect to see. In the census, census takers may have to ask neighbors or landlords for information about a household’s residents if no one has responded as the count winds down.

Using a combination of statistical methods, a more realistic (that is, statistically probable) and accurate age file is produced that reflects the natural distribution of birth years and ages. This is called *age smoothing* and was used in the 4.5 percent of households that reported age-only on the 2020 Census form. Without age smoothing, the 2020 Census data show a high degree of improbable age heaping, likely because the coronavirus pandemic led to more proxy responses than usual. The final 2020 Modified Age and Race File has the same level of accuracy as the household population that reported both age and date of birth.

[1] 2020 Modified Race Data
[2] Since the passage of Public Law 108-109, the Census Bureau has been statutorily required to include a Some Other Race category on the decennial census and the ACS.
[3] In 2024, the OMB updated Statistical Policy Directive 15, which changed the minimum number of racial and ethnic categories from five to seven.

Updated July 2025 by The Leadership Conference on Civil and Human Rights. For any inquiries, contact Meeta Anand (anand@civilrights.org).