Analysis of Census Bureau's August 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children

By
Dr. William P. O'Hare

September 2022

Analysis of Census Bureau's August 25, 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children

By
Dr. William P. O'Hare

Executive Summary

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect the confidentiality and privacy of respondents in the 2020 Census. This paper provides some information on how the use of DP in the 2020 Census is likely to impact the accuracy of data for young children (population ages 0 to 4).

The study is based on analysis of the most recent DP Demonstration Product released by the Census Bureau on August 25, 2022. The DP Demonstration Product issued on August 25, 2022 supersedes earlier DP Demonstration Products and focuses on data that will be in the 2020 Census Demographic and Housing Characteristics (DHC) file, which is scheduled to be released in May 2023.

The DHC file has most of the tables that were in Summary File 1 of the 2010 Census. The Demonstration Product released in August 2022 has data for population and housing units, but this analysis only examines data from the population file.

This paper presents analysis of the error introduced by DP by comparing the data as reported in the 2010 Census Summary File to the same data after the application of DP. According to the Census Bureau, the demonstration file released by the Census Bureau in August 2022 has been optimized for major use cases of the DHC tables.

Analysis presented in this paper found little impact of DP on data for young children for large (highly aggregated) geographic units like states or large counties.

However, the story is different for smaller geographic units. Many smaller areas have high levels of error in their data on young children after DP is applied. For example, the count of young children would exhibit absolute *percent* error of 5 percent or more in about 18 percent of Unified School Districts after DP is applied. The data also show that 64 percent of Unified School Districts had absolute *numeric* errors of 5 or more young children after DP is applied.

Errors of the magnitude shown above could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data and some schools would get less than they deserve.

Bigger absolute *percent* errors are evident for Hispanic, Black, and Asian young children in Unified School Districts. The mean absolute *percent* error for Non-Hispanic White young children was 5 percent compared to 28 percent of Hispanic young children, 35 percent for Black young children, and 45 percent for Asian young children. Differential accuracy among race and Hispanic Origin groups raises questions of data equity after DP is applied.

I also examined the accuracy/errors for the single year age 4 child population and found that errors for single year of age are particularly large. I found 52 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4, and 59 percent had absolute *numeric* errors of 5 or more children age 4

The results are similar for Places. Analysis shows that 46 percent of Places (cities, village, and towns) had absolute *percent* errors of 5 percent or more for age 0 to 4, and 38 percent of Places had absolute *numeric* errors of 5 or more young children.

I believe the most important type of error introduced by the application of DP are the large errors introduced for some geographic units. Analysis shows that 2 percent of Unified School Districts have Absolute Percent errors of 25 percent or more. In terms of *numeric* errors, 5 percent of Unified School District have absolute *numeric* errors of 25 or more young children. I urge the Census Bureau to take steps to reduce or eliminate these large errors for I believe the large errors injected by DP that will be most problematic.

The application of DP also caused a number of impossible or improbable results. After the injection of DP in the 2010 Census data included in the August 2022 Census Bureau Demonstration Product (U.S. Census Bureau 2022d Table 18), there were 163,077 blocks nationwide (1.5 percent of all blocks) that had population ages 0 to 17, but no population ages 18 or over, compared to 82 such blocks before DP was applied This result has two important implications.

First, blocks with children and no adults are a highly implausible situation and the large number of such blocks may undermine confidence in the overall Census results.

Second, these implausible results are likely due to young children being separated from their parents in 2020 Census DHC processing with DP. This separation of children and parent in data processing is an ongoing concern for data on young children and the production of future tables for children. This issue is particularly important in introducing DP into the American Community Survey, which is a key source

of child well-being measures (O'Hare 2022b). To understand the well-being of children, it is critical to understand the situation of a child's parents or caretakers. Moreover, if the same separation of children from their parents and caregivers occurs in the application of DP to the American Community Survey, it will eliminate reliable child poverty data which is based on household income. Child poverty rates are one of the most important measures of child well-being.

Based on the errors for the young child population with the privacy parameters for DP used in the August 2022 DP Demonstration Product, and the lack of clarity about the level of privacy protection from DP, I recommend the Census Bureau take steps to reduce the size of errors injected into the 2020 Census DHC file and in particular focus on trimming or eliminating the number of large errors.

This paper is meant to provide stakeholders and child advocates with some fundamental information about the level of errors DP is likely to inject into the 2020 Census data for the population ages 0 to 4. There are a couple of reasons for sharing this information with child advocates now. The 2020 Census results for some localities may include situations where the number of young children reported looks suspect. It is important to make sure child advocates are aware of the potential impact of DP so they can explain odd child statistics to local leaders.

There is a second reason for sharing this information with state and local child advocates. The U.S. Census Bureau is looking for feedback on the use of DP in the 2020 Census. The Census Bureau is looking for cases where census data are used to make decisions and the Census Bureau is asking data users to examine the DP

Demonstration Product to see if the error injected by DP make the data unfit for use. After reading this report, I hope you will convey your thoughts to the Census Bureau.

There is some latitude in how much error the Census Bureau will inject into the DHC files so feedback from census data users is important. If many users feel the current level of precision for data on young children in DP Demonstration Product is not accurate enough for some uses, there is a chance the Census Bureau could make the final data more accurate.

Stakeholders, child advocates, and data users should take advantage of this opportunity to communicate their thoughts to the Census Bureau before Census Bureau's Data Stewardship Advisory Committee makes a final decision on the privacy parameters to be used in the DHC file when it is released in May of 2023. Comments on the implications of DP in the August 2022 Demonstration File are due September 26, 2022, **Comments and responses can be sent to 2020DAS@census.gov.**

Analysis of Census Bureau's August 25, 2022 Differential Privacy Demonstration Product: Implications for Data on Young children

By
Dr. William P. O'Hare

Introduction

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect confidentiality and privacy of Census respondents in releasing data from the 2020 Census.[1] Analysis in this paper uses several measures to assess the accuracy of census data for young children after DP is applied. Young children are defined in this report as those ages 0 to 4. The analysis is based on the Demonstration Product released on August 25, 2022, which is the most recent available from the Census Bureau. This is the last Demographic and Housing Characteristics (DHC) Demonstration Product file the Census Bureau will release before they determine the final production parameters for the DHC file to be released in May 2023.

In short, DP injects errors in the data provided by respondents to make it more difficult for someone to be identified in the Census records. Adding or subtracting random numbers to the census results makes it more difficult to identify data for specific respondents because the data in the published census results no longer match what respondents submitted. The U.S. Census Bureau (2020e) provides more information

[1] The terminology in this arena can be confusing. Differential Privacy is sometimes called "formal privacy." The system developed for the 2020 Census DHC file has also been called the Top Down Algorithm or TDA. Since the application of differential privacy occurs within the Census Bureau's Disclosure Avoidance Systems (DAS) that term has sometimes been used to describe the use of differential privacy. To avoid confusion, I use the term differential privacy (DP) here to distinguish the version of DAS that includes DP from other versions of DAS.

on the use of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). In the fall of 2021, the Census Bureau released a primer on DP. (U.S. Census Bureau 2021d).

For an independent look at differential privacy see Boyd (2019) or Bouk and Boyd (2021). Hotz and Salvo (2020) offer a good review of DP early in the Census Bureau's development. A good overview of the evolution of the DP issue at the Census Bureau is provided by Boyd and Sarathy (2022).

It is fair to say that the introduction of DP in the 2020 Census has become a very controversial issue. In their review of the development of the DP issue over the past few years, Boyd and Sarathy (2022, page 1) conclude, "When the U.S. Census Bureau announced its intention to modernize its disclosure avoidance procedures for the 2020 Census, it sparked a controversy that is still underway."

One reason to focus on the impact of DP on the population ages 0 to 4 is the high net undercount of that population in the Census. Results of the 2020 Census evaluation using the Demographic Analysis method, show a net undercount of 5.4 percent for young children which was much higher than any other age group (U.S. Census Bureau 2022c).

Recent trends are also unsettling. From 1950 to 1980, the young children and adults had similar decade-to-decade improvement in terms of census coverage. However, after 1980 the trajectories were quite different. The coverage for adults continued to improve while the coverage of young children decreased dramatically (O'Hare 2022a). The net undercount of young children in the 2020 Census (5.4 percent) is higher than the young children undercount in the 1950 Census. I am not aware of any

other population group where census coverage is worse in the 2020 census than it was in the 1950 Census.

There are a couple of perspectives one could take regarding the high net undercount of young children and DP. On one hand, since the 2020 Census data for young children already has more error than data for other age groups, perhaps the amount of error injected by DP should be limited for this group. It does not seem fair to inject more error into data for groups that already have a lot of error in their census results. On the other hand, one might think that since the 2020 Census data for young children already has a lot of error, the added error from DP will not make much difference.

I focus first on data accuracy for Unified School Districts because schools are the public institution most closely associated with the child population and schools use demographics in a variety of ways. I next look at data for Places. Places include big cities and small villages. They typically have policymaking authority, and they often provide programs for young children such as childcare or preschool programs.

Several issues regarding DP are addressed in the Discussion section including the high error rate for blocks, breaking the relationship between children and parents, questions of equity, and the extent to which DP contributes to the lack of public trust in the census.

Background on Privacy in the Census

In every census, the U.S. Census Bureau faces a trade-off between privacy protection and accuracy. According to the U.S. Census Bureau (2020d),

"One of the most important roles those national statistical offices (NSOs) play is to carry out a national population and housing census. In so doing, NSOs have two data stewardship mandates that can be in direct opposition. Good data stewardship involves both safeguarding the privacy of the respondents who have entrusted their information to the NSOs as well as disseminating accurate and useful census data to the public."

The problem that DP is designed to fix is complicated as is the implementation of DP. The passage below from the U.S. General Accountability Office (2020, page 14) is the best short description I have seen on this issue.

"Differential privacy is a disclosure avoidance technique aimed at limiting statistical disclosure and controlling privacy risk. According to the Bureau, differential privacy provides a way for the Bureau to quantify the level of acceptable privacy risk and mitigate the risk that individuals can be reidentified using the Bureau's data. Reidentification can occur when public data are linked to other external data sources. According to the Bureau, using differential privacy means that publicly available data will include some statistical noise, or data inaccuracies, to protect the privacy of individuals. Differential privacy provides algorithms that allow policy makers to decide the trade-offs between data accuracy and privacy. "

It is important to note that the U.S. Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to U.S. U.S. Census Bureau (2018) some method of disclosure avoidance has been used by the U.S. Census Bureau since 1970. The 2010 Census data include some changes to original responses to help avoid disclosure of information about individual respondents, largely using a method called swapping.

The application of differential privacy allows the Census Bureau to control the amount of error injected into the data which is largely controlled by a parameter called "Epsilon." A higher-level of Epsilon means less error and more risk of violating confidentiality and a lower level of Epsilon means more error and less risk of violating confidentiality. In the latest material from the Census Bureau, Epsilon has been replaced with a term called Rho. It is my understanding Rho works the same way as

Epsilon in that a higher value means more accuracy and a lower value means more privacy protection. The point here is that the Census Bureau has control over how much error to inject into the data.

Measuring Accuracy

There is no consensus on exactly what measures should be used to assess the accuracy of DP-infused data, and there is no single benchmark to determine if DP-infused figures are "accurate enough for use." The U.S. Census Bureau (2020a) has suggested several measures of accuracy that could be used to evaluate the DP-infused data.

Like the Census Bureau's assessment of DP-infused data, I provide data for both absolute *numeric*al errors and absolute *percent* errors because either can be important and using both perspectives provide a more complete picture of the error profiles for geographic units. It may be a bit confusing presenting both *numerical* and *percent* errors, so I italicize the terms for help readers more easily distinguish which measure is being discussed.

For simplicity I only look at a few key measures here, but they provide sufficient information to reach some conclusions. The measures used here (mean absolute *numeric* error, mean absolute *percent* error, and large errors) are a subset of those discussed by the Census Bureau.

The DP demonstration file released by the Census Bureau on August 25, 2022, provides DP-infused data from the 2010 Census which can be compared to the 2010 Census data without DP to understand the likely impact DP has on data accuracy.

Errors are defined here as the difference between the data as originally reported in the 2010 Census Summary File and the same data after DP has been injected. The data from the Summary File is sometimes referred to as data without the application of DP in this report. Specifically, I subtract the value of the data with DP from the corresponding data without DP (Summary File) to find the error. For percentages, the difference is divided by the data without DP (i.e., Summary File) value.

I include a measure the Census Bureau calls the Mean Absolute Error (I label this Mean Absolute *Numerical* Error in the tables to distinguish it from the Mean Absolute *Percent* Error) and I also include the Mean Absolute *Percent* Error.

An absolute error reflects the magnitude of the error regardless of direction. A geographic unit with an absolute error of 10 percent could be 10 percent too high or 10 percent too low. Absolute errors are used to make sure positive errors and negative errors do not cancel each other out and make it appear as if there are no errors.

Percent error reflects the size of the error relative to the size of the population. An error of a given magnitude (say 10 young children) may be trivial in large Places but very significant in smaller Places. For example, a numeric error of 10 young children in a school district of 1,000 young children is only a 1 percent error, but a *numeric* error of 10 young children in a school district of 100 is a 10 percent error.

In addition to measures of average error, I include analysis on the number and percent of geographic units that have relatively large errors. I use two sets of benchmarks to identify large errors: one for absolute *numeric* errors and one for absolute *percent* errors.

The number and percent of large errors are likely to be the most important measures of accuracy in the 2020 Census. Large errors are likely to be a statistical problem and a public relationship problem for the Census Bureau, particularly if the errors are accompanied by large swings in funding. Data from the Census is often used to distribute federal and state dollars based on population (O'Hare 2020a: Reamer 2020, O'Hare and Rashid 2022: The Annie E. Casey Foundation, 2018). Large errors can result in implausible or impossible results. Such results are likely to cast suspicion on all the data from the Census Bureau and it is likely to undermine the confidence people have in all the census data.

Data Used in This Study

The Demonstration Product released in August 2022 reflects ongoing work at the Census Bureau. Starting in October 2019, the Census Bureau has released several Demonstration Products that reflect the injection of DP into 2010 Census data. The first official data from the 2020 Census with DP infused was the redistricting data file released by the Census Bureau in August 2021.

The DP Demonstration Product examined here is related to the Demographic and Housing Characteristics file that is scheduled to be released in May 2023. The Census Bureau (U.S. Census Bureau 2022d) has provided some measures of accuracy for the DP Demonstration Product, but they are somewhat limited.

Related to previous DP releases, my analysis of the DHC DP Demonstration Product released in March 2022, is available on the Count All Kids website (O'Hare 2022c). The Census Bureau's summary of all comments submitted in relation to the

March 2022 DP Demonstration Product are also available (U.S. Census Bureau 2022f).

The data used in my analysis were originally provided by the Census Bureau. The IPUMS- NHGIS unit at the University of Minnesota processed the Census Bureau files and put the data into more user-friendly tables. I analyzed the data produced by IPUMS-NHGIS unit which are available at https://nhgis.org/privacy-protected-demonstration-data

Geographic units where there were zero people ages 0 to 4 in either the 2010 data with DP or without DP were removed from the files for analysis. Observations with zeros for key measures produce very unusual results. This analysis does not include data for Puerto Rico.

Results for Age 0 to 4 in Four Kinds of Geographic Units

Table 1 provides a few key accuracy measures for the population ages 0 to 4 for four kinds of geographic units. These units were selected because they all have significant policy-making power regarding programs for children and they range widely in terms of population size.

The results shown in Table 1 indicate that DP is unlikely to have much of an impact on the young child data for states. The mean absolute *numeric* error for states for the population ages 0 to 4 is about 100 young children and the mean absolute *percen*t error rounds to zero.

Also, DP is unlikely to have much impact on young child county data for most counties. The mean absolute *numeric* error for counties is about 8 young children and mean absolute *percent* error is 0.92.

However, of the 3,142 counties examined here 36 percent (1,130) had less than 1,000 children ages 0 to 4 based on the Summary File results. For this subset of counties, DP may distort the data to a considerable degree. For the 1,130 counties with less than 1,000 young children, the mean absolute *numeric* error for ages 0 to 4 was 6 and the mean absolute *percent* error was 2.1.

Table 1 Key Statistics for Absolute *Numeric* and Absolute *Percent* Errors* for Children Ages 0 to 4 for Selected Geographic Units

	States	Counties ***	Unified School Districts ****	Places *****
Number of Units in the Analysis	50	3,141	10,864	28,548
Mean Size of District (Children ages 0-4 based on Summary File)	403,375	6,429	1,860	545
Mean Absolute *Numeric* Error**	100	8	9	5
Mean Absolute *Percent* Error	rounds to zero	0.92	4	13
Percent of Units with Absolute *Numeric* Errors of 5 or more Children	98	62	64	38
Percent of Units with Absolute *Percent* Errors of 5% or more****	0	3	18	46

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022. Data from IPUMS NHGIS, University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected (SF- DP).

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.

***DC is not included in the state data but is included in the county data

**** based on percentages rounded to two decimal points

***** this includes both incorporated places and Census Designated places

52 school districts removed from analysis because zero population age 0-4 in DP or SF

The situation is different for Unified School Districts and Places (shown in Table 1), where DP is likely to cause larger distortions (percentage-wise) for the young child

population. The mean absolute *numeric* error for Unified School Districts is 9 young children and it is 5 young children for Places. The mean absolute *percent* error for United School Districts is 4 percent and it is 13 percent for Places.

In my opinion the bigger problem is the number of extreme errors for these geographic units. For Unified School Districts and Places, the share of units that have extreme errors is substantial. Table 1 shows that 64 percent of Unified School District have absolute *numeric* errors of 5 or more children and 18 percent have absolute *percent* errors of 5 percent or more. For Places, 38 percent have absolute *numeric* errors of 5 or more children, and 46 percent have absolute *percent* errors of 5 percent or more. These extreme errors are more consequential than the mean figures. Accuracy for Unified School Districts and Places are explored in more detail in the next two sections of this report including more information on extreme errors.

Application of Differential Privacy to School District Data

The analysis first focuses on Unified School Districts since schools are the largest public institution focused on children. The Census Bureau reports there were 61.6 million children ages 3 to 17 enrolled in schools in 2019 (U.S. Census Bureau 2021a).

Schools often provide preschool programs for those under age 5. The Census Bureau shows there were over 5 million children enrolled in preschool in 2019, and more than half of all children age 3 and 4 are in preschool or nursey school (McElrath et al. 2022)

Reamer (2020) shows that $39 billion of federal funds were distributed by the U.S. Department of Education to states and localities in FY 2017 based on census-

derived data. Table 2 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data. In addition, many other government programs also use census-derived data to distribute funds targeted to children. This underscores why the accuracy of the population figures from the Census are so important.

Overall, Reamer (2020) identified 316 federal programs that use census-derived data to distribute about $1.5 trillion to states and localities in Fiscal Year 2017. About two-thirds of the 315 programs use substate data which underscores the importance of small area census data. When one is talking about billions of dollars, a small percent error can translate into a large dollar amount.

Table 2. Federal Programs in the U.S. Department of Education that Distribute Funds to States and Localities based on Census-derived Data	
	Amount Distributed in FY 2017
Adult Education - Basic Grants to States	$581,955,000
Title I Grants to LEAs	$15,459,802,000
Special Education Grants	$12,002,848,000
Career and Technical Education - Basic Grants to States	$1,099,381,000
Vocational Rehabilitation Grants to the States	$3,121,054,000
Rehabilitation Services - Client Assistance Program	$13,000,000
Special Education - Preschool Grants	$368,238,000
Rehabilitation Services - Independent Living Services for Older Individuals Who are Blind	$33,317,000
Special Education-Grants for Infants and Families	$458,556,000
School Safety National Activities	$68,000,000
Supported Employment Services for Individuals with the Most Significant Disabilities	$27,548,000
Program of Protection and Advocacy of Individual Rights	$17,650,000
Twenty-First Century Community Learning Centers	$1,179,756,000
Gaining Early Awareness and Readiness for Undergraduate Programs	$338,831,000
Teacher Quality Partnership Grants	$43,092,000
Rural Education	$175,840,000
English Language Acquisition State Grants	$684,469,000
Supporting Effective Instruction State Grants	$2,055,830,000
Grants for State Assessments and Related Activities	$369,051,000
Teacher Education Assistance for College and Higher Education Grants	$90,955,000
Preschool Development Grants	$250,000,000
Student Support and Academic Enrichment Program	$392,000,000
Total	$38,831,173,000
Source: Counting for Dollars. https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds	

It is also clear that census-related data are often used by states to distribute state government money, but as far as I can tell, there is no systematic data on how much money is distributed by states based on Census data (O'Hare 2020a).

At the National Academy of Sciences, Committee on National Statistics workshop on DP which was held in December 2019 there were several presentations reflecting implications of DP-infused data for young children and school districts (Vink

2019; O'Hare 2019; Nagle and Kuhn 2019). Note that some of these analyses are now outdated but they may be useful for framing issues.

O'Hare (2021) focuses on the accuracy of population ages 0 to 17 for Unified School Districts based on data from the Census Bureau's redistricting file released in August 2021. In addition, my analysis of the DP Demonstration Product's impact on data for young children based on the DP file released in March 2022 by the Census Bureau can be found on the Count All Kids website (O'Hare 2022c).

Demographic data are used for several important school district applications. Population projections are often used to plan for expanding (or reducing) school facilities, staff, and other school-related needs. Demographic projections are typically based on Decennial Census data. Current and projected demographic data are often used to construct attendance boundaries to keep classrooms from becoming overcrowded. Constructing attendance boundaries often include sensitivity to racial composition, so small area demographics by race are important. Such activities often require very small area data such as census blocks. Demographers who work extensively with school districts report that census blocks are a critical geographic unit for their work (Cropper et al. 2021).

Many school districts are governed by school boards which are often elected from single member districts. Such districts must meet the usual legal requirements of redistricting such as having districts with equal population size. Such redistricting must also meet the requirements of the Voting Rights Act, which means small area tabulations of population by race and Hispanic origin are important.

Once children get into the K-12 school system, school systems have pretty good data for forecasting the number of children to expect in each grade the following year. From that perspective it is the cohort age 0 to 4 that is the biggest unknown for many school systems. Therefore, this is the most important age group for examining the amount of error injected by DP.

DP has a bigger impact, percentage-wise, in smaller populations and the majority of Unified School Districts are relatively small. Many of the 10,864 Unified School Districts in this analysis are very small; 7,475 (69 percent of all Unified School Districts) had a young child population of less than 1,000, and 1,454 districts (13 percent of all districts) had a young child population less than 100 in the 2010 Census. The translation of small *numeric* errors into large *percent* errors is also more apparent in looking at data for Hispanic, Black, and Asian groups within Unified School Districts because those are typically smaller population groups.

Table 3 shows several measures of accuracy/error for 10,864 Unified School Districts in the 2010 Census used in this analysis.[2] The data are provided for all young children (all races) as well as for Non-Hispanic White Alone young children, Hispanic young children, Black Alone young children, and Asian Alone young children. For the remainder of this report when I use the term Black or Asian, it means Black alone or Asian alone. Other race groups were not examined here because the numbers were small, they were often highly clustered, and time was limited.

[2] Recall that districts where there was a zero for population age 0 to 4 in the DP or SF file were not included in the analysis. Also, recall Puerto Rico is not included.

Data in Table 3 show the majority of Unified School Districts have at least one Black child, one Hispanic child, and one Asian child. But many districts have relatively few young children of color. The average number of Hispanic young children in Unified School Districts where there was at least one Hispanic was 521, for Blacks it was 384. and for Asians it was 143. These numbers are well below the overall average of 1,860 young children. The relatively small number of Black, Hispanic, and Asian young children in many districts results in these groups having larger absolute *percent* errors.

Table 3 shows the mean absolute *numeric* error for all young children (all races) in Unified School Districts is 9 young children. Data in Table 3 shows for all children, the mean absolute *percent* error was 4. But these measures mask big differences among race and ethnic groups.

The mean absolute *numeric* errors for race and Hispanic Origin groups are smaller than for all children (8 for Hispanic young children 6 for Black young children, and 4 for Asian young children), compared to 9 for all children, as these are smaller population groups in general

On the other hand, mean absolute *percent* error was 4 percent for all children, 28 percent for Hispanic, 35 percent for Blacks young children, and 45 percent for Asian young children (Table 3).

Table 3 Summary of Key Error* Statistics for Children Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin					
	All young children	Non-Hispanic White Alone	Hispanic	Black**	Asian**
Number of units in the analysis	10,864	10,841	10,238	7,548	6,251
Mean number of young children in district (in group column heading)	1,860	945	521	384	143
Mean absolute numeric error***	9	9	8	6	4
Mean absolute percent error	4	5	28	35	45
Percent of units with errors of 5 or more young children	64%	64%	49%	40%	31%
Percent of units with errors of 5% or more	18%	23%	66%	63%	70%
Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org					
Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.					
* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.					
** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean					

Recall that absolute errors reflect the magnitude of the error without regard to the direction of the error. Absolute errors are used in this analysis so that positive and negative errors do not cancel each other out in constructing an average or mean.

Large Errors in Unified School Districts

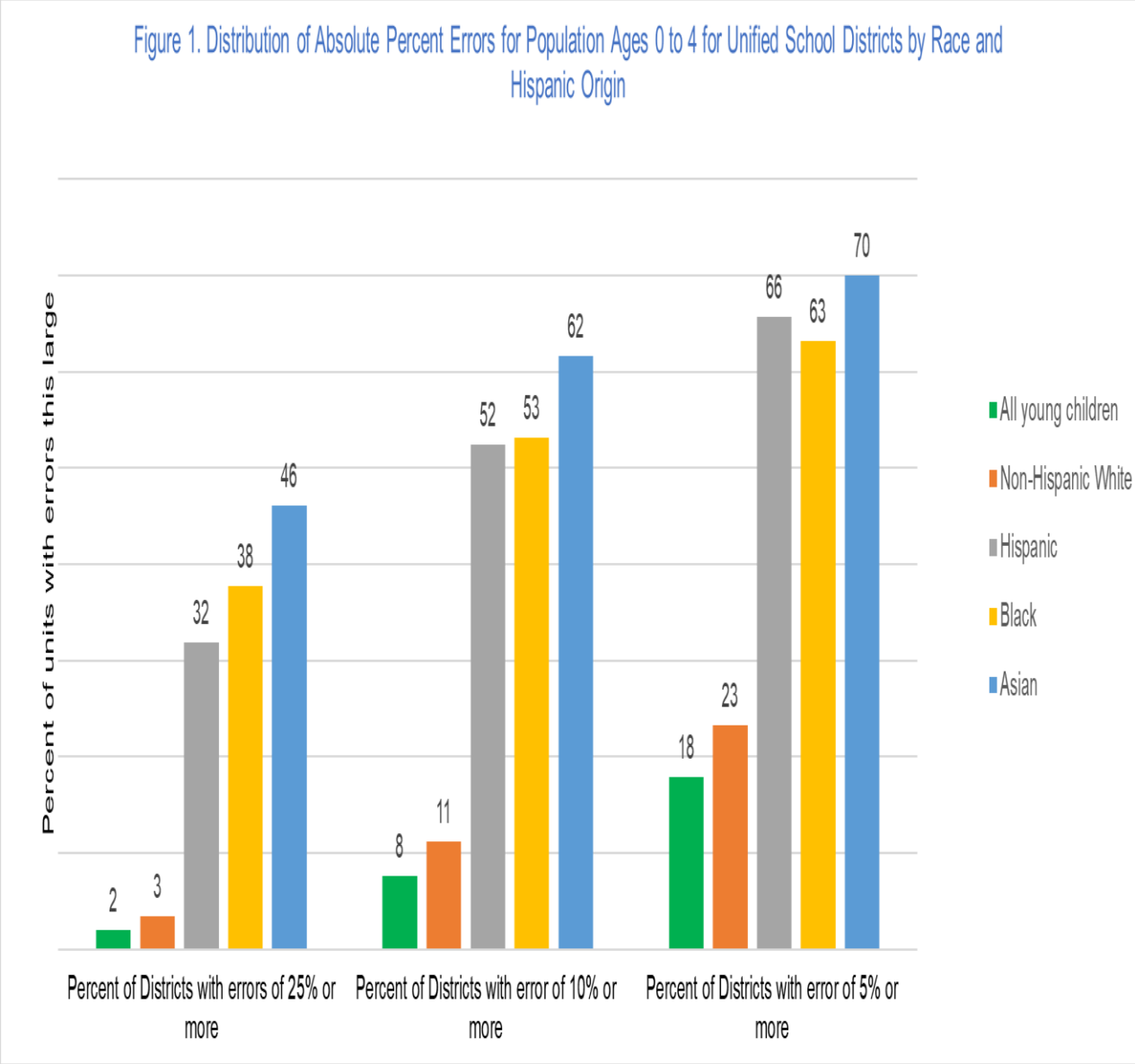
Means or averages are helpful, but they do not reveal the full story. Large errors can be problematic even if the overall mean error is relatively low. An examination of the distribution of Unified School Districts by error size can provide more information on the relative accuracy of the DP-infused data.

There is no consensus on what constitutes a large error and definitions probably vary with different applications. I show three benchmarks for large absolute *percent* errors. The 5 percent or more and 10 percent or more categories are used in several publications. I added the 25 percent plus category to look at the most extreme errors. Errors of 25 percent or more are likely to be very problematic. These thresholds are judgmental, but they provide a reasonable range of errors.

To be clear, the districts with more than 25 percent with large errors are also counted in the categories for more than 10 percent error and more than 5 percent error.

Distributions of absolute *percent* errors are shown in Figure 1 which shows that for all young children, 18 percent of districts had absolute *percent* errors of 5 percent or more, compared to 23 percent of Non-Hispanic White Alone, 66 percent for Hispanic young children, 63 percent for Black young children, and 70 percent for Asian young children. Since minority groups are smaller in population size, it is not surprising that there are more extreme absolute *percent* errors. There is a similar pattern by race and Hispanic Origin for other benchmarks.

In the largest error category (25 percent or more) the numbers are quite low for all young children and non-Hispanic whites alone young children, but quite high for Black, Hispanic, and Asian young children. Figure 1 shows that 32 percent of Unified School Districts have absolute *percent* errors of 25 percent or more for Hispanics, compared to 38 percent for Blacks and 46 percent for Asians. Figure 1 also shows that for young children of color, absolute *percent* errors of 25 percent or more are not unusual. Only two percent of Unified School Districts have absolute *Percent* Errors of 25 percent or more, but this amount to about more than 200 Districts nationwide.



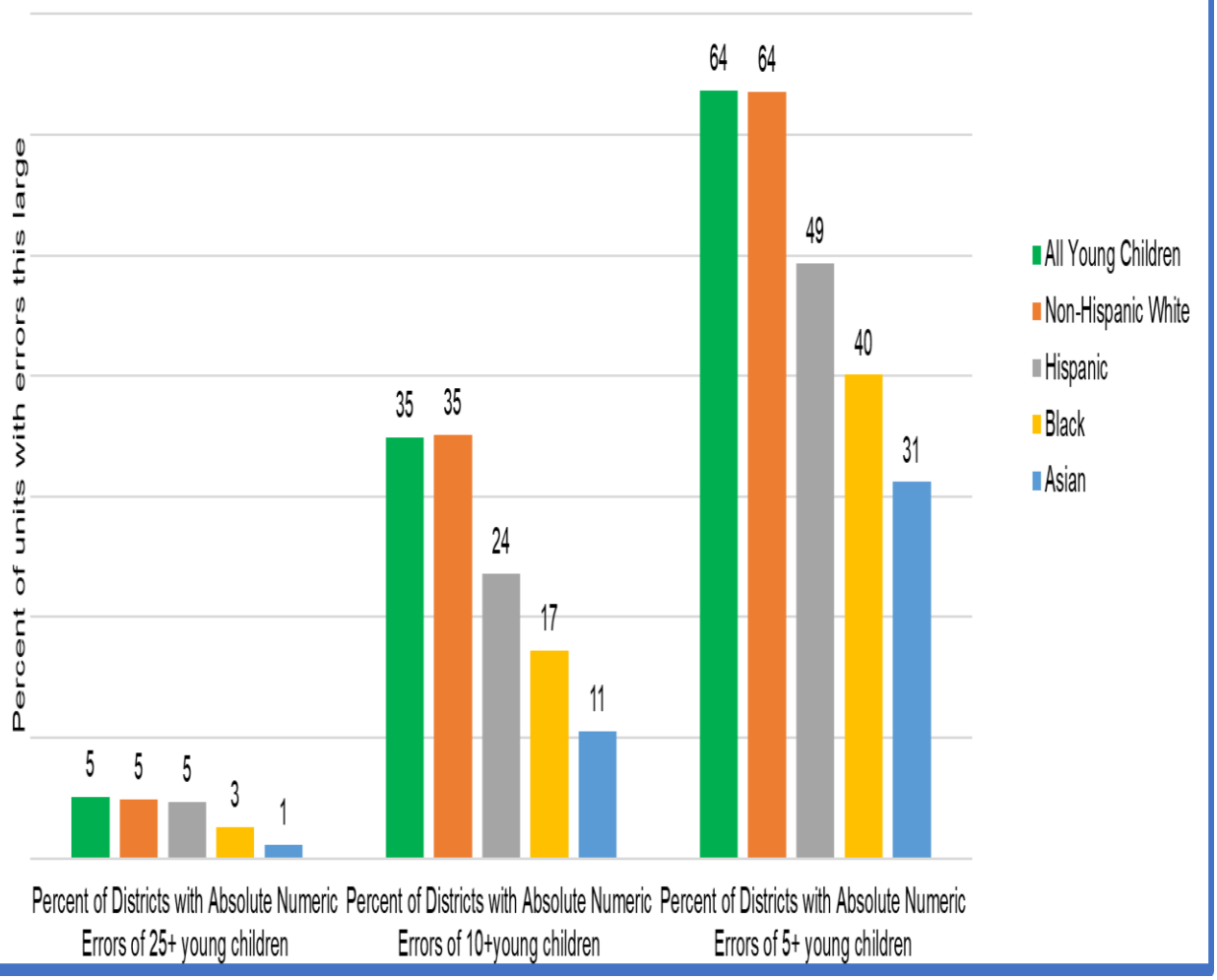
I use three benchmarks for large absolute *numeric* errors. The 5 persons and 10 persons categories of error have been used in other publications. I added the 25 persons plus category to look at the most extreme errors. Errors of 25 or more young children are likely to be very problematic in many Unified School Districts.

Figure 2 shows 64 percent of the Unified School Districts had errors of 5 young children or more for young children of all races but the figures for minority groups are

smaller: 49 percent for Hispanic young children, 40 percent for Black young children, and 31 percent for Asian young children.

In Figure 2, in each category of absolute *numeric* errors (5 young children, 10 young children, and 25 young children), there are many fewer districts that have this level of error for Hispanic, Black, and Asian young children than there are districts that have this level of error for all young children or Non-Hispanic White young children. This is because these are generally smaller populations.

Figure 2. Distribution of Absolute *Numeric* Errors for Population Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin



There are relatively few Unified School Districts with very large absolute *numeric* errors. Only 5 percent of Unified School Districts have errors of 25 young children or more, compared to 5 percent of Hispanic young children, 3 percent for Black young children, and 1 percent for Asian young children. For those districts that have errors of 25 percent or more because of the application of DP, the results are likely to be a substantial problem

The national numbers shown above mask a lot of variation across states. Table 4 shows states ranked on two key measures of accuracy (mean absolute *numeric* error and mean absolute *percent* error) for Unified School Districts. The mean absolute *percent* error ranges from a low of 3.4 for Vermont to a high of 16.9 percent in California. The mean absolute *percent* error for states ranges from a low of 0 for Hawaii (Hawaii only has one unified school district) to a high of 15.8 for Montana.

Table 4. States Ranked on Mean Absolute Numeric Error and Mean Absolute Percent Error for Children ages 0 to 4 by Unified School Districts

Rank*		Average of Absolute *Numeric* Difference		Rank*		Average of Absolute *Percent* Difference
1	California	16.9		1	Montana	15.8
2	Delaware	13.9		2	Alaska	14.0
3	Arizona	13.3		3	Maine	11.2
4	Hawaii	13.0		4	Vermont	10.1
5	Michigan	11.5		5	Washington	8.1
6	New York	11.2		6	North Dakota	7.4
7	Florida	10.1		7	Nebraska	6.6
8	Texas	10.1		8	South Dakota	6.6
9	Mississippi	9.9		9	Oregon	6.2
10	Arkansas	9.8		10	Colorado	6.1
11	Washington	9.8		11	New Mexico	5.8
12	Illinois	9.8		12	Oklahoma	5.3
13	Oregon	9.7		13	Idaho	5.2
14	South Carolina	9.6		14	Texas	5.1
15	Missouri	9.5		15	Kansas	4.9
16	Oklahoma	9.4		16	Indiana	4.6
17	North Carolina	9.3		17	Wyoming	4.4
18	Utah	9.3		18	Iowa	4.2
19	Minnesota	9.3		19	New Hampshire	4.1
20	Wisconsin	9.2		20	Missouri	4.1
21	Ohio	9.0		21	Minnesota	3.4
22	Idaho	8.6		22	Ohio	3.0
23	Iowa	8.6		23	Wisconsin	3.0
24	New Mexico	8.5		24	New York	2.9
25	Louisiana	8.3		25	Arkansas	2.9
26	Colorado	8.2		26	Illinois	2.7
27	Indiana	7.9		27	Michigan	2.5
28	Kansas	7.9		28	Arizona	2.1
29	Maryland	7.9		29	California	1.8
30	Tennessee	7.8		30	New Jersey	1.6
31	Georgia	7.8		31	Nevada	1.5
32	Wyoming	7.8		32	Mississippi	1.4
33	Alabama	7.7		33	Kentucky	1.3
34	Pennsylvania	7.5		34	Pennsylvania	1.0
35	Connecticut	7.4		35	Utah	0.9
36	Nebraska	7.3		36	Massachusetts	0.9
37	Virginia	7.2		37	Tennessee	0.8
38	Nevada	7.2		38	Virginia	0.8
39	South Dakota	7.0		39	Alabama	0.7
40	Kentucky	7.0		40	South Carolina	0.7
41	Massachusetts	6.5		41	Georgia	0.7
42	New Jersey	6.2		42	Rhode Island	0.7
43	Montana	5.6		43	Connecticut	0.6
44	North Dakota	5.3		44	West Virginia	0.6
45	New Hampshire	5.1		45	Delaware	0.6
46	West Virginia	5.0		46	North Carolina	0.5
47	Alaska	4.9		47	Louisiana	0.4
48	Rhode Island	4.8		48	Florida	0.3
49	Maine	3.6		49	Maryland	0.2
50	Vermont	3.4		50	Hawaii	0.0
U.S. Average		9				3.7

Source: Authors analysis of Demonstration Product released by the Census Bureau August 25, 2022 after process by IPUMS NIHGIS at the University of Minnesota

* Ranking is based on unrounded data.

Analysis for Age 4

In the Demonstration Product released in August 2022, the Census Bureau provided data by single year of age and sex for the population under age 20. I analyze this data for age 4 for Unified School Districts. I selected age 4 because that is often used by school systems to predict the number of kindergarteners to expect in the following school year. I do not see any reason why the metrics for age 4 would be much different than the metrics for any other single year of age.

Table 5 provides the key metrics for the comparison of age 4 in Unified School Districts in the 2010 Census file with and without DP. Districts with no people age 4 in the DP or SF file were not used in the analysis. The mean absolute *numeric* error was 9 and the mean absolute *percen*t error was 11 percent for age 4

A large share of Unified School Districts had large errors in both *numeric* and *percent* terms. About three out of five (59 percent) of Unified School System had absolute *numeric* errors of 5 or more children and 52 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4.

With errors of this magnitude for single year of age, one has to wonder if this data is worth producing. This is particularly true for smaller districts where the errors are likely to be larger percentage-wise. It is not clear how users are supposed to manage data with this degree of uncertainty

.

Table 5. Unified School District Error* Metrics for Age 4	
Number of Units in Analysis	10,782
Mean number of children age 4 in Summary File	376
Mean Absolute *Numeric* Error	9
Mean Absolute *Percent* Error	11
Percent of units with Absolute *Numeric* Error 5+ children age 4	59
Percent of units with Absolute *Percent* Error 5%+ **	52
Source: Author's analysis of Demonstration Product released by the Census Bureau on August 25, 2022 after processing by IPUMS NHGIS at the, University of Minnesota www.nhgis.org	
* In this paper, errors reflect the difference between the 2010 Census data without and with DP injected.	
Data in this table does not include Puerto Rico or geographic units with zero population age 4 in 2010 Summery File or DP-Infused file.	
** analysis is based on figures rounded to two decimal points	

Data for Places

Census Places are geographic units used by the U.S. Census Bureau to publish data. They range from Places with millions of people such as Los Angeles and New York City, to the smallest villages and towns.

Places include both incorporated Places and Census Designated Places (CDPs). There are a little more than 29,000 Places for which the infusion of DP data was produced in the August 16, 2022 (DP Demonstration Product) and most of them (over 19,000) are Incorporated Places rather than Census Designated Places (CDPs). Incorporated Places are legally bounded entities such as cities, boroughs, towns, or villages (names may vary depending on the state). Census Designated Places (CDPs) are statistical entities used in the Census. They are unincorporated communities where

there is a concentration of population, housing, and commercial structures and they are identifiable by name. There are nearly 10,000 CDPs for 2010 Census data.

Cities, villages, and towns might want to know about the number of young children in their area for things like planning youth activities, child facilities, and day care centers. The preschool-age population is also useful for forecasting future school enrollments.

Table 1 shows the mean absolute *numeric* error for Places was 5 and the mean absolute *percent* error was 13 percent. The high percent error is not surprising because many of these Places are small. There were 1,422 Places where the number of young children was less than 100, and 9,012 Places where the number of young children was less than 500, based on the 2010 Summary File.

Figure 3 shows the distribution of Places by absolute *percent* error using the same thresholds used for Unified School Districts. The data in Figure 3 shows that almost half (46 percent) of Places had absolute *percent* errors of 5 percent or more for the young child population and 15 percent had absolute *percent* errors of 25 percent or more. Since Places are generally smaller (in population size) than Unified School Districts, it is not surprising that the percentages are larger for Places than for Unified School Districts.

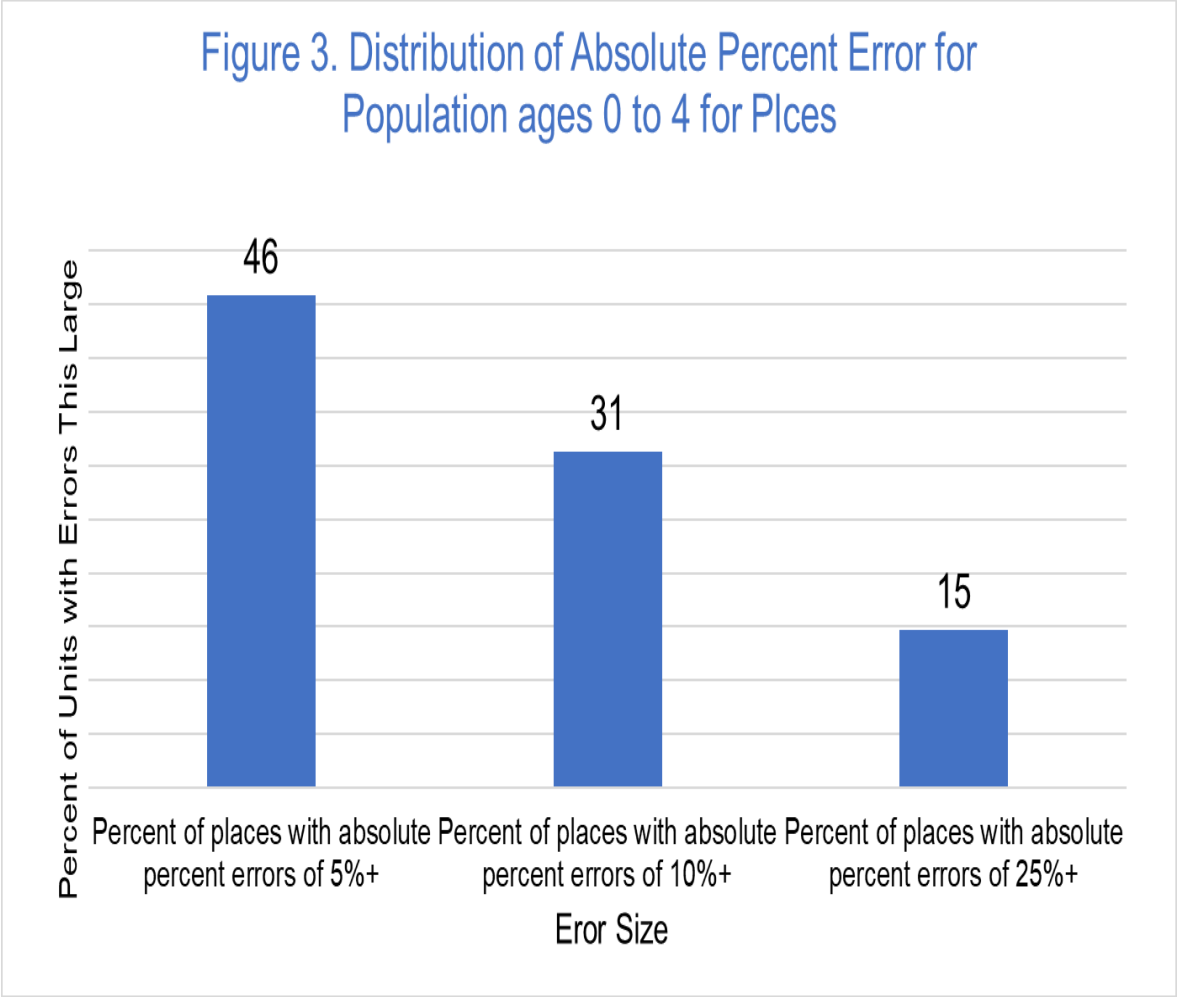


Figure 4 shows the distribution of Places by absolute *numeric* errors using the same categories as Figure 2. Data show 38 percent of the Places had absolute *numeric* errors of 5 or more young children, and only 2 percent had absolute *percent* errors of 25 or more young children.

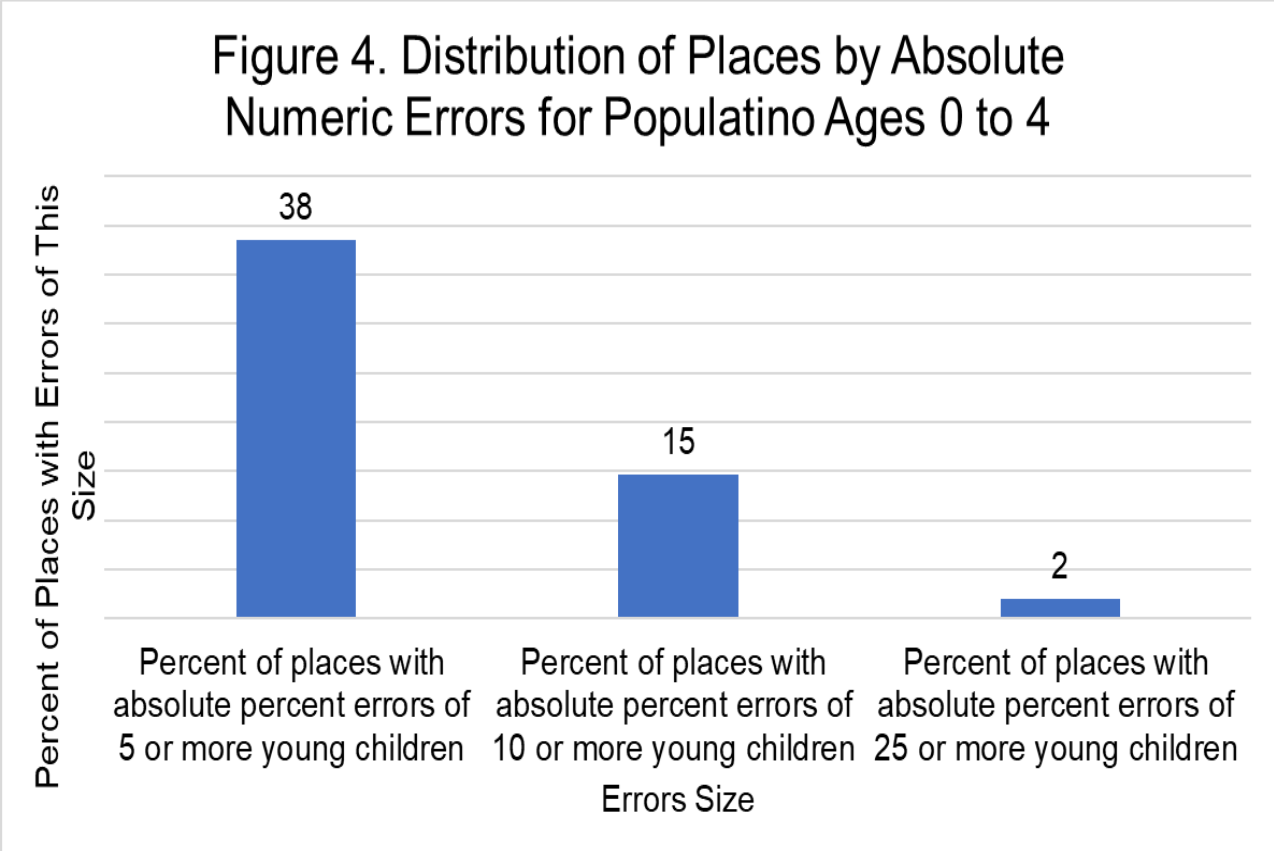


Table 6 shows states ranked on the percent of places in a state with absolute *percent* errors of 5 percent or more. Data for errors of 10 percent or more and 25 percent or more are also provided in the Table 6.

There is a lot of variation across the states. For example, 67 percent of the Places in Nebraska had absolute *percent* errors of 5 percent or more, compared to 25 percent of Places in New Jersey.

Table. 6 States Ranked by Percent of Places in State with Absolute Percent Errors of 5 Percent or More for Ages 0 to 4

Rank*	State	Number of Places in State	Errors of 5% +	Errors of 10% +	Errors of 25% +
1	Nebraska	543	67	50	26
2	New Mexico	419	64	50	28
3	Vermont	116	63	49	27
4	North Dakota	346	63	51	30
5	South Dakota	353	62	47	26
6	Montana	341	62	51	34
7	Wyoming	184	60	48	28
8	Alaska	310	60	45	28
9	West Virginia	397	60	42	20
10	Oklahoma	708	57	42	20
11	Arizona	429	56	41	23
12	Maine	130	55	35	5
13	New Hampshire	96	55	38	22
14	Kansas	648	55	41	20
15	Iowa	976	54	36	18
16	Missouri	994	51	38	17
17	Arkansas	529	51	33	14
18	Pennsylvania	1742	50	34	15
19	Colorado	435	48	35	20
20	Nevada	123	48	36	22
21	Minnesota	895	46	31	14
22	North Carolina	733	46	30	12
23	Massachusetts	242	46	29	8
24	Virginia	590	45	34	16
25	Kentucky	520	45	27	13
26	Idaho	215	45	27	10
27	Wisconsin	761	44	28	12
28	South Carolina	392	44	27	12
29	New York	1181	43	27	10
30	Alabama	574	43	30	13
31	Texas	1714	43	29	14
32	Oregon	365	43	32	19
33	Delaware	75	43	31	15
34	Utah	318	42	28	10
35	Washington	610	42	30	15
36	Indiana	677	42	25	11
37	Ohio	1198	40	26	11
38	Illinois	1360	40	26	10
39	Michigan	684	39	26	10
40	Maryland	507	38	32	18
41	Rhode Island	34	38	35	18
42	Louisiana	472	38	22	8
43	California	1466	38	26	15
44	Georgia	622	38	23	9
45	Tennessee	427	38	22	8
46	Mississippi	360	36	20	8
47	Hawaii	150	35	21	8
48	Connecticut	142	35	20	4
49	Florida	908	31	19	8
50	New Jersey	536	25	17	9
	U.S. Total	28547	46	31	15

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.

Table 7 shows states ranked on the percent of places in the state with absolute *numeric* errors of 5 or more young children. Data for 10 or more and 25 young children or more are also shown in the table. There is a lot of variation among the states. For example, 69 percent of places in Massachusetts have absoltue *numeric* errors of 5 or more young children compared to12 percent of North Dakota.

Table 7. States Ranked by Percent of Places in State with Absolute Numeric Errors

Rank*	State	Number of Places in State	Errors of 5+ young children	Errors of 10+ young children	Errors of 25+ young children
1	Massachusetts	242	69	43	7
2	Hawaii	150	65	37	9
3	Maine	130	65	29	6
4	Connecticut	142	64	35	9
5	California	1,465	58	28	4
6	New Hampshire	96	58	26	4
7	Florida	908	55	30	5
8	Rhode Island	34	53	18	0
9	Virginia	590	51	25	7
10	Maryland	507	49	22	6
11	Vermont	116	49	20	1
12	Washington	610	49	23	3
13	New Jersey	536	49	22	3
14	Arizona	429	48	23	3
15	New York	1,181	48	23	3
16	Utah	318	46	17	2
17	Nevada	123	46	20	4
18	Texas	1,714	44	14	2
19	South Carolina	392	43	19	3
20	Michigan	684	43	14	2
21	Delaware	75	43	19	8
22	North Carolina	733	42	15	1
23	Louisiana	472	42	14	3
24	New Mexico	419	40	16	1
25	Pennsylvania	1,742	40	18	2
26	Oregon	365	39	14	2
27	Tennessee	427	38	12	1
28	Georgia	622	38	14	1
29	Colorado	435	37	14	2
30	Ohio	1,198	36	11	1
31	Wisconsin	761	34	9	1
32	Alabama	574	34	11	1
33	West Virginia	397	34	14	1
34	Mississippi	360	33	11	1
35	Kentucky	520	33	9	1
36	Montana	341	32	12	1
37	Illinois	1,360	31	9	1
38	Oklahoma	708	31	7	0
39	Indiana	677	31	9	1
40	Alaska	310	27	9	2
41	Wyoming	184	27	7	0
42	Idaho	215	26	6	1
43	Minnesota	895	25	5	0
44	Missouri	994	25	6	0
45	Arkansas	529	22	4	0
46	South Dakota	353	20	6	0
47	Nebraska	543	20	3	0
48	Kansas	648	19	4	0
49	Iowa	976	18	2	0
50	North Dakota	346	12	2	0
	U.S. Total	28,546	38	14	2

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

DC is not included in the state data but is included in the county data

Discussion

It is clear that the introduction of DP into the 2020 Census has caused a lot of controversy. I have been following the U.S. Census since 1970, and I do not remember any issue that has caused as much discussion, concern, and debate among data users as the decision to implement DP in the 2020 Census.

Below I review a couple of issues regarding DP that were not addressed in my analysis but may impact stakeholder's view of DP

Block-Level Data

Blocks are the smallest geographic unit used in the Census and there are about 8 million blocks in the 2020 Census but only about 6 million are occupied. The average block has a total population of about 41 people and about 3 young children. The small population size of blocks makes them susceptible to large percent errors when random numbers are injected with DP.

Assessment of Census accuracy using the two standard Census Bureau methods (Demographic Analysis and Post-Enumeration Survey) are not available at the sub-state level. But the DP Demonstration Product allows one to look at errors attributable to DP for all levels of Census geography down to the census block level and there are some very troublesome issues regarding the use of DP at the census block level.

There are two broad perspectives on the error DP injects into census blocks. One perspective is that data for census blocks are among the most important data supplied by the Decennial Census, and they need to be as accurate as possible. One of

the primary purposes of the Decennial Census is to provide comparable population figures for small areas across the country. To the best of my knowledge, there is no other data source that provides demographic data for all the blocks in the country other than the Decennial Census. Consequently, census accuracy for blocks is especially important. O'Hara (2022) makes a strong case for why block level data are important in terms of creating special or custom districts. The need for such data is often not apparent until well after the Census data has been collected and reported.

Another perspective holds that blocks are typically aggregated into larger units like congressional districts, cities, and counties and in those aggregations the random error injected into individual blocks cancel each other out and produce relatively accurate data for larger units. From this perspective, errors at the block level are not so important.

Regarding the usability of block level data, the Census Bureau (Devine 2022, slide 17) recently stated, "Block-level data are fit-for-use when aggregated into geographically contiguous larger entities. They are not intended to be fit-for-use as a unit of analysis."

I do not think there is any dispute that the error injected by DP for blocks produces a relatively high absolute *percent* error and that these errors typically cancel each other out when blocks are aggregated into larger areas. Because the error is random, the amount of error does not become cumulative. It is an open question about how important census block level data are for making decisions.

One problem with use of DP for small areas is the implausible or impossible results produced. For example, more than 163,000 blocks have children (population

age 0 to 17) but no adults (population age 18 and over) after DP is applied compared to just 82 such blocks before DP was applied (U.S. Census Bureau 2022d). Many such cases are highly unlikely and raise questions about who these children are living with if there are no adults in their household. The Census Bureau (2022d) offers several other examples of implausible or impossible results in the data after DP is applied.

It is not clear to me exactly what statistical problems might be caused by these results, but they undermine the veracity of the census data broadly. A high number of improbable results is identified as a problem of "legitimacy" rather than statistical accuracy by Hogan (2021) and is likely to undermine the confidence the public has in the Census results. When data users see highly implausible results like the large number of blocks with children and no adults, they often wonder what other errors are in the data that are not so apparent.

Despite the statement by the Census Bureau about using block-level data and misgivings among some demographers about the quality of census block data, many data users routinely use the block level data, either because they do not realize the level of potential errors, or because it is the best (or only) data they have at that level of geography.

The data indicate the average percent errors for census blocks is relatively high but does not address how often block-level data are used in decision-making. Readers may have their own answer to that question.

Breaking the Link Between Child and Parents

The production of many blocks where there are children, but no adults may be related to the link between children and adults in a household that is broken when 2020 Disclosure Avoidance System(DAS) with Differential Privacy (DP) was applied to the DHC file. DP is administered to children (population age 0 to 17) and parents (population age 18 and over) independently, so it may eliminate the adults in a household that has children by randomly subtracting data from the number of adults. If the processing retained the link between young children and their parents in a household, it is doubtful that there would be such a high number of blocks with children and no adults.

This statistical disconnection of children and parents is an on-going concern and is likely to have important impacts in later Census products which have more detailed data on young children.[3] For example the connection between children and parents is critical for a lot of data from the American Community Survey. Child poverty is probably the single most important measure of child well-being and determining poverty status requires linking a child to the income of the adults in the households.

The Census Bureau says it will use a different method of DP in the Detailed Demographic and Housing File which will retain the connection between children and parents. Hopefully, that will alleviate concerns. But data that links children and adults in the Detailed Demographic and Housing file will not be available until late 2023 or 2024. That is getting very close to the date (2025) the Census Bureau said it might start applying DP to the American Community Survey (ACS). Translating the application of

[3] It is my understanding that the use of DP does not necessarily require the disconnect between children and parents in a household. The break between children and parents in the redistricting file and the DHC is a result of the particular DP-related processing chosen by the Census Bureau.

DP from the Census to the ACS, is likely to be a difficult process because the ACS is a sample survey rather than a census and the ACS measures more than 40 topics.

Accuracy and Equity

The focus of this report is on census accuracy, but the differential accuracy revealed in my analysis raises the issue of equity. Equity in terms of data provision has become a more visible aspect of data collection and reporting in the federal government recently (White House Equitable Data working Group 2022). According to the U.S. Census Bureau (2021e, pages 1) " The Census Bureau has an ongoing commitment to producing data that depict an accurate portrait of America, including its underserved communities." Data equity has become a part of broader equity questions. This suggests all results should be examined through the lens of equitable data.

In terms of equity, Table 3 shows substantial differential accuracy for Unified School Districts by race and ethnicity after DP is applied in terms of absolute *percent* errors. For Hispanic young children, the mean absolute *percent* error was 28, for Black young children the mean absolute *percent* error was 35, and for Asian young children was 45, compared to 5 for non-Hispanic white children. What does this say about the equity of using the DP method?

There is already differential accuracy in census results before DP is applied but it may be the case that DP exacerbates such inequities. Is it fair to inject as much error for groups that already have a lot of error in census data as for those groups that do not

have much error? Did the Census Bureau examine equity concerns when they decided to use DP in the 2020 Census?

Selection of a Disclosure Avoidance System and Public Trust

Disclosure avoidance is not just a statistical issue and examining it only from a statistical perspective may be problematic. Another dimension for assessing alternative DAS methods is the extent to which a given DAS method undermines public trust in the Census data and the Census Bureau itself. There has been a great deal of concern about the erosion of public trust in the Census Bureau recently. According to the National Academy of Sciences, Engineering and Medicine panel assessing the 2020 Census (2022, page 6),

"We are very concerned, based on presentations to the panel and our knowledge of reactions to previous demonstration data, that the Census Bureau's adoption of differential privacy-based disclosure avoidance has increased the level of public mistrust in the 2020 Census and the Census Bureau itself."

A recent statement from the Federal States Cooperative Program for Population Estimate (FSCPE 2022) states, "Differentially private algorithms have appropriate applications, but they are not a panacea. The evidence and experience to date indicate that they are not capable of handling the complexity of the nation's political and statistical geography and hence do not provide usable data for key constituents."

In their review of the impact DP has had on the Census Bureau credibility and trust among data users, Boyd and Sarathy (2022, page 1) state, "We argue that

rebuilding trust will require more than technical repairs or improved communication: it will require reconstructing what we identify as a “statistical imaginary.”

Summary

This report provides information on the accuracy of DP-infused data and provides a profile of the likely errors for young children that will be seen in data for in the 2020 Census if the Census Bureau uses the privacy protection parameters reflected in the August 2022 Demonstration Product.

It is important to note that the analysis provided in this paper is just a sample of analyses that could be done. But I believe the data analyzed in this study a relatively good sample of the broader implications of using a DAS method with DP in the Demographic and Housing Characteristics file with the privacy protection parameters used in this Demonstration Product.

The question that is not addressed in the previous sections is whether the level of error reflected in this analysis would make 2020 Census for data on young children “unfit for use.” Each person will probably have a different answer to how much error in census data for young children is too much error.

Like all disclosure avoidance systems, the use of DP involves a trade-off between privacy protection and census accuracy. There have always been errors in the Census data, but in the 2020 Census, the Census Bureau is trying to decide how much additional error to add to the data in order to enhance privacy protection. By setting privacy parameters, the Census Bureau has control over the level of accuracy and level of privacy protection in the 2020 Census.

Given this balancing act, it would be useful to have more information about metrics on privacy protection. It would be helpful if we could compare the metrics of accuracy like those in this report to metrics of privacy protection in the August 2022 Demonstration Product. I see many measures of accuracy based on the Demonstration Product. However, I do not see any privacy protection metrics produced by the Census Bureau nor do I see a way to explore the privacy protection aspect with the Demonstration Product. It seems the balance of accuracy and privacy protection is the key reason for using a given disclosure avoidance system but without metrics for privacy protection I am not sure how to do that. When I have asked experts about the level of privacy protection afforded by an Epsilon of 19.6 in the redistricting data in terms I can understand it seems like I always get a variation of "it depends." But no metrics.

On the other hand, the problems that are likely to be caused by inaccurate census data on young children are clearer to me. The data in this paper, and many other analyses, provide a rich set of metrics showing the magnitude of error DP injects into Census data and I can envision problems such errors might cause.

When the number of young children in a school district is under-reported by 5 or 10 percent, that could have big implications for their funding and when the number of young children in a community is off by 10 percent or more, that could impact planning in ways that waste taxpayer money and undermine quality education for young children. If the number of young children reported in the Census for a Unified School District is 10 percent too low, it may not automatically translate into 10 percent less money for that

jurisdiction. But there is a strong link between underreporting the number of young children and the loss of money in a general sense.

In addition to the money distributed on the basis of census-derived data, Census data are used for many decisions in the public and private sector. The more errors there are in the data and the larger the errors in the data, the less likely those decisions will be correct ones.

Given the level of errors in Unified School Districts and Places using the privacy protection level in the most recent DP Demonstration Product, and the lack of clear evidence or measurements about the level or impact of privacy loss, I recommend that the Census Bureau increase the level of accuracy used in the DHC to provide more accurate small area data for young children. And reduce or eliminate large errors caused by application of DP.

References

Bouk, D. and Boyd, D. (2021*). Democracy's Data Infrastructure.; The technologies of the U.S. Census*. https://knightcolumbia.org/content/democracys-data-infrastructure

Boyd. D. (2019). "Balancing Data Utility and Confidentiality on the 2020 US Census," Data and Society, https://datasociety.net/library/balancing-data-utility-and-confidentiality-in-the-2020-us-census/ .

Boyd, D. and Sarathy, J. (2022) "Differential Perspectives: Epistemic Disconnects Surrounding the US Census Burau's Use of Differential Privacy," *Harvard Data Science Review* (forthcoming)

Committee on National Statistics (2019). "Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations," presentations are available at https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

Cropper, M., McKibben, J, and Stojakovic, Z. (2021). The Importance of Small Area Census Data for School Demographics, Count all Kid website https://ednote.ecs.org/counting-all-kids-how-the-census-impacts-education/

Federal State Cooperative Population Estimates (FSCPE) (2022). "Letter to Census Bureau Director Robert Santos," https://docs.google.com/forms/d/e/1FAIpQLScU7bK9yIAy9YV-WIVjIJhx-b05-IB2el8M47Cg1jZu3Sa5tA/viewform

Hogan, H. (2021). "The History of Assessing Census Quality, Presentation at 2021 Population of Association of America Conference, May 5, 2021.

Hogan, H. (2021). "The History of Assessing Census Quality, Presentation at 2021 Population of Association of America Conference, May 5, 2021.

Hotz, J. and Salvo J. (2020). Addressing the Use of Differential Privacy for the 2020 Census: Summary of What We Learned from the CNSTAT Workshop. https://www.apdu.org/2020/02/28/apdu-member-post-assessing-the-use-of-differential-privacy-for-the-2020-census-summary-of-what-we-learned-from-the-cnstat-workshop/ ,

McElrath, K. Bauman, K., and Schmidt, E (2022) Preschool Enrollment in the United states,: 2055 to 2019," U.S. Census Bureau https://www.census.gov/content/dam/Census/newsroom/press-kits/2021/paa/paa-2021-presentation-preschool-enrollment-in-the-united-states.pdf

Nagle, N. and Kuhn, T. (2019). “Implications for School Enrollment Statistics.” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations.

National Academy of Sciences, Engineering and Medicine, (2022). *Understanding the Quality of the 2020 Census, Interim Report* , Washington Dc. The National Academy Press, https://nap.nationalacademies.org/catalog/26529/understanding-the-quality-of-the-2020-census-interim-report

O’Hara, A. (2022) presentation at Analysis of Census Noise Measurements Workshop, April 28-29, Rutgers University.

O’Hare, W.P. (2019). “Assessing 2010 Census Data with Differential Privacy for Young Children,”” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

O’Hare W. P. (2020a). “Many States Use Decennial Census Data to Distribute State Money,” The Census Project Website https://thecensusproject.org/2020/01/09/many-states-use-decennial-census-data-to-distribute-state-money/

O’Hare, W.P (2020b). “Implications of Differential Privacy for Reported Data on Young children in the 2020 U.S. Census,” Posted on Count All KIDS Website Implications-of-Differential-Privacy-for-kids-11-17-2020-FINAL-00000003.pdf (myftpupload.com) .

O’Hare, W.P. (2021). “Analysis of Census Bureau’s August 2021 Differential Privacy Demonstration Product: Implications for Data on Children,” Count All Kids website November *https://countallkids.org/resources/analysis-of-census-bureaus-august-2021-differential-privacy-demonstration-product-implications-for-data-on-children/*

O’Hare, W. P. (2022a). “New Census Bureau Data Show Young Children Have a High Net Undercount in the 2020 Census, “ Posted on Count All Kids website , March, https://countallkids.org/resources/new-census-bureau-data-show-young-children-have-a-high-net-undercount-in-the-2020-census/

O’Hare , W. P. (2022b). “U.se of the American Community Survey Data by State Child Advocacy Organizations.” Count All Kids website, https://countallkids.org/resources/use-of-the-american-community-survey-data-by-state-child-advocacy-organizations/

O’Hare W. P and A. Rashid (2022). Selected Federal Programs that Use Figures for the Population Ages 0 to 5 for Distribution of Federal Funds to States and Localities, Posted on Count All Kids website July 5 https://countallkids.org/selected-federal-programs-that-use-the-population-size-for-ages-0-to-5-for-the-distribution-of-federal-funds-to-states-and-localities/

O'Hare, W.P. (2022c). "Analysis of Census Bureau's March 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children," Posted on the Count All Kids website, https://countallkids.org/resources/analysis-of-census-bureaus-march-2022-differential-privacy-demonstration-product-implications-for-data-on-young-children/

Reamer, A. (2020). Counting for Dollars, George Washington University https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds .+

The Annie E. Casey Foundation (2018). *KID COUNT DATA BOOK 2018*, https://www.aecf.org/resources/2018-kids-count-data-book

U.S. Census Bureau (2018), "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," THE RESEARCH AND METHODOLOGY DIRECTORATE, Mc Kenna, L. U.S. Census Bureau, Washington DC., https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf .

U.S. Census Bureau (2019). "2010 Demonstration Data Products," U.S. Census Bureau, Washington DC., October, https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html .

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S Census Bureau, Washington DC., August18, https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?# .

U.S. Census Bureau (2020b). "2020 Census Data Products and the Disclosure Avoidance System", Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, August26.

U.S. Census Bureau (2020c). "DAS Updates, U.S Census Bureau," Hawes M. June 1 Washington DC., https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?# .

U.S. Census Bureau (2020d). "Disclosure Avoidance and the Census," Select Topics in International Censuses, U.S. Census Bureau, October 2020. https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html .

U.S. Census Bureau (2020e). "Disclosure Avoidance and the 2020 Census, U.S. Census Bureau," Washington DC., Accessed November 2, https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html .

U.S. Census Bureau (2020f). "Error Discovered in PPM," U.S. Census Bureau, Washington DC. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2020g). "2020 Disclosure Avoidance System Updates," U.S. Census Bureau, Washington DC., https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2021a). School Enrollment in the United States: October 2019 - PowerPoint Presentation (census.gov Detailed Tables, School Enrollment in the United States: October 2019 - Detailed Table 1, FEBRUARY 02, 2021.

U.S. Census Bureau (2021b). Developing the DAS: Demonstration Data and Progress Metric, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html .

U.S. Census Bureau (2021c). "Differential Privacy 101." Webinar May 4, 2021, Michael Hawes. https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/differential-privacy-101.html

U.S. Census Bureau (2021d). "Disclosure Avoidance for the 2020 Census: An Introduction," U.S. Census Bureau, Washington, DC. November https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Bureau (2021e). "Advancing Equity with Census Bureau Data." Census Bureau Blog, November 2, 2021, Ron Jarmin , Acting Director Advancing Equity with Census Bureau Data

U.S. Census Bureau (2021f). "Disclosure Avoidance for the 2020 Census: An Introduction," November 2021, U.S. Census Bureau, Washington DC https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Bureau (2022a). "Understanding Disclosure Avoidance- Related Variability in the 2020 Census Redistricting data, " U.S. Census Bureau, Washington DC. January 28. https://www.census.gov/library/fact-sheets/2022/variability.html

U.S. Census Bureau (2022b). "Revised Data Metrics for 2020 Disclosure Avoidance," U.S. Census Bureau, Washington DC.

U.S. Census Bureau (2022c) Post-Enumeration Survey and Demographic Analysis Help Evaluate 2020 Census Results, August10 , Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census

U.S. Census Bureau (2022d). Detailed Summary Metrics , https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-08-25_Summary_File/2022-08-25_Detailed_Summary_Metrics_Overview.pdf

U.S. Census Bureau (2022e) "Just Released: New Demonstration Data for the DHC; webinar August 31,

U.S. Census Bureau (2022f). Summary of Feedback on DHC Demonstration Data JUNE 23, 2022 https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/newsletters/summary-of-feedback-on-dhc-demonstration-data.htm

U.S. Census Burau (2022g) Just Released: New Demonstration Data for the DHC; Webinar August 31, August 25, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/newsletters/New-2010-DHC-Demonstration-Data-Coming-August-25-Webinar-August-31.html

U.S. General Accountability Office (2020). "COVID-19 Presents Delays and Risks to Census Counts," U.S. General Accountability Office, Washington, DC., https://www.gao.gov/products/GAO-20-551R .

Vink, J. (2019). "Elementary School Enrollment," https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

White House Equitable Data Working Group (2022) "A Vision for Equitable Data : Recommendations from the Equitable Data Working Group," https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf