Methods Used by the Census Bureau to Measure the Accuracy of the 2020 Census Count and Estimate the Population Beyond

Joseph Salvo
Fellow
Social and Decision Analytics Division
University of Virginia Biocomplexity Institute

Abstract

Understanding the process and methods used to evaluate the 2020 Census count is integral to well-informed applications of the data. This applies not only to the census itself but to sample surveys – inside and outside of the Census Bureau – where the decennial census defines the universe for estimates. Further, understanding the methods and issues involved in the preparation of estimates in the post-census years is equally important, since these estimates are used for a wide variety of purposes, most importantly the distribution of federal funds to states and localities. The goal is to provide participants with a high-level view of how the Census Bureau evaluated the 2020 Census and constructs post-census population estimates. Emphasis is placed on the interpretation of general concepts and on the language used by the Census Bureau to describe their methods.

Methods of Coverage Evaluation

1. Demographic Analysis (DA)[1]

Overview: The DA program uses current and historical vital records, data on international migration, and Medicare records to produce national

[1] Demographic Analysis (capitalized) is a term coined by the U.S. Census Bureau to denote a method used to produce national estimates of the population independent of the decennial census for evaluating the completeness of census coverage of the population. It can be distinguished from the generic use of "demographic analysis," which refers to general methods used by or analyses conducted by demographers.

estimates of the population on April 1, 2020, by age, sex, the DA race categories, and Hispanic origin. The results will be compared to the census counts to evaluate net coverage error. It is essentially providing a measure of *net coverage* by comparing the population from the census enumeration with an independent count of the population. This method cannot produce estimates of gross errors – erroneous enumerations/duplicates (overcounts) and omissions (undercounts).

a. **Frequency:** Conducted every 10 years just prior to the release of the census count for apportionment of the seats in the House of Representatives
b. **Geographic Detail:** National only, since estimates of interstate migration are not sufficiently accurate for the preparation of state estimates.
c. **Subgroup Detail:** Single years of age, sex, limited race (black/non-black) and Hispanic origin for ages 0-29.
d. **General Approach:** Four components:
    i. Births from vital records
    ii. Deaths from vital records
    iii. International Migration (subpopulation counts):
        1. Stock of foreign-born persons now living in the U.S. (largest group by far, includes people in the U.S. legally and without documentation)[2]
        2. Native-born who emigrated (remove)
        3. Migration from Puerto Rico to the U.S. Mainland (add)
        4. Born abroad of American citizen parents and now living in the U.S. (add)
    iv. Use Medicare enrollment for older people (65 years and older in 2010, 75 years and over in 2020, will be 85 years and over in 2030)

[2] The American Community Survey (ACS) which is used to estimate the size of the foreign-born population, collects data for all immigrants, irrespective of legal status.

e. **Published Information:** three series: Low, middle, and high are based on different counts of International Migration and other components in what is referred to as a "sensitivity analysis."

f. **2020 Components of Demographic Analysis (DA)**

Population Components from Demographic Analysis		
United States		
April 1, 2020 (Middle Series)		
	Number	
Total Population (000s)	332,601	
Births	288,908	
Deaths	(22,412)	
International Migrants	44,256	
Medicare Enrollees	21,849	
Source: U.S. Census Bureau. Population Division,		
2020 Census Demographic Analysis		

The population estimate for the U.S. for April 1, 2020, was 332.6 million, based on the count from the DA middle series: 288.9 million births minus 22.4 million deaths, plus 44.3 million international migrants and 21.8 million Medicare enrollees. This compared with an enumerated population of 331.5 million from the 2020 Census, which translates into an undercount of some 1.1 million or 0.35 percent. This compares with a 2010 Census overcount of 0.13 percent, using the middle series.

Depending on the series, **DA results** show both an undercount and overcount nationally. DA provides a range of net coverage errors — low, middle and high. This range is produced by varying the level of historical births, international migration, and Medicare enrollment records across the three series. The DA's net coverage error estimate was 0.22% (a slight overcount) for the low series, which includes less international migration, fewer people in the oldest ages, and fewer historical births than the other series. The net coverage error estimate for the middle series is -0.35% (a slight undercount). The

high series, which has the highest international migration, more population in the oldest ages, and more historical births, shows a net coverage error of -1.21%. In contrast, the 2010 DA showed a 1.00% overcount in the low series, a 0.13% overcount in the middle series, and a 1.27% undercount in the high series.

2020 and 2010 Census Coverage Results: DA

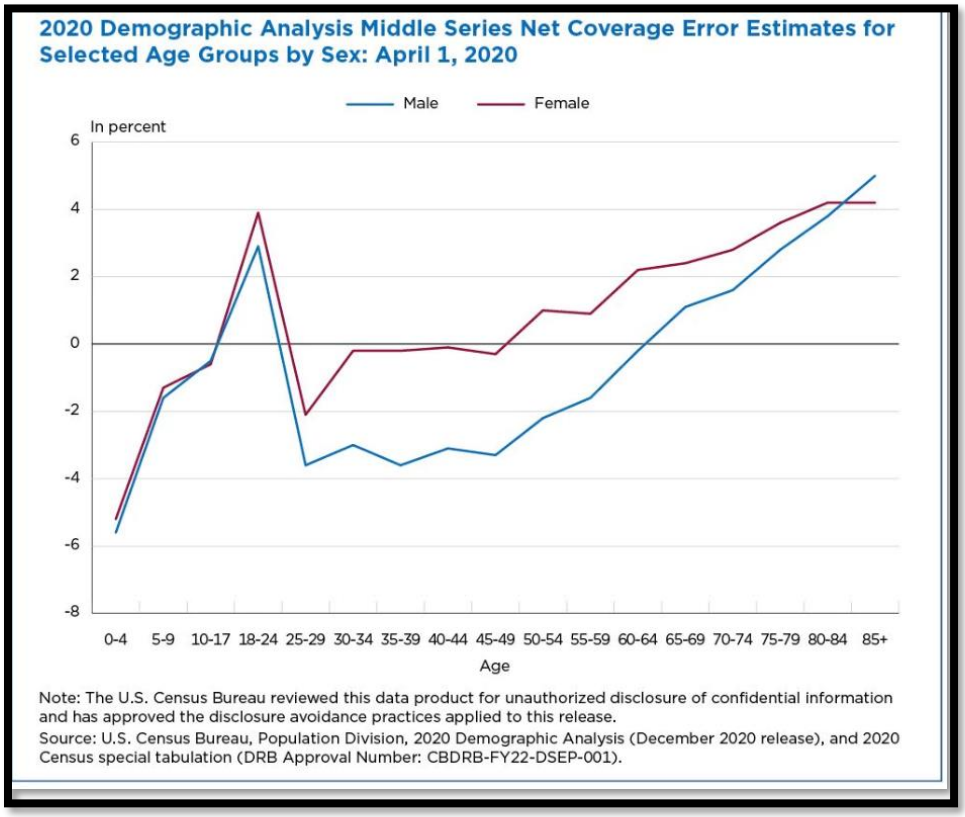
Year	DA net coverage error		
	Low series	Middle series	High series
2010 Census	1.00	**0.13**	-1.27
2020 Census	0.22	**-0.35**	-1.21

*Three series were produced for the 2020 DA. This change was made because a new methodology was used to estimate international migration, which is the largest source of uncertainty in the DA estimates.

Source: U.S. Census Bureau, 2010 and 2020 Demographic Analysis Estimates.

g. Advantages of DA
    i. Almost completely independent from the census enumeration
    ii. Detail available for selected population subgroups: single years of age, sex, selected race (black, non-black), Hispanic origin for young people (0-19 years in 2010, 0-29 years in 2020, and will be 0-39 years in 2030)
h. Limitations of DA
    i. National level only
    ii. Lack of accurate vital records, which is one reason why counts of older people are derived using Medicare records.
    iii. Sampling error due to the use of survey data for estimating the foreign-born population from the American Community Survey.
    iv. Differences in the reporting of race between the census and the data used for the DA estimates.
i. Age/Sex

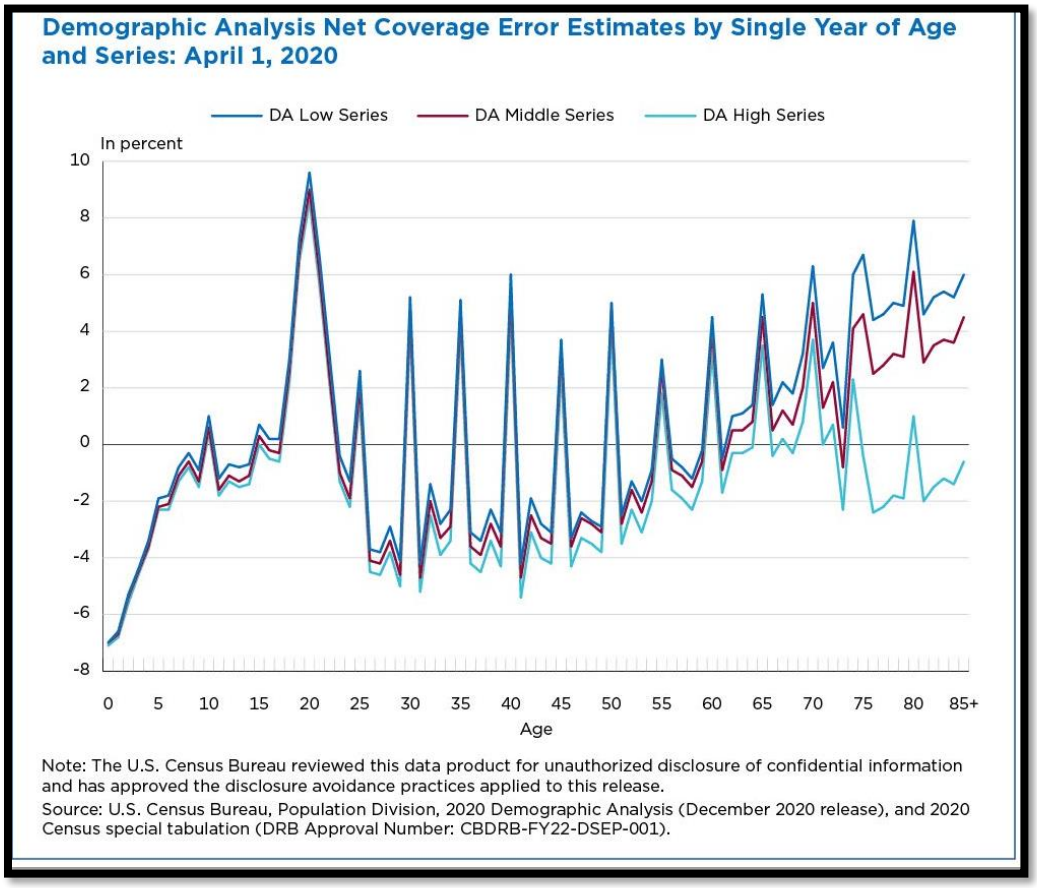
DA is the preferred method for obtaining coverage estimates by age and sex, shown below. The largest net undercount was for those under 5 years of age, with substantial overcounts for those in the college ages – 18 to 24 years. Net undercounts for males in the working ages were apparent, while for women a net undercount occurred just for those 25 to 29 years of age. Overcounts for both men and women were apparent from age 65 on.



j. Age Heaping

It is common for persons who respond on behalf of others in the census to report an age ending in 0 or 5, especially when responses are obtained from neighbors or in other situations where exact date of birth is not available. While age heaping occurs in every census, DA results show that the level of age-heaping in 2020 was much higher than in the 2010 Census, likely the result of the poorer quality

of proxy responses associated with the pandemic. In such situations, demographers may "smooth" age distributions to make them conform closer to reality.



k. Race

Unlike the decennial census, it is not possible for DA to include a category for "Some Other Race" (SOR). Therefore, in order for comparisons to be done, census data for SOR needs to be reallocated to specific races used by DA, namely the federal Office of Management and Budget categories.[3] After each census, the Census Bureau issues a special file which uses census data to reclassify SOR into the OMB categories by age/sex. The Modified Race File (MRF)

[3] These are White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander.

uses yet-to-be-released data from the 2020 Census, and is expected to be released in the summer of 2023.

Further Reading:

Overall: https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/da.html

Census Coverage Results: https://www.census.gov/newsroom/press-releases/2022/2020-census-estimates-of-undercount-and-overcount.html#:~:text=DA%20results%20by%20single%20year,higher%20than%20would%20be%20expected.

Detailed Methodology: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020da_methodology.pdf

2. Post-Enumeration Survey (PES)

a. **Overview:** The purpose of the Post-Enumeration Survey is to measure the accuracy of the census by independently enumerating a sample of blocks in the census, and comparing those results with the same blocks in the decennial census. Interviews are conducted with all housing units in each of the PES blocks and occupants are matched to their respective records in the decennial census. PES estimates are available at subnational levels -- each of the 50 states and DC for 2020.

b. **Method:** Prior to the beginning of the 2020 census, a specially trained group of fieldworkers canvassed the 10,000 blocks designated for the PES sample, leading to the creation of an independent list of addresses. Immediately after the census, field staff interviewed

housing units in these blocks, asking residents where they lived on April 1, 2020, about 161,000 housing units in total. The information they collected for the housing unit and household occupants was then matched to information collected in the decennial census to determine whether people were or were not counted. The PES uses a technique that is referred to as "dual-system estimation," with the two "systems" being the Post-Enumeration Survey and the census itself. After matching and field follow-up to resolve unmatched cases, an estimate of the total U.S. population is then derived, called the Dual-System Estimate (DSE) of the population. Essentially, the estimation takes advantage of the algebraic relationship that the ratio of people matched between the census and the PES to correct enumerations in the census equals the ratio of correct enumerations in the PES to the unknown true total population.[4]

Net Coverage Rate = 100 x (Census Total[5] – DSE Total) / DSE Total

Neither the census enumeration nor the DSE need to be perfect for this estimation process to work so long as they are independent of each other; but, together, they produce a better estimate of the population and an accurate idea of census coverage.

Efficient matching of housing unit addresses from the PES with the decennial census is key to this method. Lack of a match could represent a household or group of persons who were missed in the census, erroneously enumerated (e.g., duplication) in the census, or it could be the result of an error in the matching process. As the PES sample interviews move away from the actual census enumeration, recall error increases, as does the probability that households and/or persons have moved, which creates measurement problems. The

[4] DSE builds on the capture-recapture method used to estimate wildlife, such as fish in a lake—one captures a bunch of fish, tags them, returns them to the lake, and then takes a second bunch of fish, sees how many have tags and how many do not, and applies the DSE algebra.

[5] The Census Total includes not only correct enumerations, but also erroneous enumerations (e.g., duplicates) and whole-person imputations (people for whom nothing is known except that there is a person at the address).

pandemic created matching problems in 2020 -- 18 percent of addresses had an "undetermined" match status in 2020, compared to 8 percent in 2010.

2020 Census Coverage Results: PES

Year	PES	
	Net coverage error	Standard error
2010 Census	**0.01**	0.14
2020 Census	**-0.24**	0.25

Source: U.S. Census Bureau, 2010 and 2020 Post-Enumeration Surveys and Demographic Analysis Estimates.

The PES found that the 2020 Census had neither an undercount nor an overcount for the nation. It estimated a net coverage error of -0.24% (or 782,000 people) with a standard error of 0.25% for the nation, which was not statistically different from zero. Similarly, the PES did not show a statistically significant undercount or overcount for the 2010 Census.

Race/Hispanic Origin

DA estimates are not available for the major race categories or for older Hispanic people, mostly due to the deficiencies in the reporting of race and Hispanic origin in vital records. Thus, there is a greater reliance on the PES for estimates of coverage by race and Hispanic origin. As further explained in the *Using Demographic Benchmarks to Help Evaluate 2020 Census Results* blog[6], DA will first need to reconcile differences in how vital records categorize race with census

[6] https://www.census.gov/newsroom/blogs/random-samplings/2021/11/demographic-benchmarks-2020-census.html

results before any estimates of coverage by race can be released.[7] For this reason, analysis of coverage by race and Hispanic origin is derived from the PES.

The PES results show that Black or African American alone or in combination population had a statistically significant net undercount of 3.30%. Although this is not statistically different from the 2.06% undercount in 2010, it remains high relative to net coverage overall. The Hispanic population had a statistically significant undercount rate of 4.99%, a very high level and more than triple the 1.54% undercount in 2010. Only for American Indians on reservations was the undercount as high, at 5.6 percent.

Net Coverage Error Rates (Percents) for the Population by Race/Hispanic Origin					
United States					
2020 Census					
		2020	**SE**	**2010**	**SE**
Total (Household Population)		**-0.24**	**0.25**	**0.01**	**0.14**
White Non-Hispanic		***1.64**	**0.21**	***0.83**	**0.15**
Black or African American AOIC		***-3.30**	**0.61**	***-2.06**	**0.50**
Asian AIOC		***2.62**	**0.77**	**0.00**	**0.52**
American Indian AOIC on Reservations		***-5.64**	**2.72**	***-4.88**	**2.37**
Hispanic or Latino		***-4.99**	**0.53**	***-1.54**	**0.33**
***Percent net coverage error is significantly different from zero**					
AOIC=alone or in combination					
Source: Khubba, S, K. Heim, and J. Hong. (2022) "2020 Post-Enumeration					
Survey Estimation Report," PES20-G-01, U.S. Census Bureau, March					

Further Reading:

Overall: https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html

[7]This process is taking much longer for the 2020 census than in previous censuses because of the new Disclosure Avoidance System in use for 2020.

More Detailed Look at the Method: Marra, Elizabeth and Timothy Kennel, U.S. Census Bureau, 2020 Post-Enumeration Survey Methodology Report, PES20-J-01, *Source and Accuracy of the 2020 Post-Enumeration Survey Person Estimates*, U.S. Government Publishing Office, Washington, DC, March 2022. https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/2020-source-and-accuracy-pes-estimates.pdf

Characteristics: Khubba, Shadie, Krista Heim, and Jinhee Hong. (2022) "2020 Post-Enumeration Survey Estimation Report: National Census Coverage Estimates for People in the United States by Demographic Characteristics," U.S. Census Bureau, PES20-G-01, Washington, DC, U.S. Government Publishing Office, March. https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/national-census-coverage-estimates-by-demographic-characteristics.pdf

Methods of Estimating the Post-Census Population

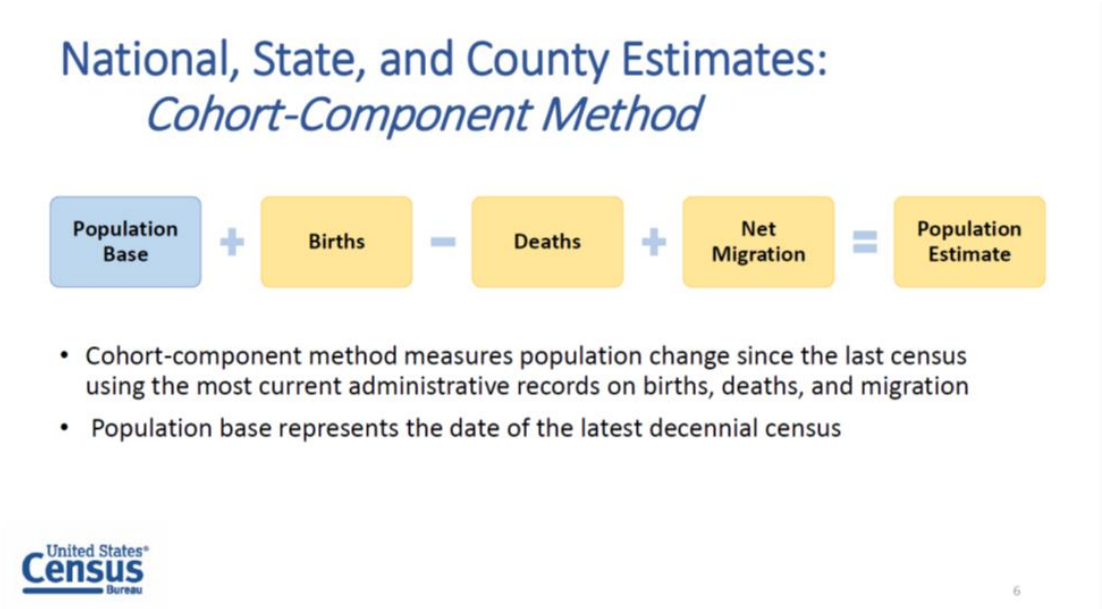
1. Nation, States, and Counties

a. Objective: to estimate the intercensal population annually
b. Relevance: the intercensal estimates are used in countless formulas for the distribution of more than 1.5 trillion dollars to states and localities annually and are used as "controls" for surveys (e.g., creation of weights and post-stratification adjustments for nonresponse). This includes the American Community Survey (ACS) and the Current Population Survey (done for the Bureau of Labor Statistics).
c. Referred to as the "Component Method" because estimates are constructed by adding or subtracting the components of population change, using the last census enumeration as a base.
d. Two very important principles govern the creation of population estimates: 1) Estimates for larger geographic areas are considered to be more reliable than those for smaller areas; 2) Estimates for small areas must be controlled or "raked" to larger areas, ensuring that all

estimates are consistent (e.g., counties add to states and states add to the national total).

e. All estimates are created by age/sex, race/Hispanic origin and must be made consistent (e.g., groups add to the totals for each geographic area).

f. The Census Bureau assumes that the population in Group Quarters (GQ)[8] remains constant throughout the decade unless they receive updated data on GQ population change. Such information can come from the military and/or Department of Veterans Affairs, and state partners as part of the Federal-State Cooperative for Population Estimates (FSCPE).



g. P2 = P1 (Base) + (Births – Deaths) + (In-migration – Out-migration)
where:

P2 - July 1 of estimate year (referred to as “vintage” year)

P1 – July 1 of previous year with the decennial census population used for the initial or base year of a decade

[8]Group quarters are defined as places where people live or stay in a group living arrangement that is owned or managed by an organization providing housing and/or services for the residents.

Births – Deaths – Derived from Vital Statistics reports from the National Center for Health Statistics (NCHS).

In-migration – Out-migration

- IRS tax returns (domestic migration)
- Medicare enrollment from the Center for Medicare and Medicaid Services (CMS) for persons 65 and Over
- Social Security Administration's Numerical Identification File.[9]
- ACS data on residence one year ago (abroad)
- ACS data on birthplace and year of arrival
- Estimates of emigration from the U.S. using survival rates calculated using mortality data from NCHS.

h. Because of delays in the release of data from the 2020 Census and problems with undercounts and overcounts by age, shown using DA at the national level, the Census Bureau created a modified base population from the 2020 Census. This "Blended Base" integrates three sources of data:
    1. **2020 Census Data:** 2020 Census data from the internal Census Edited File (CEF) tabulated into 2022 geographies at the subcounty level, infused with differentially private noise, and then aggregated to create county, state, and national estimates of total residents, households, and group quarters (by facility type)
    2. **2020 Demographic Analysis (DA) Estimates**: National population estimates by age and sex
    3. **Vintage 2020 Population Estimates for April 1, 2020**: Nation, state, and county population estimates by age, sex, race, and

[9] NUMIDENT is a database of all Social Security Numbers ever assigned, which is updated annually with new entries and any changes to a person's record.

Hispanic origin. (Note: the base for these estimates was the 2010 Census.)

2. Subcounty areas: Incorporated Places and Minor Civil Divisions

a. Use what is called the d*istributive housing unit method* where population change at the county level is "distributed" to subcounty jurisdictions as a function of change in the number of housing units.
b. For each subcounty area a count of housing units and a ratio of persons per housing unit count are calculated, to produce an "uncontrolled" population in households.
c. The "uncontrolled" calculation is done for each subcounty area and then all of the subcounty units are controlled (aka "raked") to the total for the county.
d. GQ population is held constant from the decennial census base or modified to reflect change based on the same inputs provided for counties.

Further Reading:

Vintage 2022 Population Estimates Methods Statement:
https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020-2022/methods-statement-v2022.pdf

Vintage 2021 Subcounty Population Estimates:
https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020-2021/2021-subco-method.pdf

Presentation on the 2020 Census "Blended Base":
https://www.census.gov/library/visualizations/2022/comm/creating-the-vintage-2021-blended-base.html