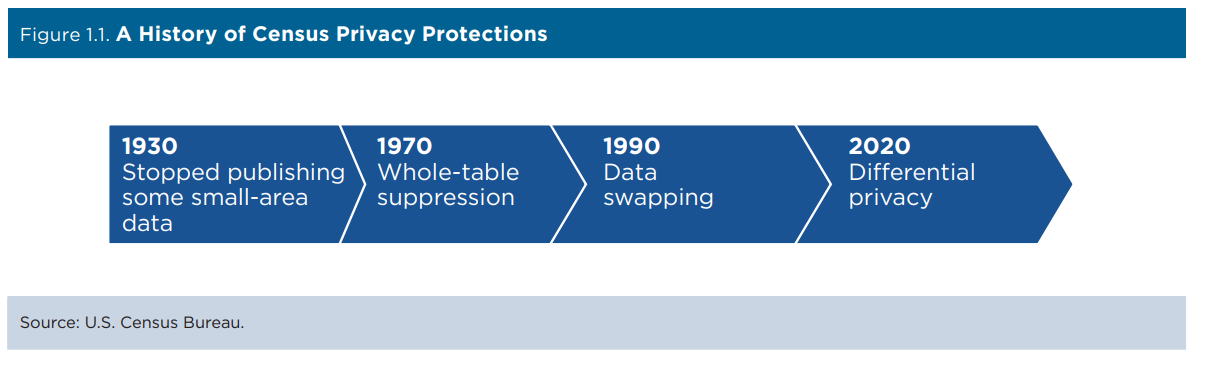
Differential Privacy Notes for CDPH Population Task Force

January 2023

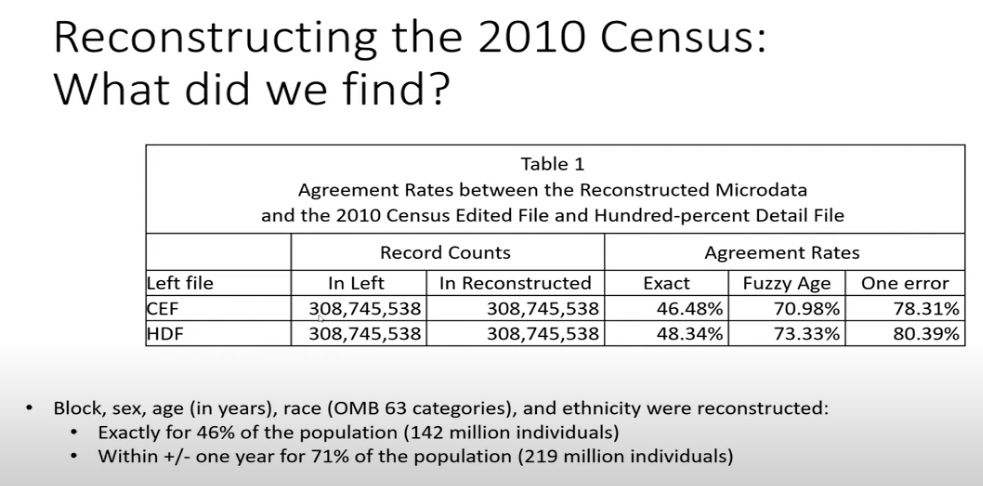
Submitted by Liz Reosti, Yasmine Castro, Pam Lough

**Executive Summary: Data products from the 2020 Census will have noise injected under a framework of Differential Privacy (DP). This means that, with a few exceptions, the population counts released in the 2020 Census Data Tables will not reflect the actual enumerated count. The Census Bureau advises against treating block-level data as accurate and recommends aggregating block-level data to a grouping of at minimum 450 people to avoid introducing excess error into analyses. Some scholars have found aggregations of data with DP applied with fewer than 1,000 people can significantly alter rate calculations. Data users should exercise caution when using or interpreting 2020 census data representing small populations.**

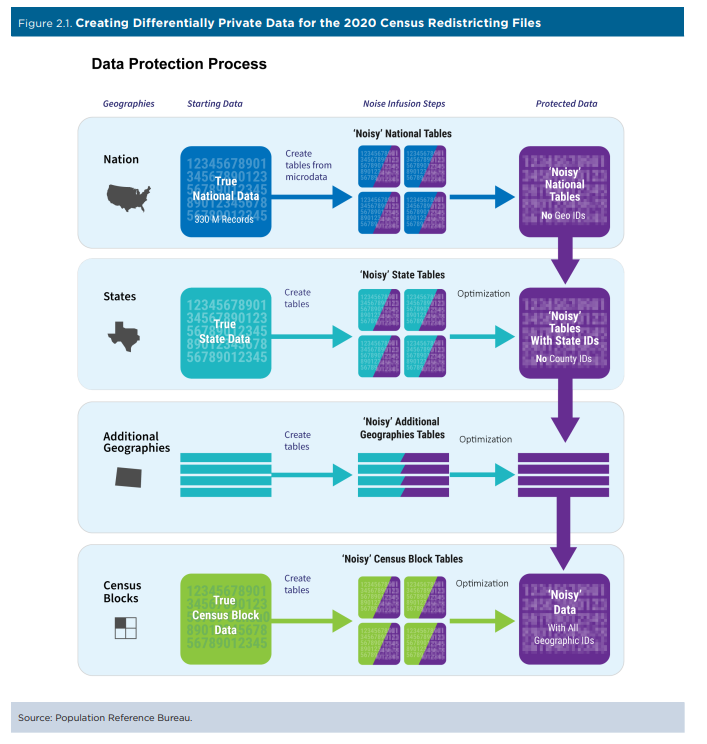
**Introduction:**  The Census Bureau has adopted a new Disclosure Avoidance System (DAS) based on a theory of differential privacy (DP). Differential Privacy refers to procedures that inject noise into data in order to protect privacy of individual respondents. Differential Privacy will be applied to all tables released as part of the 2020 Census. As a result, published data from the 2020 census will be a less accurate representation of actual microdata than previous decennial censuses.



**Background:** In the years since release of 2010 data tables, Census Bureau has demonstrated that micro-data could be reconstructed from published tables. This reconstruction violated disclosure rules according to Census Bureau. This reconstruction also demonstrated that previous DAS like swapping and small cell suppression were not sufficient to protect individual privacy considering ~ 50% of population are, “population uniques on the basis of block, sex, age, race, (OMB 63 categories[[1]](#endnote-1)) and ethnicity” (Abowd 2021). With the availability of commercial data sources, someone could correctly identify by name and address 77% of the population represented in 2010 census.



Using straightforward or naïve DP, noise is added at the most granular level (this would be block-level for Census data). This approach does not satisfy needs of Census Bureau, which still wants highly aggregated data (e.g., county or state-level) to be highly accurate and which needs data tables to be used to satisfy Voting Rights Act Section 2 majority-minority district identification. Instead the Bureau uses a Top-Down DP Algorithm (TDA) which injects noise first at the largest geographic level then to smaller sub-geographies. The process also involves post-processing the noise-injected data to ensure that a) no negative population counts and b) population counts at high levels of aggregation or relatively accurate. Noise is injected at the table-cell level with some exceptions (total state population, housing unit count/type) but not to the microdata.



The level of noise introduced to the data is guided by the privacy loss budget that is defined as the absolute upper bound of privacy loss that is deemed acceptable. Improving the accuracy in one dimension (such as a smaller level of geography) means there is less budget for accuracy in another dimension (such as detailed race categories). The amount of noise added is independent to the size of the population of the cell (relative error will be higher for smaller populations as the underlying population (denominator) is smaller. Smaller population groups, such as AIAN or NHPI groups, have a smaller budget as a result of their small numbers.

Graphical user interface, diagram

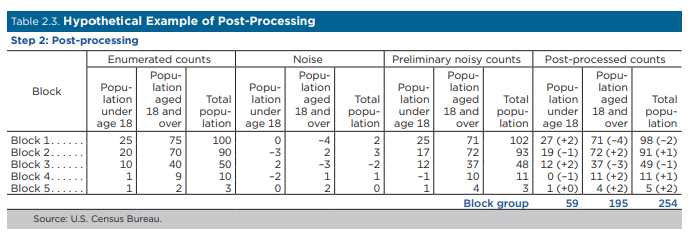
Description automatically generated

Source: U.S. Census Bureau (2021)

Post-processing is conducted to ensure tables are internally and hierarchicaly consistent and that non-negative integers are not published. The following statistics are considered by the bureau to be ‘invariants’ meaning the data will be released without noise infused.

* Total population (at the state and state-equivalents level) ​
* Total housing units (at the census block level)​
* Number of group quarters facilities by type (at the census block level)

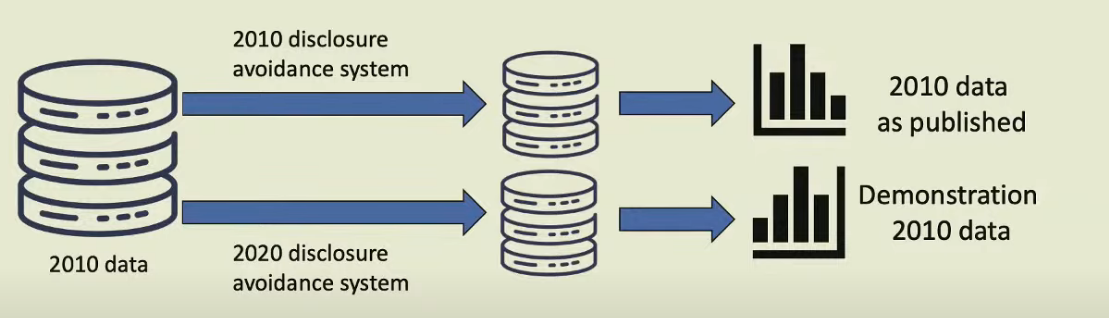
Additionally, AIAN tribal area data have a separate and greater privacy-loss budget than other census data resulting in greater accuracy for these data.



Some scholars (see Ruggles et. al 2019) have challenged the Census Bureau’s assertions about risk of disclosure in 2010 tables and necessity/appropriateness of DP.

**Demonstration Data:** Census Bureau released  [2010 Demonstration Data Products](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html) which reflect the 2010 census data with differential privacy applied allowing researchers to investigate impacts of differential privacy on analyses.

* + The initial 2010 demonstration data were released in May 2020
  + In March 2022 a modified version of the 2020 demonstration data was released with a larger privacy-loss budget and therefore less noise

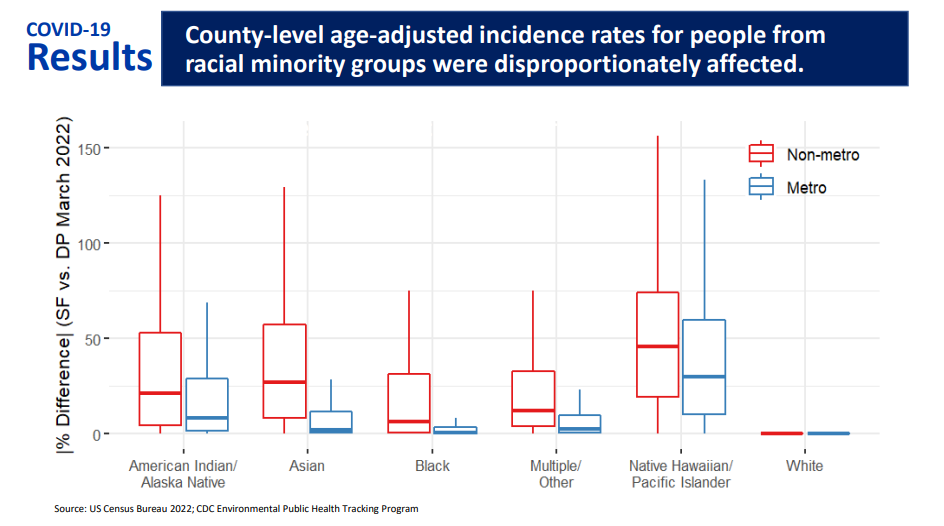


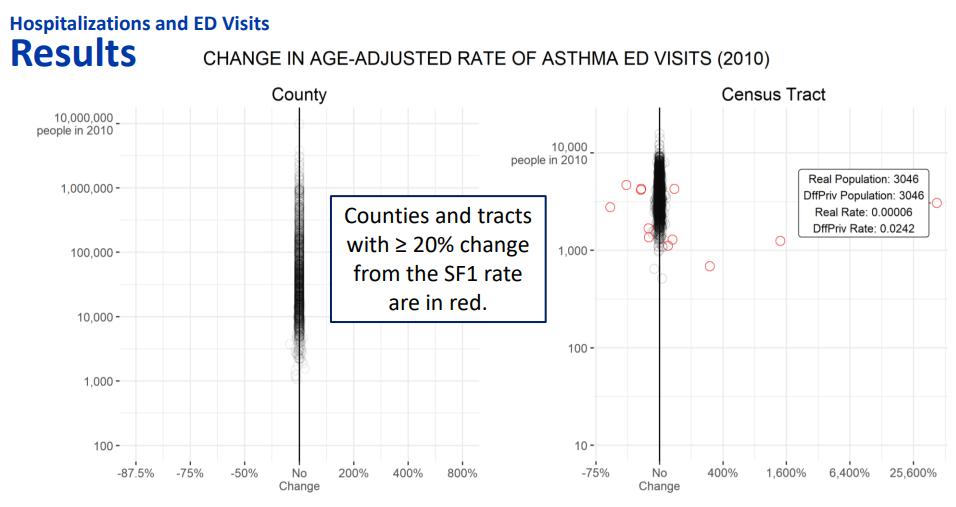
**Assessment/Key Findings:** Many analyses of the initial May 2020 demonstration data found that relatively small aggregations of the data contained significant error, with some variation in terms of what level of aggregation is required for accurate measurement. Hauer and Santos-Lozada (2020) show that population groupings with fewer than 1,000 people are subject to significant error. A later paper by the same authors (2021), however, indicates groupings smaller than 2,500 people will distort outcome measurements. For reference in the 2010 Census Data, >99% of California blocks had fewer than 1,000 people and 23% of block groups had fewer than 1,000 people.

In their 2020 paper, Hauer and Santos-Lozada suggest that differential privacy will more strongly affect mortality rate estimates for non-Hispanic blacks and Hispanics than estimates for non-Hispanic whites, with significant changes in estimated mortality rates for less populous areas and more pronounced changes when stratified by race/ethnicity. They also find that the error introduced through DP will disproportionately impact rural areas.

Krieger et. al. (2021) find that measures of health inequities at the census-tract level were not impacted by using 2010 Census data with DP applied.

The CDC found that county-level Covid-19 incidence rates for racial minority groups were disproportionality impacted using the March 2022 released demonstration data.





**Other Relevant Findings:** Kenny et. al. (2021) find “(differential privacy)… has a tendency to transfer population across geographies in ways that artificially reduce racial …. heterogeneity. “

Dyda, Purcell, and Curtis, et al (2021) pilot a new differential privacy tool, COVID-19 Real-Time Information System for Preparedness and Epidemic Response (CRISPER), to provide real-time data for the health sector.

Hodges and Cortes (2020) from Claritas found inconsistencies in the person and household counts in the initial demonstration data – cases where there were households but no population or more householders than there were population in block groups.

Kurz, König, and Emmert-Fees, et al, (2022) find the DP method introduces errors up to 10% into counts and proportions of Medicaid participation rate accuracy at the county level, especially for small subpopulations and racial and ethnic minority groups.

Petti and Flaxman (2020) list seven key choices that balance accuracy and privacy when implementing the Census Bureau’s TopDown approach to differential privacy.

Walter Schwarm and Jonathan Buttle of the California Department of Finance Demographic Unit compare an aspect of differential privacy to a privacy budget that is reduced by increased accuracy in data. They describe two methods of allocating the privacy budget, sequential and parallel, and propose methods to measure accuracy, bias and outliers. The 2020 Census was planning to address issues caused by differential privacy methods.

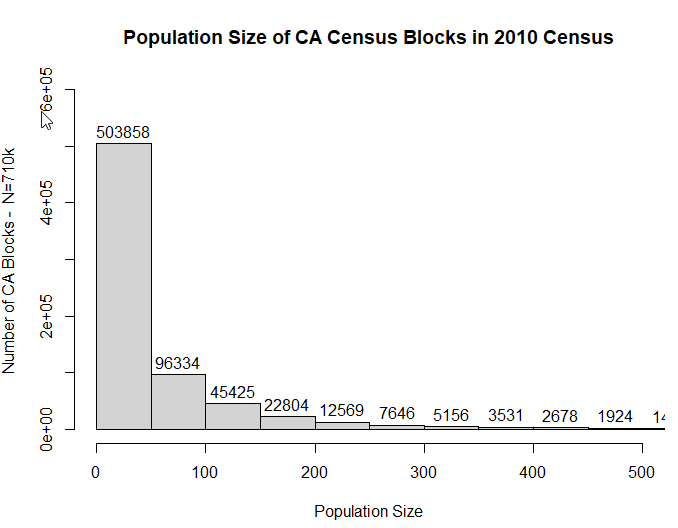
**Recommendations:** The Census Bureau recommends aggregating block-level data before use. They note that block groups of minimum size of 450-499 people are sufficiently large to provide reliable data. For census place and civil divisions a minimum population size of 200-249 is needed for accurate data. Census Bureau indicates, “Counts are consistent within tables, across tables, and across geographies. For example, rows within a table sum up to the parent row and universe.” Users should be cautious about drawing inferences based on observed changes between 2010 census data and 2020 census data at very small geographies (i.e. blocks). They also underscore that coverage error and sampling error exist in all census data previously released and that their estimates are that the errors due to coverage and sampling will be greater than the error introduced due to DP.

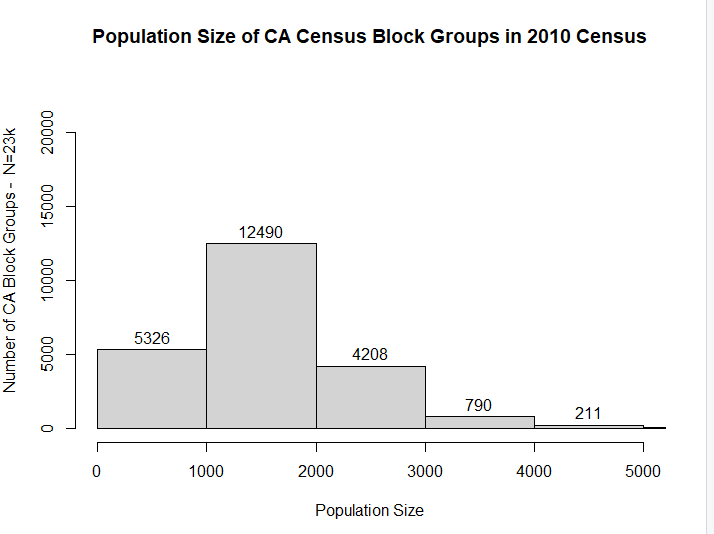
Cohen et. al. (2022) note that excluding and/or weighting observations with small counts can reasonably correct the attenuation bias you would see in a regression model using precinct-level data.

Dyrting, Flaxman and Sharygin (2022) show that smoothing age distributions from the 2022 demonstration data

Jiang, Feuer, and Li, et al, (2020) propose a bias-corrected rate estimator as well as its corresponding variance estimator that takes into account sampling errors in the denominators for age-standardized cancer rates.

**Future Impacts:** The Census Bureau has indicated that they do intend to eventually apply DP to the American Community Survey (ACS) but that that would not occur prior to 2025.





# **References:**

<https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan5-and-memo.pdf>

Census Bureau, Disclosure Avoidance for the 2020 Census: An Introduction, U.S. Government

Publishing Office, Washington, DC, November 2021.

Census Bureau (2022, April 12). Protecting the Confidentiality of 2020 Census Redistricting Data. <https://www.census.gov/library/fact-sheets/2021/protecting-the-confidentiality-of-the-2020-census-redistricting-data.html>

Abowd, John (2020) The U.S. Census Bureau Adopts Differential Privacy. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, <https://dl.acm.org/doi/abs/10.1145/3219819.3226070>

Abowd, John (2021) Differential Privacy and the 2020 Census in the United States. Presentation for American University <https://www.youtube.com/watch?v=u5h8ZC8C1uE>

Cohen, Aloni, Duchin, M, Matthews, JN and Bhushan Suwal; 2nd Symposium on Foundations of Responsible Computing (FORC 2021). Census TopDown: The Impacts of Differential Privacy on Redistricting <https://arxiv.org/pdf/2203.05085.pdf>

Cohen, A., Duchin, M., Matthews, J. N., & Suwal, B. (2022). Private numbers in public policy: Census, differential privacy, and redistricting. *Harvard Data Science Review*, (Special Issue 2).

Dyda A, Purcell M, Curtis S, Field E, Pillai P, Ricardo K, Weng H, Moore JC, Hewett M, Williams G, Lau CL. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. Patterns (N Y). 2021 Dec 10;2(12):100366. doi: 10.1016/j.patter.2021.100366. PMID: 34909703; PMCID: [PMC8662814](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8662814/).

# Dyrting, S., Flaxman, A. & Sharygin, E. Reconstruction of age distributions from differentially private census data. Popul Res Policy Rev 41, 2311–2329 (2022). <https://doi.org/10.1007/s11113-022-09734-2>

Hotz, V. J., & Salvo, J. (2022). A Chronicle of the Application of Differential Privacy to the 2020 Census. Harvard Data Science Review, (Special Issue 2). <https://doi.org/10.1162/99608f92.ff891fe5>

# Jiang J, Feuer EJ, Li Y, Nguyen T, Yu M. Inference about age-standardized rates with sampling errors in the denominators. Stat Methods Med Res. 2021 Feb;30(2):535-548. doi: 10.1177/0962280220962516. Epub 2020 Oct 15. PMID: 33059531.

Kenny, Christopher T., Shiro Kuriwaki, Cory McCartan, Evan Rosenman, Tyler Simko, and Kosuke Imai. (2021) .``The Use of Differential Privacy for Census Data and its Impact on Redistricting: The Case of the 2020 U.S. Census.'' Science Advances, Vol. 7, No. 7 (October), pp. 1-17.

Krieger N, Nethery RC, Chen JT, Waterman PD, Wright E, Rushovich T, Coull BA. Impact of Differential Privacy and Census Tract Data Source (Decennial Census Versus American Community Survey) for Monitoring Health Inequities. Am J Public Health. 2021 Feb;111(2):265-268. doi: 10.2105/AJPH.2020.305989. Epub 2020 Dec 22. PMID: 33351654; PMCID: PMC7811099.

Kurz CF, König AN, Emmert-Fees KMF, Allen LD. The effect of differential privacy on Medicaid participation among racial and ethnic minority groups. Health Serv Res. 2022 Dec;57 Suppl 2:207-213. doi: 10.1111/1475-6773.14000. Epub 2022 May 20. PMID: 35524543.

Hauer, M. E., & Santos-Lozada, A. R. (2021). Differential Privacy in the 2020 Census Will Distort COVID-19 Rates. Socius, 7. <https://doi.org/10.1177/2378023121994014>

Mueller, J.T., Santos-Lozada, A.R. The 2020 US Census Differential Privacy Method Introduces Disproportionate Discrepancies for Rural and Non-White Populations. Popul Res Policy Rev 41, 1417–1430 (2022). <https://doi.org/10.1007/s11113-022-09698-3>

# Petti S, Flaxman A. Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff. Gates Open Res. 2020 Apr 6;3:1722. doi: 10.12688/gatesopenres.13089.2. PMID: 32478311; PMCID: PMC7216402.

# Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research." AEA Papers and Proceedings, 109: 403-08. DOI: 10.1257/pandp.20191107

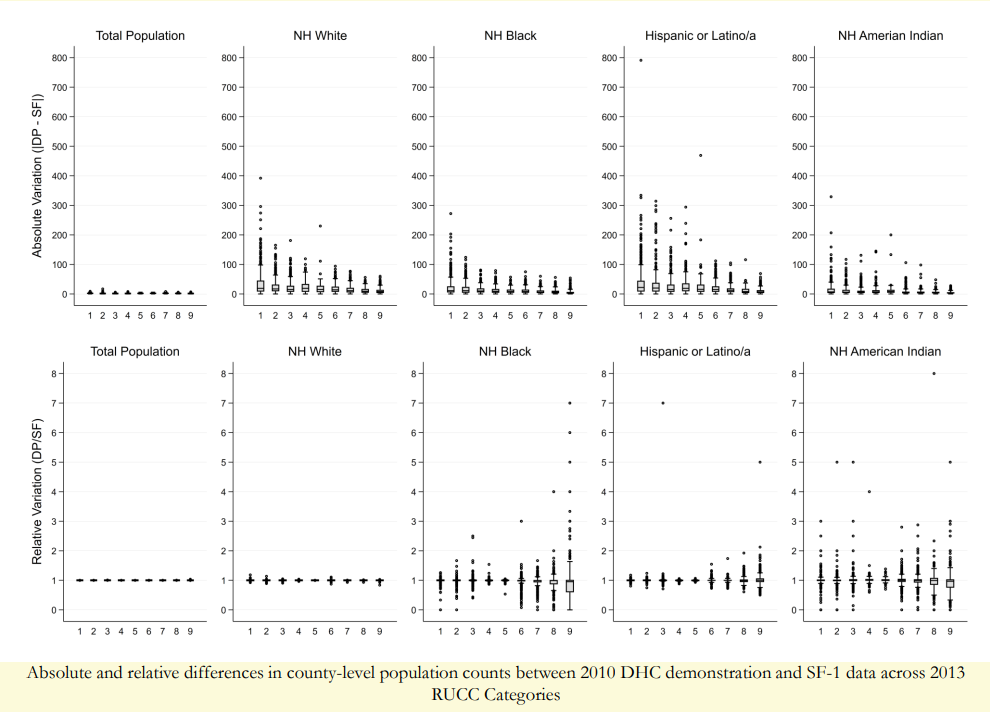
Santos-Lozada AR, Howard JT, Verdery AM. How differential privacy will affect our understanding of health disparities in the United States. Proc Natl Acad Sci U S A. 2020 Jun 16;117(24):13405-13412. doi: 10.1073/pnas.2003714117. Epub 2020 May 28. PMID: 32467167; PMCID: PMC 7306796. [P￼](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7306796/)

Walter Schwarm, Jonathan Buttle. California Department of Finance Demographic Research Unit. Differential Privacy and Census Data. [2020?] Available at: <https://scag.ca.gov/sites/main/files/file-attachments/jonathanbuttle.pdf?1604614125>

Wines, Michael (2022, April) “The 2020 Census Suggests That People Live Underwater. There’s a Reason.” New York Times

Presentations from the June 2022 Meeting of the National Academy of Sciences Committee on National Statistics, “2020 Census Data Products—Public Workshop on the Demographic and Housing Characteristics Files” https://www.nationalacademies.org/event/06-21-2022/2020-census-data-products-workshop-on-the-demographic-and-housing-characteristics-files?utm\_campaign=20220719msdecs1ccdtars&utm\_medium=email&utm\_source=govdelivery

1. There are 63 potential single and multiple race categories -- 6 categories for those who marked exactly one race and 57 categories for those who marked two or more races. These 57 categories of two or more races include the 15 possible combinations of two races (for example, Asian and White), the 20 possible combinations of three races, the 15 possible combinations of four races, the 6 possible combinations of five races, and the 1 possible combination of all six races.

   Source: Mueller, Tom County-level Discrepancies between SF-1 and DHC across the Rural-Urban Continuum. Presentation to the June 2022 Meeting of the National Academy of Sciences Committee on National Statistics, [↑](#endnote-ref-1)