2020 Modified Age and Race Census (MARC) File

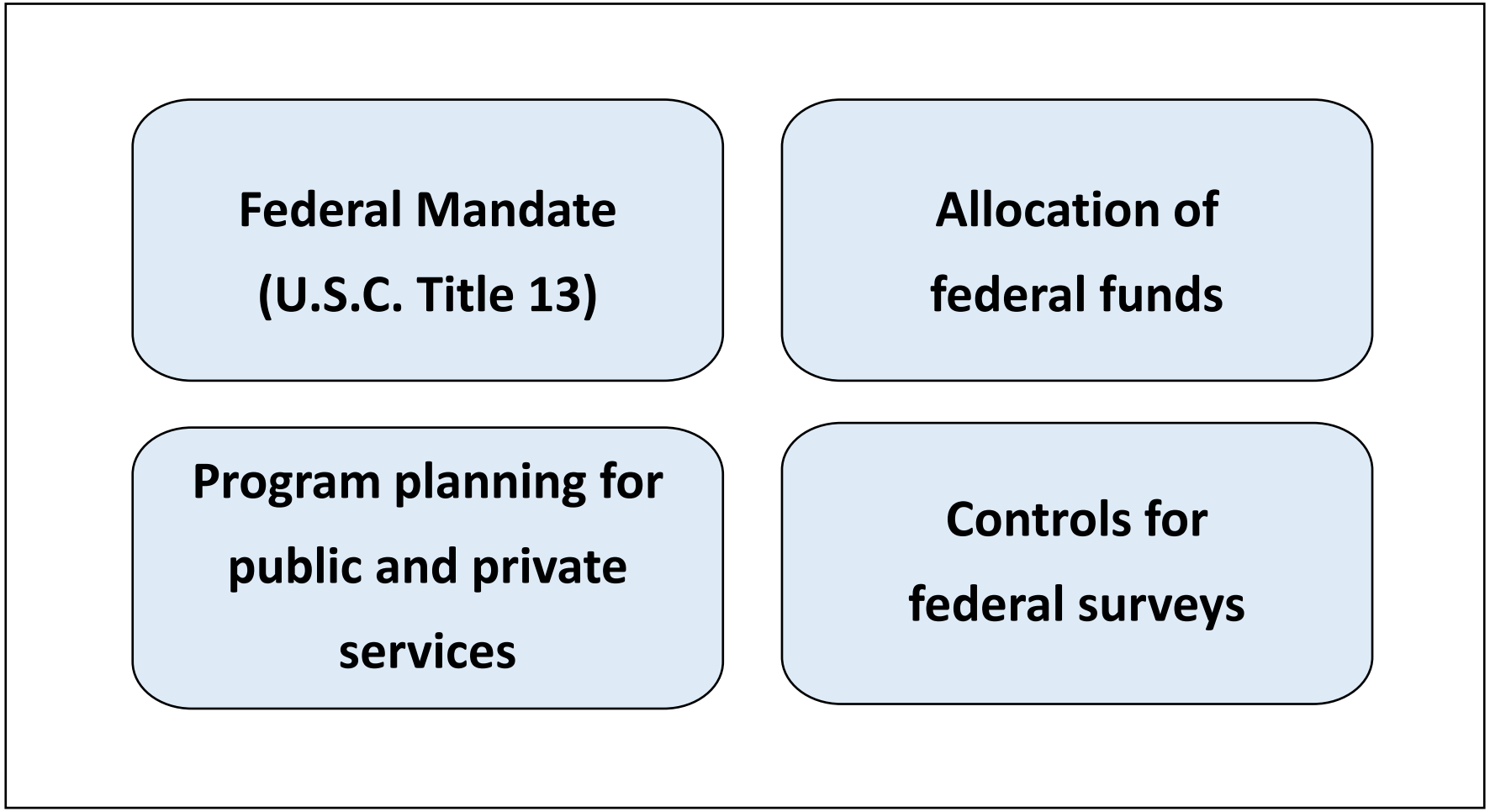
Eric Jensen
Senior Research Scientist for Population Estimates and Coverage Measurement
Population Division
U.S. Census Bureau

California State Data Center Spring Meeting
May 22, 2025

The Population Estimates Program

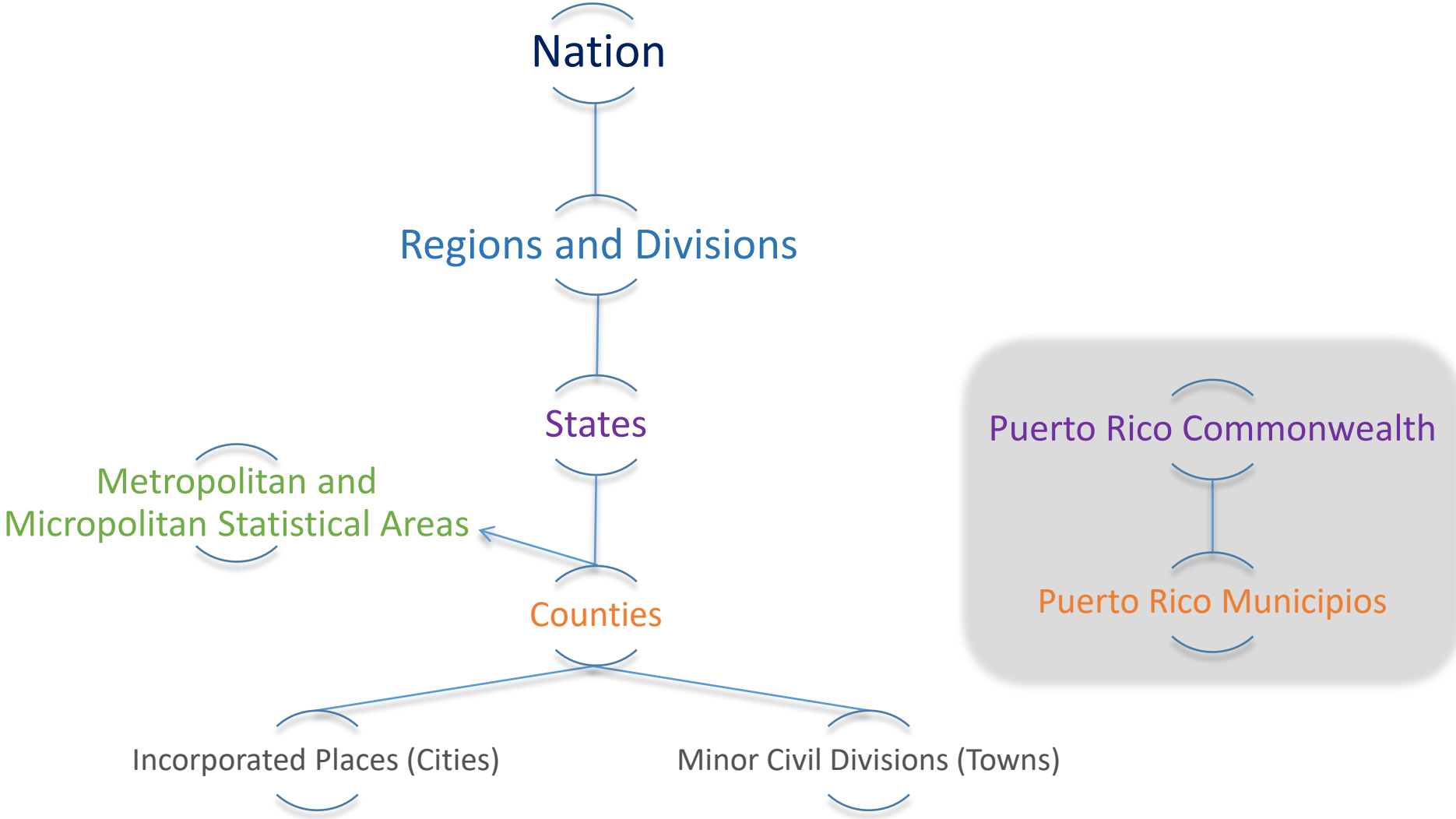
- The Population Estimates Program (PEP) disseminates official measures of population and housing units between decennial censuses
- Developed as per Title 13, U.S. Code, Section 181
- Use cases include:
  - Population controls and denominators
  - Academic and business research
  - Program planning in the public and private sectors
- Estimates extend back to 1900
- Time series are released annually featuring data for the date of the last census through the vintage year, which represents the latest year of estimates available
- Vintage 2024 is the current series
  - Contains estimates for April 1, 2020 to July 1, 2024

Why We Disseminate Estimates



United States® Census Bureau

PEP Estimates Geographies



80,000+ Geographic Areas / 40,000+ Communities

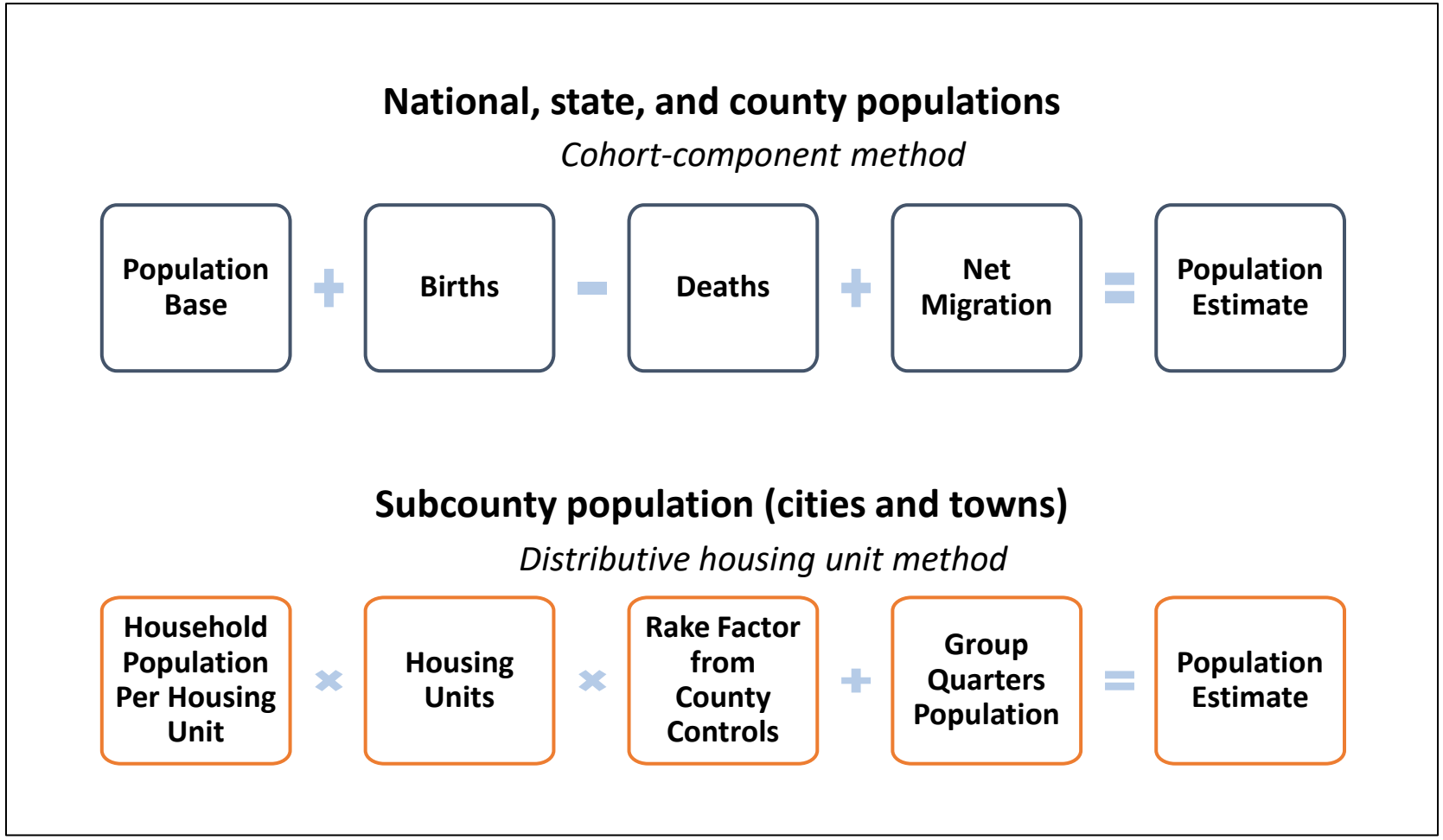
Population and Housing Unit Estimates by Type and Geography

Estimate Type	U.S.	States	Metro/ Micro Statistical Areas	Counties	Cities and Towns	Puerto Rico	Puerto Rico Municipios
Total Resident Population	X	X	X	X	X	X	X
Components of Change (Births, Deaths, Net Migration)	X	X	X	X		X	
Population Characteristics (Age, Sex, Race, Hispanic Origin)	X	X	X	X		X (age & sex)	X (age & sex)
Monthly Population (Five Universes*)	X						
Group Quarters Population	X (by char.)	X (totals)		X (totals)			
Housing Units	X	X		X			

*These five estimates universes are: resident, resident plus Armed Forces overseas, civilian, civilian noninstitutionalized, household.

United States® Census Bureau

Population Estimates Methodology



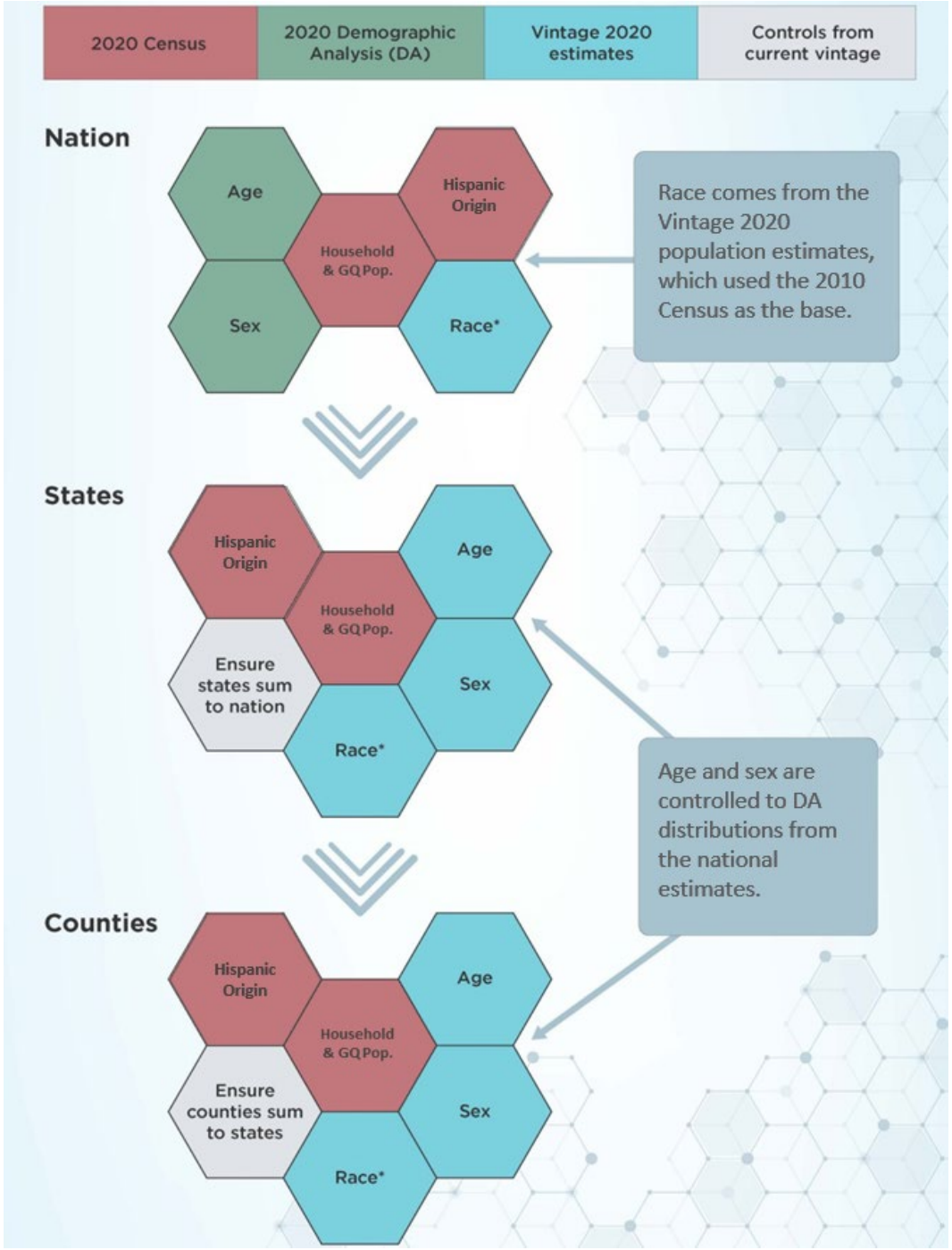
United States® Census Bureau

Population Estimates Input Data

Component	Source of Data
Base population	Blended Base: 2020 Census, Vintage 2020, 2020 Demographic Analysis
Births	National Center for Health Statistics
Deaths	National Center for Health Statistics
Domestic migration	IRS Tax Records and Medicare Enrollment Records
International migration	American Community Survey, Puerto Rico Community Survey, Foreign Censuses and Population Registers, Internal Benchmark Database

United States® Census Bureau

Blended Base Approach



United States® Census Bureau

Terminology (not confusing at all…)

- **Modified Race:** The process of reconciling the race categories from the Census with those that appear in the population estimates. This was the name of the file we released in 2000 and 2010
- **IMPRACE:** The name of the race variable with the five 1997 Office of Management and Budget (OMB) categories and all possible combinations (31 categories)
  - White, Black, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander
- **RACE2010:** The name of the IMPRACE variable on the 2020 Census Edited File (CEF)
- **MARC File:** The Modified Age and Race Census file, which contains the modified race data for the 2020 Census

United States® Census Bureau

What is Modified Race?

- The 2020 Census included an option for respondents to identify as Some Other Race (SOR)
- The vital records and other input data for the population estimates do not have SOR
- We need to reconcile the race categories from the Census with those that appear in the population estimates
- We create the modified race data that allow for comparisons between the 2020 Census results and the population estimates

United States® Census Bureau

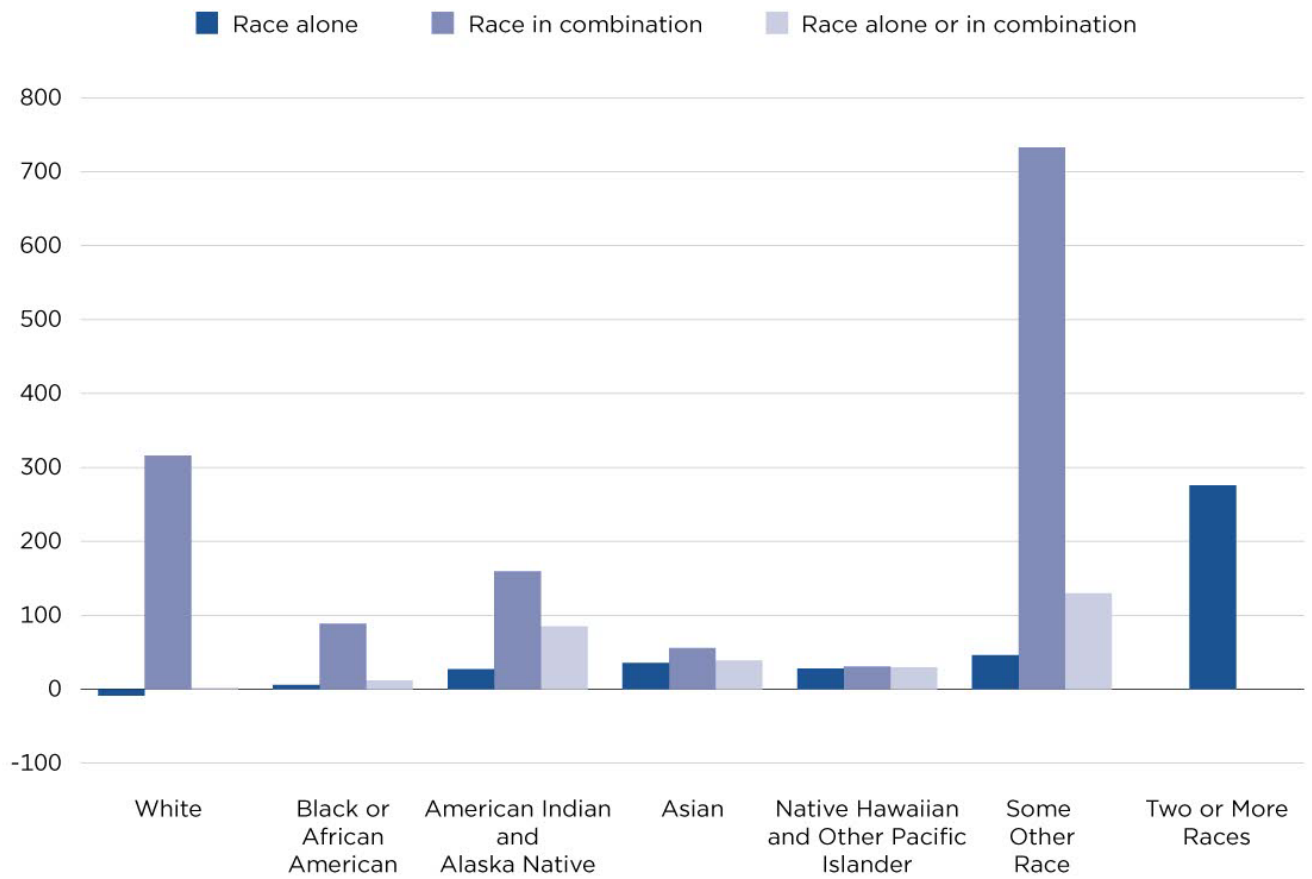
SOR in 2020

- In the 2020 Census, there were 27,915,715 people who identified as SOR alone. This was 8.4% of the total population.
- There were 49,902,536 people who identified as SOR alone or in combination, or 15.1% of the total population (**up from 21 million in 2010**)
- The SOR population overwhelmingly identifies as Hispanic.
  - 94% of responses to the race question that are classified as Some Other Race alone are from people of Hispanic or Latino origin
  - 42% of the Hispanic population identifies as SOR alone, and 73% of Hispanics identify as SOR alone or in combination
- To produce the modified race data, we needed to reassign these observations to the 1997 OMB race categories.

United States® Census Bureau

Figure 3.

Percentage Change in Race Groups: 2010 and 2020



Note: Data users should use caution when comparing 2010 Census and 2020 Census race data because of improvements to the question design, data processing, and coding procedures for the 2020 Census.

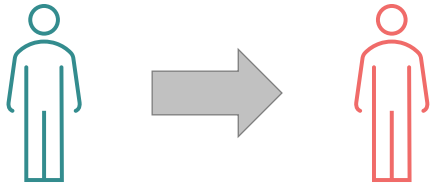
Information on confidentiality protection, nonsampling error, and definitions is available at <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/>.

Source: U.S. Census Bureau, 2010 Census Redistricting Data (Public Law 94-171) Summary File; 2020 Census Redistricting Data (Public Law 94-171) Summary File.

United States® Census Bureau

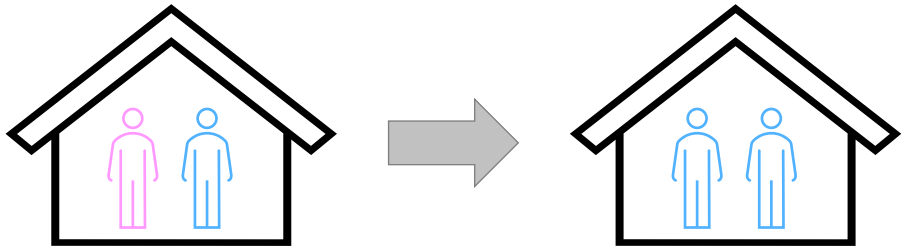
Creating RACE2010

1. If SOR and another race category, then drop SOR



Asian *and* SOR → Asian

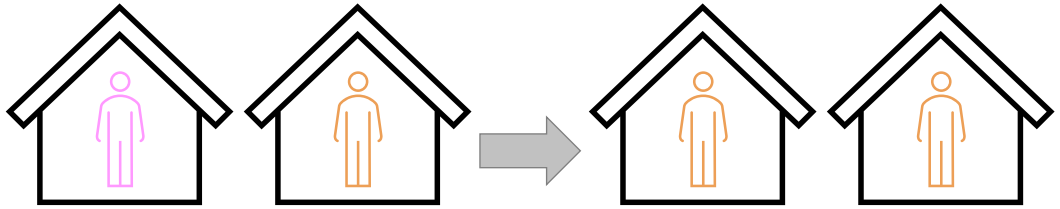
2. If SOR alone, then use information from a household member



SOR alone → Black or African American alone

Black or African American alone → Black or African American alone

3. If SOR alone and no information from a household member is available, then use information from a neighbor in the same age group and Hispanic or Latino origin group



SOR alone, White alone → White alone, White alone

United States® Census Bureau

MARC File

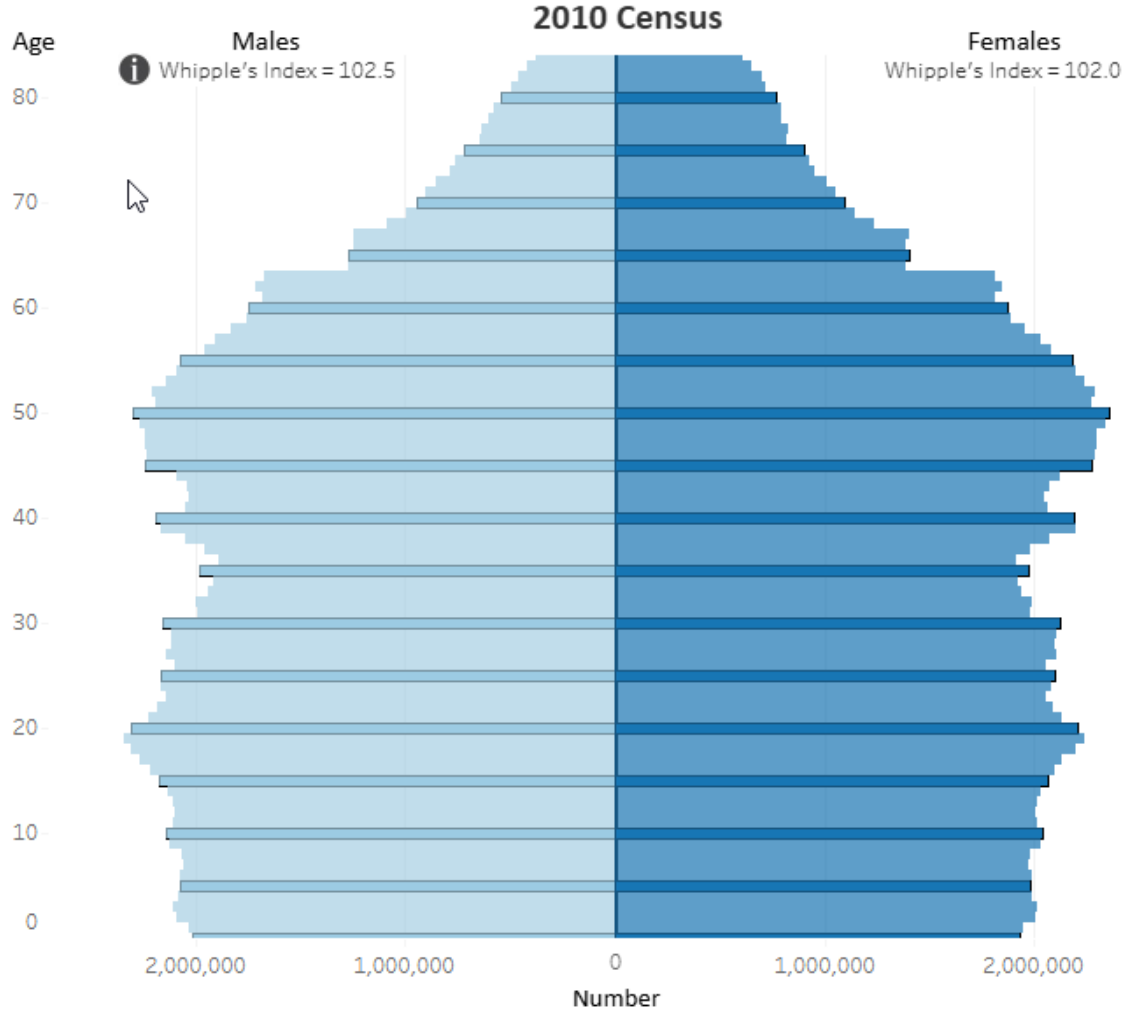
- The Census Bureau produced the 2020 Modified Age & Race Census (MARC) File
- The primary use for the MARC file is to have race data in the imputed race, or “IMPRACE,” categories
- A secondary benefit of the MARC file is smoothed age distributions

United States® Census Bureau

What is Age Heaping?

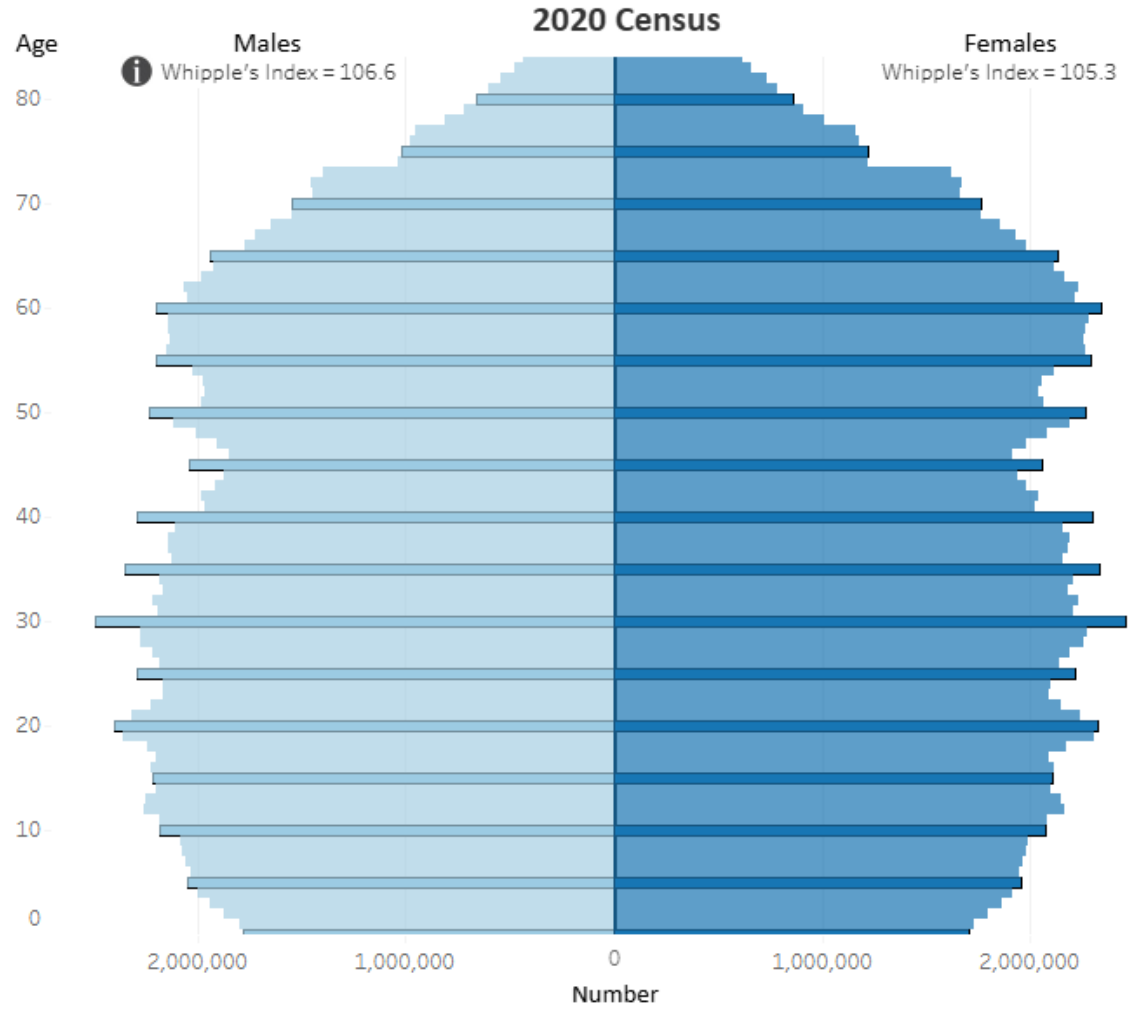
- Age heaping is a phenomenon where there are higher values reported for ages ending in 0 or 5 than would be expected
- While all censuses show some age heaping, it was much more pronounced in the 2020 Census than in past decades
- The “Whipple’s index” is one way to measure age heaping
  - A value of 100 indicates no heaping
- The population that only reported an age and not a date of birth (“age only”) showed the most heaping

United States® Census Bureau



United States® Census Bureau | U.S. Department of Commerce U.S. CENSUS BUREAU census.gov

Source: 1980 Census Summary File 2C (SF2C), 1990 Census Summary File 2C (SF2C), Census 2000 Summary File 1 (SF1), 2010 Census Summary File 1 (SF1), 2020 Census Demographic and Housing Characteristics File (DHC)



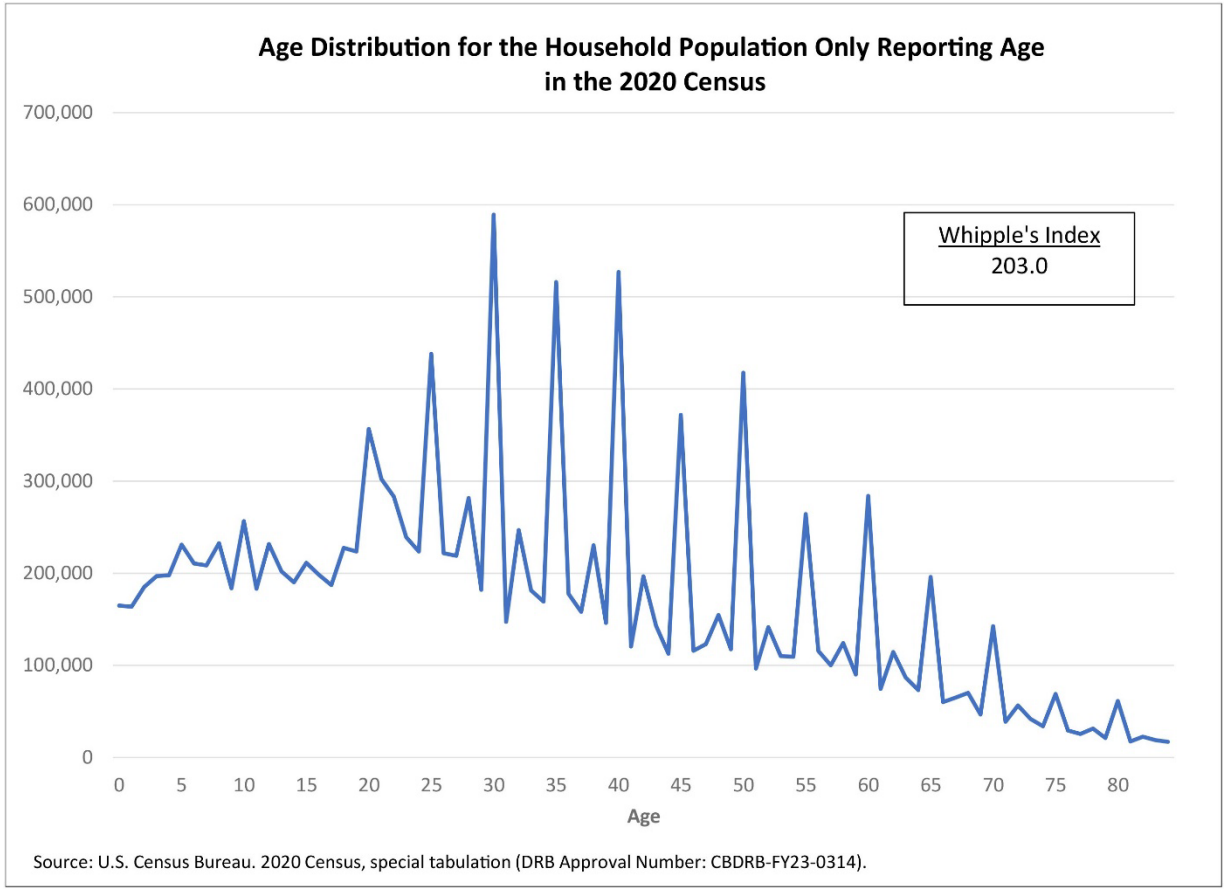
United States® Census Bureau | U.S. Department of Commerce U.S. CENSUS BUREAU census.gov

Source: 1980 Census Summary File 2C (SF2C), 1990 Census Summary File 2C (SF2C), Census 2000 Summary File 1 (SF1), 2010 Census Summary File 1 (SF1), 2020 Census Demographic and Housing Characteristics File (DHC)

United States® Census Bureau

Addressing Age Heaping

- The age heaping was greatest for the “age only” population
  - Approximately 14.7 million people or 4.5% of the household population
- We pulled these records out of the data and smoothed the age using a Discrete Gaussian Kernel approach
- Next, we put them back into the total population before we passed the data to the Disclosure Avoidance System (DAS) Science Team



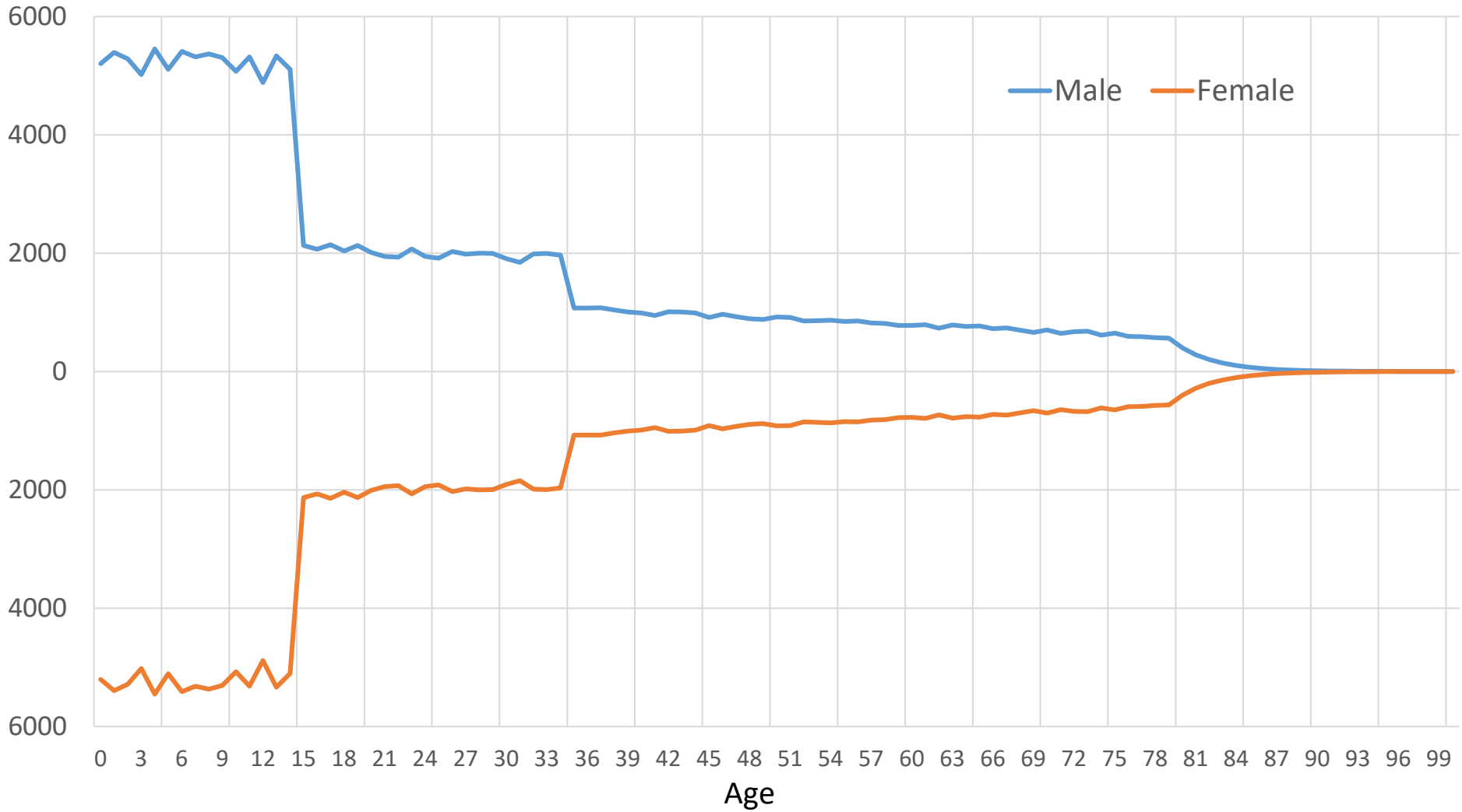
United States® Census Bureau

Issue with the RACE2010 data

- We had planned to release the file on September 5, 2024
- In reviewing the 2010-2020 Intercensals file, an issue was discovered that impacts all race groups but is most noticeable in the Hispanic data by race
- The review found demographically implausible age distributions for many of the race groups by Hispanic origin
- We traced the issue back to the edit process used to create the imputed race data in the 2020 Census Edited File (CEF)

United States® Census Bureau

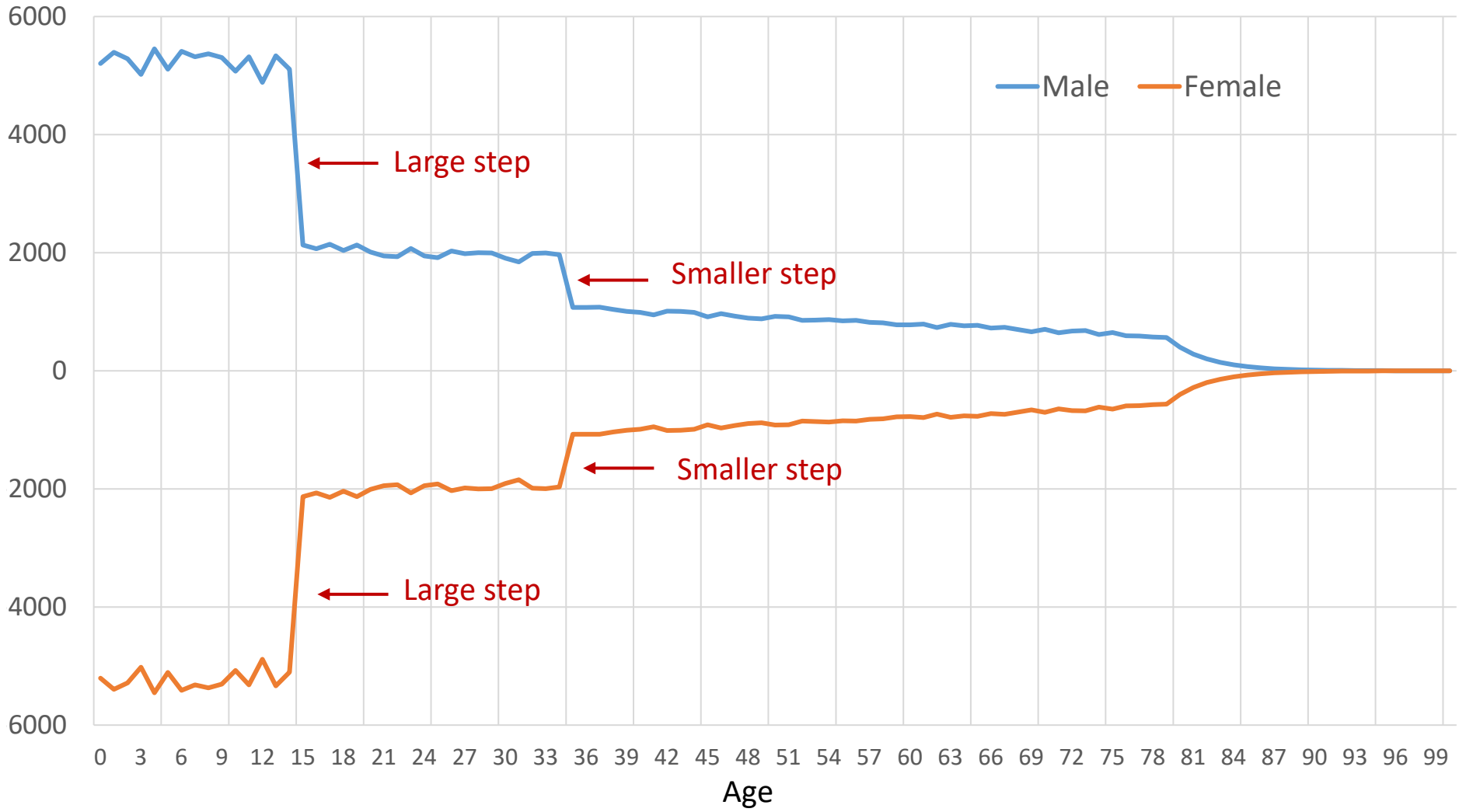
Figure 1. Illustration of the Data Issue Impacting the 2020 Modified Age & Race Census (MARC) File



Source: Completely made-up data for illustration purposes only.

United States® Census Bureau

Figure 1. Illustration of the Data Issue Impacting the 2020 Modified Age & Race Census (MARC) File

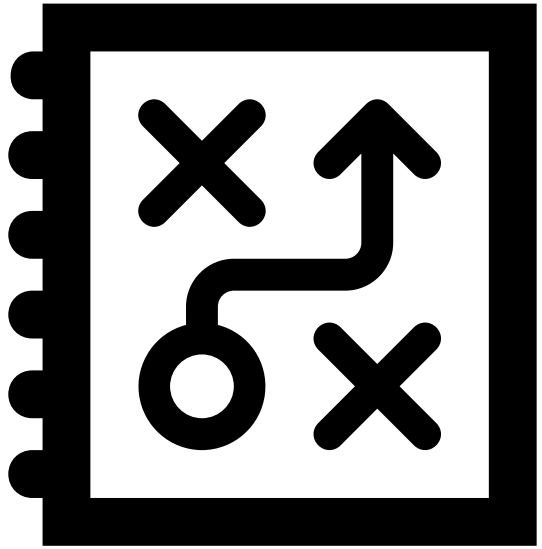


Source: Completely made-up data for illustration purposes only.

United States® Census Bureau

Revising the RACE2010 data

- After we found the issue, the Base Evaluation and Research Team (BERT) started working to revise the RACE2010 data
- BERT looked at several options:
  - Smoothing the impacted ages
  - Revising the hotdeck imputation method
  - Developing a county proxy distribution method
  - Using statistical models to impute a new race value



United States® Census Bureau

Mlogit model

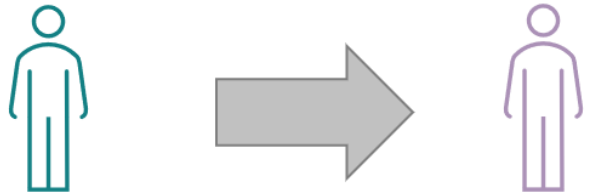
- Starts by adding a new edit step that assigns the RACE2010 value based on past census or survey response and other administrative records
  - This is similar to the administrative record edit that was done for the race data in the 2020 Census Demographic and Housing Characteristics File (DHC)
- Multinomial logistic regression model
  - The variables used to fit this model were:
    - Edited age
    - Edited Hispanic origin
    - Group quarters type
  - The dependent variable was the 31 modified race outcomes accounting for all combinations of the 5 major race groups: White, Black, AIAN, Asian, and NHPI
- The approach only impacts the RACE2010 values for the SOR alone population that were hotdecked in the CEF

Mlogit model

- Mlogit results:
  - The Mlogit model fixed the age issue
  - The race and age distributions were very similar to RACE2010 results
  - The model ran faster than other models
- Mlogit limitations:
  - Originally processed state by state, but the model didn’t converge for some of the smallest states
  - We grouped the states that didn’t converge by region and reran the model

Revising RACE2010

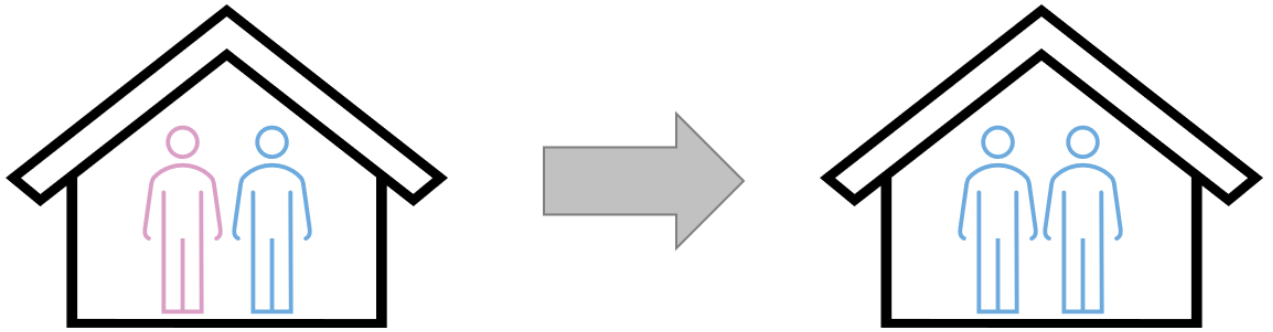
1. If SOR and another race, then use the non-SOR race category



American Indian and Alaska Native *and* SOR

American Indian and Alaska Native

2. If SOR alone, then use information from a household member



SOR alone

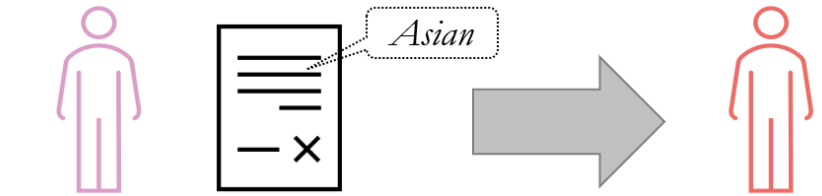
Black or African American alone

Black or African American alone

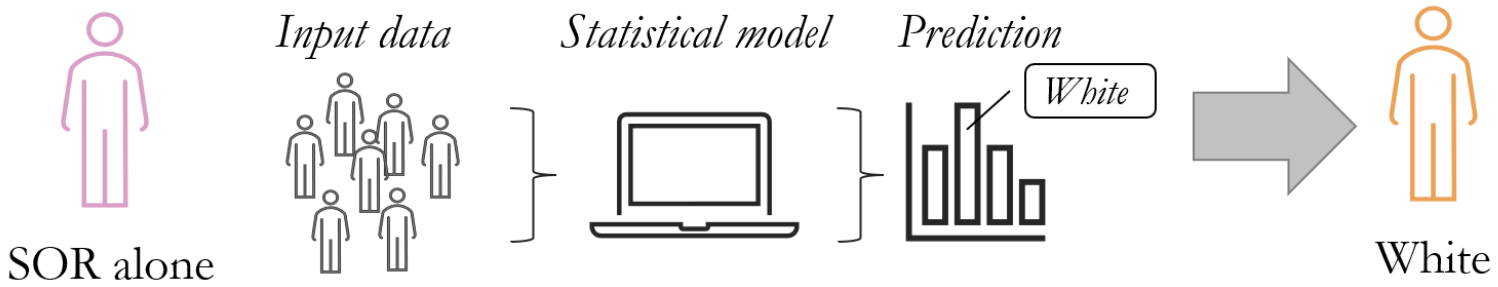
Black or African American alone

United States® Census Bureau

3. If SOR alone *and* no information from a household member is available, then use information from administrative data for that individual



4. If SOR alone *and* no information is available from administrative records, then apply a statistical model that uses age, Hispanic origin, and group quarters type to make a prediction based on the population within the state with similar characteristics



United States® Census Bureau

Differences between the MARC and DHC

MARC

- Imputed race categories (31)
- Age data have been smoothed to reduce heaping
- Data available for the nation, states, and counties
- Disclosure Avoidance System (DAS) parameter settings optimized for the population estimates use case

DHC

- Census race categories (63)
- Age data show heaping for values ending in 0 and 5
- Data available for the nation, state, counties, MSAs, congressional districts, tracts, block groups, and blocks
- DAS parameter settings optimized for a wide range of use cases

United States® Census Bureau

2030 Census Modified Race process

- Producing the 2020 Census Modified Race variable was challenging because the number of SORs increased substantially.
- However, with a combined race and ethnicity question (SPD-15) the SOR population should be much smaller.

Table 4. Race/Ethnicity Group Distribution by Question Format for Internet

Alone or in Combination Groups	Separate Question	Combined Question with Write-In Response Areas	Combined Question with Detailed Checkboxes
White	78.6% (0.46)	75.2% (0.51)	75.0% (0.54)
Hispanic	11.3% (0.27)	12.5% (0.26)	12.3% (0.27)
Black	8.3% (0.36)	8.1% (0.36)	8.3% (0.36)
Asian	7.7% (0.17)	7.7% (0.19)	7.8% (0.16)
AIAN	4.2% (0.08)	4.6% (0.08)	3.6% (0.07)
MENA	0.9% (0.04)	1.0% (0.04)	1.1% (0.04)
NHPI	0.4% (0.02)	0.4% (0.02)	0.3% (0.02)
SOR	10.2% (0.25)	1.5% (0.05)	1.0% (0.04)
Invalid	0.6% (0.03)	0.3% (0.02)	0.3% (0.02)
Missing	1.0% (0.04)	0.8% (0.04)	0.7% (0.04)

Source: 2015 National Content Test data. Note: Estimates are weighted with standard errors in parentheses.

United States® Census Bureau

2030 Census Modified Race process

- Producing the 2020 Census Modified Race variable was challenging because the number of SORs increased substantially
- However, with a combined race and ethnicity question (SPD-15) the SOR population should be much smaller

Table 4. Race/Ethnicity Group Distribution by Question Format for Internet

Alone or in Combination Groups	Separate Question	Combined Question with Write-In Response Areas	Combined Question with Detailed Checkboxes
White	78.6% (0.46)	75.2% (0.51)	75.0% (0.54)
Hispanic	11.3% (0.27)	12.5% (0.26)	12.3% (0.27)
Black	8.3% (0.36)	8.1% (0.36)	8.3% (0.36)
Asian	7.7% (0.17)	7.7% (0.19)	7.8% (0.16)
AIAN	4.2% (0.08)	4.6% (0.08)	3.6% (0.07)
MENA	0.9% (0.04)	1.0% (0.04)	1.1% (0.04)
NHPI	0.4% (0.02)	0.4% (0.02)	0.3% (0.02)
SOR	10.2% (0.25)	1.5% (0.05)	1.0% (0.04)
Invalid	0.6% (0.03)	0.3% (0.02)	0.3% (0.02)
Missing	1.0% (0.04)	0.8% (0.04)	0.7% (0.04)

Source: 2015 National Content Test data. Note: Estimates are weighted with standard errors in parentheses.

United States® Census Bureau

Questions?

Eric Jensen
eric.b.jensen@census.gov

United States® Census Bureau