Analysis of Census Bureau's April 2023 Differential Privacy Demonstration Product: Implications for Data on Young Children

By
Dr. William P. O'Hare

May 2023

Analysis of Census Bureau's March 2023 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). This paper supersedes papers I wrote on this topic based on the Demonstration Products released in August 2022 and March 2022 (O'Hare 2022d; O'Hare 2022f) .

This study is based on analysis of the most recent DP Demonstration Product for the Demographic and Housing Characteristics (DHC) file released by the Census Bureau on April 6, 2023. The DP Demonstration Product issued on April 6, 2023 supersedes earlier DP Demonstration Products from the Census Bureau 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 which means it contains a lot of detailed cross-tabulated data from the 2020 Census. The Demonstration Product released in April 2023 has data for population and housing units, but this analysis only examines population data from the 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. 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 data show that 69 percent of Unified School Districts had absolute *numeric* errors of 5 or more young children after DP is applied. Also, the count of young children would exhibit absolute *percent* error of 5 percent or more in about 21 percent of Unified School Districts after DP is applied.

Errors of the magnitude shown above could have important implications for educational planning and for federal and state funding received by schools and. 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 6 percent compared to 28 percent of Hispanic young children, 37 percent for Black young children, and 47 percent for Asian young children. Differential accuracy among race and Hispanic Origin groups raises questions of data equity after DP is applied.

I believe the most important type of error introduced by the application of DP are the large errors introduced for some geographic units. In terms of *numeric* errors, 9 percent of Unified School District have absolute *numeric* errors of 25 or more young children. Analysis also shows that 2 percent of Unified School Districts have absolute *percent* errors of 25 percent or more. Two percent of all Unified School District is more than 200 districts.

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 42 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4, and 48 percent had absolute *numeric* errors of 5 or more children age 4.

The results are similar for Places. Analysis shows that 48 percent of Places (cities, village, and towns) had absolute *percent* errors of 5 percent or more for age 0 to 4, and 43 percent of Places had absolute *numeric* errors of 5 or more young children.

Evidence shows smaller (in population size) places have high absolute *percent* errors. 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 April 2023 Census Bureau Demonstration Product (U.S. Census Bureau 2023c Table 18), there were 162,743 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 parents in data processing is an ongoing concern for data on 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 (Gutierrez 2022; O'Hare 2022b). To understand the well-being of children, it is critical to understand the circumstances 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.

The Census Bureau (2023b, page 10) describes the situation this way,

"Finally, regarding (iv) the Disclosure Avoidance System (DAS) TopDown Algorithm does break the connection between people and households, however, we recognize the importance of accurate data on children and families, so we have included some ''lagged variable' tables that restructure the tables so users can still get those links as count queries. It is important to note that differential privacy itself doesn't eliminate these links. The DHC data uses the TopDown Algorithm so those links are broken, but the Supplemental DHC (S-DHC) uses a different implementation of differential privacy that preserves links between children an n families."

It was reassuring when the Census Bureau issued a note in December 2022, saying it would not use DP on the ACS until the science is ready. The Bureau (U. S. Census Bureau (2022h) stated, "Our current assessment is that the science does not yet exist to comprehensively implement a formally private solution for ACS," In addition they stated, "Strengthening confidentiality protections for the ACS is a continuous process, and we are taking the time to carefully research options and engage with the data user community."

The first data from the 2020 Census showing relationships between children and the adults in the household will be the Supplemental-Demographic and Housing Characteristics (S-DHC) file. The Census Bureau will use a variant of DP called PHSafe on the S-DHC file. Examination of errors for young children introduced by DP will help us gain a better understanding of the possible impact of DP on ACS data.

To be clear, the Census Bureau has set the parameters for production of the DHC file, and this paper will not make a difference in those parameters. But there are

three reasons for producing this paper. First, I hope the results in this paper will provide readers with some guidance about the likely errors in the DHC data for young children. This paper is meant to provide stakeholders, data users, 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 in the DHC file. 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.

Second, the U.S. Census Bureau is still 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. The Census Bureau will be releasing three more files from the 2020 Census later this year or early next year. Feedback from data users may help shape the way DP is applied to those files. 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.**

Third, I also hope this paper may be of some use when the DP issue is re-visited in the context of the 2030 Census and/or major Census Bureau data collection efforts such as the American Community Survey. Thus, the third reason for posting this paper is to help build a record regarding the impact of DP on young children. It is very likely that the question about applying DP to the 2030 Census will arise in a few years in the context of the 2030 Census. In addition, there is an ongoing question about using DP

on Census Bureau surveys. Having a record of the impact of DP on data for young children from the 2020 Census will be helpful as future use of DP is considered.

Analysis of Census Bureau's April 6, 2023 Differential Privacy Demonstration Product for DHC: 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 (in Census Bureau terminology, children less than one year old are referred to as age 0) The analysis is based on the Demonstration Product released by the Census Bureau on April 6, 2023. This DP Demonstration Product file is based on the production-level parameters for DP used in the Demographic and Housing Characteristics (DHC) file.

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. Ruggles and Van Riper (2022) offers another view on the use of DP by the Census Bureau A good overview of the evolution of the DP issue at the Census Bureau is provided by Boyd and Sarathy (2022). Several papers have examined the impact of DP on census data (Swanson and Cossman (2021), Winkler et al; 2022). The Harvard Data Science Review (Special Issue II I 2022) devoted a complete issue to Differential Privacy

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 U.S. Decennial 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 on the undercount of young children in the Census are also unsettling. From 1950 to 1980, the young children and adults had similar undercount

rates and 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 net 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 or eliminated for this group. It does not seem fair to inject more error into data for groups that already have a high level 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.

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 offered by the Census Bureau.

The DP demonstration file released by the Census Bureau on April 6, 2023, 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 2020 Census 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 value in the Summary File to calculate the error. For percentages, the difference is divided by the value in the Summary File.

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 liable 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 are 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 April 2023 is part of a series of such products the Census Bureau has released. Starting in October 2019, the Census Bureau has released several Demonstration Products that reflect the injection of DP into 2010 Census data to assess the implications of DP. 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 (DHC) file that is scheduled to be released in May 2023. According to IPUMS-NHGIS (2023, no page), "The privacy loss budget (epsilon) assigned to the person-level and housing unit level counts in the 2023-04-03 vintage was 26.43 and 34.33 respectively." If I understand epsilon correctly, this means the data from the DHC files offer somewhat more accuracy and somewhat less privacy protection than the PL 94-171 file in which the epsilon was 19.6.

The data used in my analysis were originally released by the Census Bureau. IPUMS- NHGIS (National Historic Geographic Information Systems) 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 my analysis. Observations with zeros for key measures produce very unusual and unusable results. There were very few units with zero young children in the overall analysis, so this had very little impact on the analysis. On the other hand, for the analysis by race and Hispanic Origin shown in Table 3 many units were removed from analysis because they did not have any young children in the specific race/Hispanic origin group. 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 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 11 young children and mean absolute *percent* error is 1.

However, of the 3,142 counties examined here 36 percent (1,138) 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 1,138 counties with less than 1,000 young children, the mean absolute *numeric* error for ages 0 to 4 was 8 and the mean absolute *percent* error was 3. For 197 counties with less than 200 young children, the mean absolute *numeric* error was 7 and the mean absolute *percent* error was 7.

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

	States****	Counties***	School Districts	Places
Number of Units in the Analysis	50	3,142	10,862	28,526
Average Size of District (Children ages 0-4 based on Summary File)	403,375	6,429	1,860	545
Average Absolute Numeric Error**	100	11	11	6
Average Absolute Percent Error	rounds to zero	1	4	14
Percent of Units with Absolute Numeric Errors of 5 or more	98	72	69	43
Percent of Units with Absolute Percent Errors of 5% or more	rounds to zero	5	21	48

Source: Author's analysis of Demonstration Product data released by the Census Bureau on April 6, 2023. Data are taken from website at IPUMS NHGIS, University of Minnesota www.nhgis.org

* 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.

*** includes county equivalents and one county removed because zero young children

**** state data was not put up on the IPUMS website for this demonstration product because DP has very little impact on state data. The data shown here for states is from the demonstration produce released in August 2022.

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

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 file

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 11 young

children and it is 6 young children for Places. The mean absolute *percent* error for United School Districts is 4 percent and it is 14 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 69 percent of Unified School District have absolute *numeric* errors of 5 or more children and 21 percent have absolute *percent* errors of 5 percent or more. For Places, 43 percent have absolute *numeric* errors of 5 or more children, and 48 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 will be 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 one reason 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 316 programs use substate data which underscores the importance of small area census data (Reamer 2019). When one is talking about billions of dollars, a small percent error can translate into a large dollar amount.

Many of the funding formulas that use census-derived data are complicated. For example, Gordon and Reber (2023) provide an in-depth analysis of how Census-derived data are used in Title I of the Elementary and Secondary Act of 1965.

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).

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	

At the National Academy of Sciences, Committee on National Statistics workshop on DP (Committee on National Statistics 2019) which was held in December 2019 there were several presentations reflecting implications of DP-infused data for 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.

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. Of the 10,862 Unified School Districts in this analysis; 7,473 (69 percent of all Unified School Districts) had a young

child population of less than 1,000, and 1,452 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 10,862 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 Non-Hispanic White, Black or Asian, it means Non-Hispanic White alone Black alone or Asian alone. Other race groups were not examined here because the numbers were small, they were often highly clustered.

Table 3. 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	10,862	10,840	10,240	7,577	6,188
Average number of young children in district (in group column hea	1,860	946	499	383	145
Average absolute numeric error***	11	10	9	7	5
Average absolute percent error	4	6	28	37	47
Percent of units with errors of 5 or more young children	69%	68%	54%	44%	35%
Percent of units with errors of 5% or more	21%	27%	69%	66%	72%

Source: Author's analysis of Demonstration Product data released by the Census Bureau on April 6, 2023, 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.

** these are black alone and Asian alone

*** 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

[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 Non-Hispanic white child. Black child, one Hispanic child, and one Asian child. But many districts have relatively few young children of color. The average number of Non-Hispanic white young children was 946, the average number of Hispanic young children in Unified School Districts where there was at least one Hispanic was 499, for Blacks it was 383 and for Asians it was 145 These numbers are well below the overall average of 1,860 young children for all districts. 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 11 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 (10, for Non-Hispanic white, 9 for Hispanic young children,7 for Black young children, and 5 for Asian young children), compared to 11 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, 6 for Non-Hispanic white, 28 percent for Hispanic, 37 percent for Blacks young children, and 47 percent for Asian young children (Table 3).

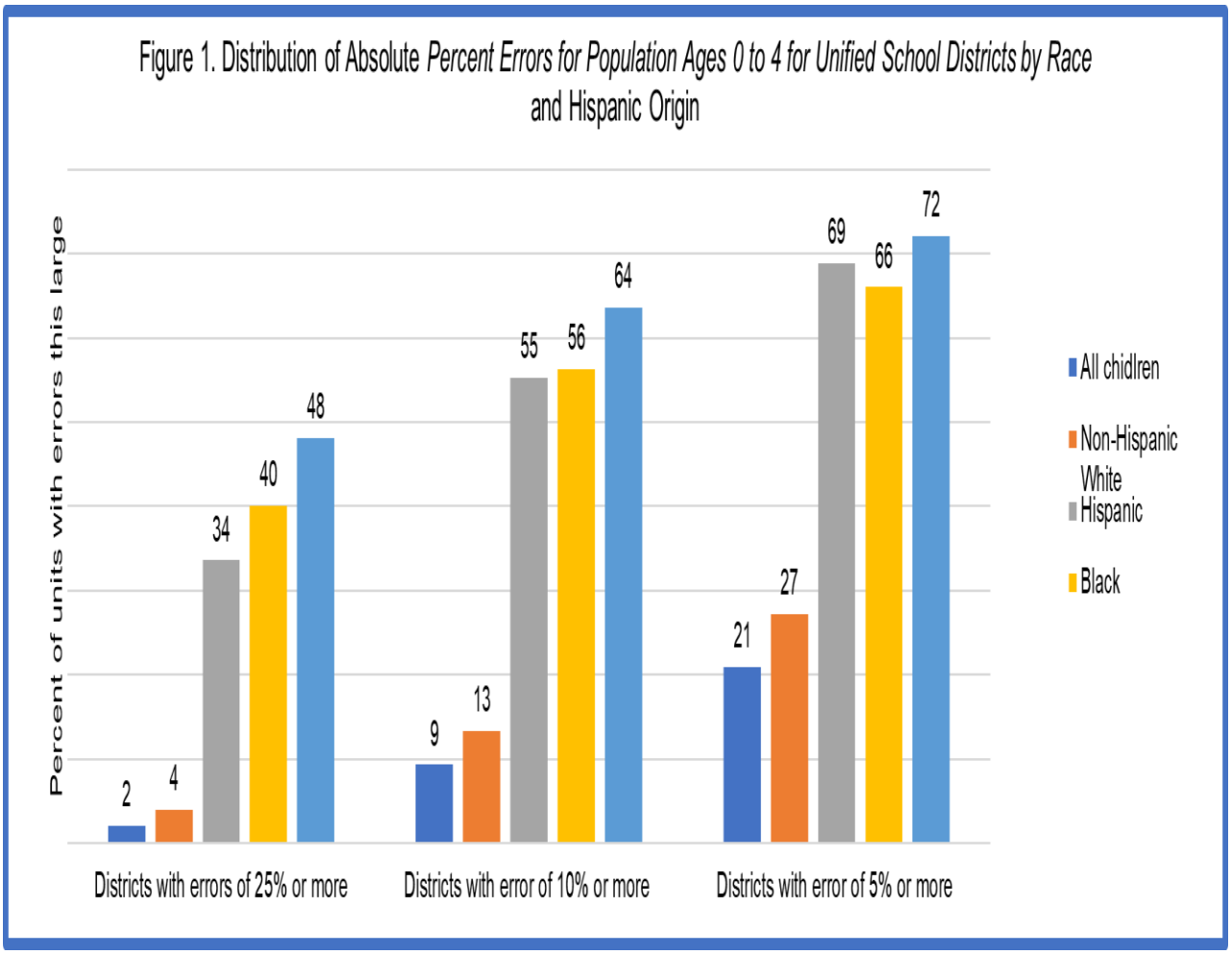
Large Errors in Unified School Districts

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 Census Bureau 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 large 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, 21 percent of districts had absolute *percent* errors of 5 percent or more, compared to 27 percent of Non-Hispanic White Alone, 69 percent for Hispanic young children, 66 percent for Black young children, and 72 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.

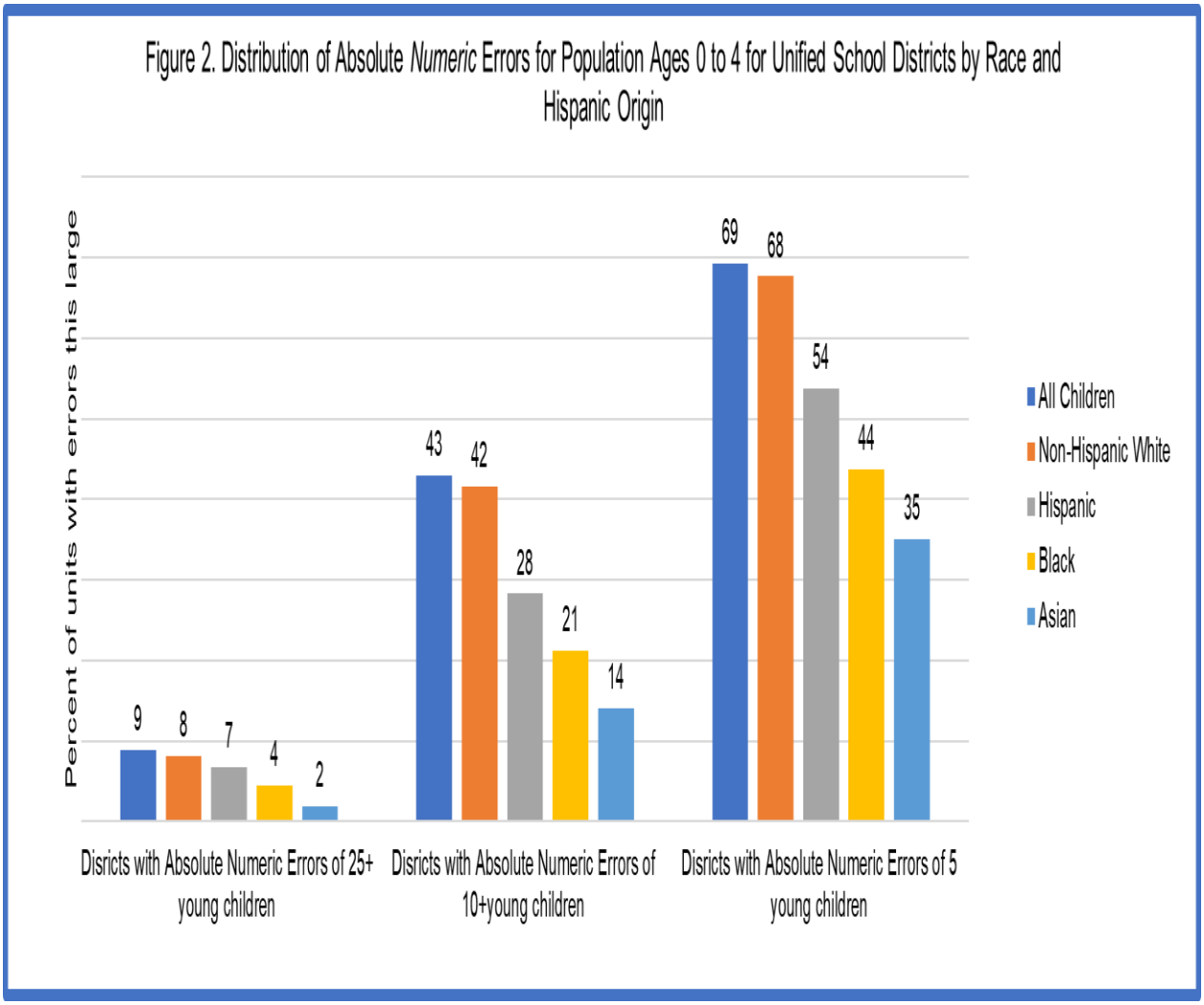


In the largest absolute *percent* error category (25 percent or more) the numbers are quite low; 2 percent for all young children and 4 percent for non-Hispanic whites alone young children, but quite high for Black, Hispanic, and Asian young children. Figure 1 shows that 34 percent of Unified School Districts have absolute *percent* errors of 25 percent or more for Hispanics, compared to 40 percent for Blacks and 48 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 amounts to 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 69 percent of the Unified School Districts had errors of 5 young children or more for young children of all races and 68 percent of Non-Hispanic white but the figures for racial and Hispanic minority groups are smaller: 54 percent for Hispanic young children, 44 percent for Black young children, and 35 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.

There are relatively few Unified School Districts with very large absolute *numeric* errors. Only 9 percent of Unified School Districts have errors of 25 young children or

more, compared to 7 percent of Hispanic young children, 4 percent for Black young children, and 2 percent for Asian young children.

A few examples might make the situation more understandable. In one school district the 2010 Census reported 37 children age 0 to 4, but after DP was applied the number was only 20. In another Unified School District, the 2010 Census reported 83 children age 0 to 4, but after DP was applied the number was only 56. It is difficult for me to believe that errors of this magnitude will not be problematic, and these types of situations are likely to be experienced by many Unified School Districts in the 2020 Census. For those districts that have errors of 25 percent or more because of the application of DP, the results are likely to be a significant problem. Nine percent of all Unified School Districts is about 900 districts.

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. For the average absolute *numeric* error, the highest state was Hawaii at 72.0 and the lowest state was Vermont at 3.8. For average absolute *percent* error, the highest state was Vermont at 15 and the lowest state was Hawaii at 0.1.

Table 4 States Ranked by Mean Absolute Numeric Error and Absolute Percent Error for Ages 0 to 4 by Unified School Districts.						
Rank*	State	Average absolute numeric error		Rank*		Average absolute percent error
1	HAWAII	72.0		1	VERMONT	15.0
2	CALIFORNIA	21.0		2	MAINE	12.7
3	ARIZONA	16.7		3	IDAHO	9.6
4	MARYLAND	14.8		4	WASHINGTON	9.3
5	SOUTH CAROLINA	14.3		5	NORTH DAKOTA	8.8
6	DELAWARE	14.3		6	MONTANA	8.7
7	NEW YORK	13.5		7	OREGON	8.0
8	UTAH	13.2		8	ALASKA	7.7
9	MICHIGAN	13.2		9	NEBRASKA	7.4
10	TENNESSEE	13.2		10	SOUTH DAKOTA	7.1
11	FLORIDA	12.6		11	NEW MEXICO	6.7
12	TEXAS	12.6		12	COLORADO	5.9
13	NORTH CAROLINA	12.6		13	TEXAS	5.8
14	WASHINGTON	12.5		14	OKLAHOMA	5.8
15	ILLINOIS	12.4		15	KANSAS	5.0
16	ARKANSAS	12.4		16	MISSOURI	4.6
17	GEORGIA	11.5		17	IOWA	4.6
18	MISSISSIPPI	11.4		18	NEW YORK	4.2
19	MINNESOTA	11.4		19	WYOMING	3.9
20	KENTUCKY	11.3		20	MINNESOTA	3.7
21	MISSOURI	10.9		21	ARKANSAS	3.6
22	WISCONSIN	10.8		22	ILLINOIS	3.4
23	NEVADA	10.8		23	NEW HAMPSHIRE	3.3
24	OHIO	10.7		24	INDIANA	3.3
25	OREGON	10.6		25	WISCONSIN	3.2
26	NEW MEXICO	10.5		26	ARIZONA	2.6
27	COLORADO	10.5		27	MICHIGAN	2.6
28	ALABAMA	10.3		28	OHIO	2.3
29	LOUISIANA	10.3		29	CALIFORNIA	2.2
30	IDAHO	10.1		30	NEVADA	2.2
31	OKLAHOMA	10.0		31	KENTUCKY	1.6
32	INDIANA	9.8		32	TENNESSEE	1.6
33	IOWA	9.7		33	NEW JERSEY	1.5
34	WEST VIRGINIA	9.5		34	UTAH	1.5
35	PENNSYLVANIA	9.2		35	MISSISSIPPI	1.4
36	NEW JERSEY	9.2		36	PENNSYLVANIA	1.2
37	NEBRASKA	9.0		37	MASSACHUSETTS	1.2
38	VIRGINIA	8.8		38	RHODE ISLAND	1.2
39	WYOMING	8.7		39	ALABAMA	1.0
40	RHODE ISLAND	8.4		40	GEORGIA	1.0
41	KANSAS	8.1		41	SOUTH CAROLINA	0.9
42	MASSACHUSETTS	8.1		42	VIRGINIA	0.9
43	CONNECTICUT	8.1		43	WEST VIRGINIA	0.9
44	SOUTH DAKOTA	8.0		44	CONNECTICUT	0.7
45	ALASKA	7.2		45	DELAWARE	0.7
46	NORTH DAKOTA	6.5		46	NORTH CAROLINA	0.6
47	MONTANA	5.7		47	LOUISIANA	0.6
48	NEW HAMPSHIRE	5.7		48	FLORIDA	0.4
49	MAINE	4.7		49	MARYLAND	0.3
50	VERMONT	3.8		50	HAWAII	0.1
	U.S. Average	11.1			U.S. Average	4.4

Source: Authors analysis of Demonstration Produce released by the Census Bureau on April 6. 2023

*Ranks are based on unrounded data

Analysis for Age 4

In the Demonstration Product released in April 2023, the Census Bureau provided data by single year of age 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 error 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. There were more than 100 school districts not analyzed because they have zero children age 4 in the Summary File or DP file.

The mean absolute *numeric* error was 6 and the mean absolute *percen*t error was 9 percent for age 4.

Table 5. Unified School District Error* Metrics for Age 4	
Number of Units in Analysis	10,775
Average number of 4 year old's in Summary File	377
Average Absolute *Numeric* Error	6
Average Absolute *Percent* Error	9
Percent of units with Absolute *Numeric* error 5+ children age 4	48
Percent of units with Absolute *Percent* error 5%+	42
Source: Author's analysis of Demonstration Product released by the Census Bureau on April 6, 2023, 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 0 to 4 in 2010 Summery File or DP-Infused file.	

A substantial share of Unified School Districts had large errors in both *numeric* and *percent* terms. About half (48 percent) of Unified School System had absolute *numeric* errors of 5 or more children and 42 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4.

With errors of this magnitude for a 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.

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. There were 27,963 places with adequate data for my analysis.

Table 1 shows the mean absolute *numeric* error for Places was 6 and the mean absolute *percent* error was 14 percent. The high percent error is not surprising because many of these Places are small. There were 15,906 Places where the number of young children was less than 100, and 23,209 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 (48 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.

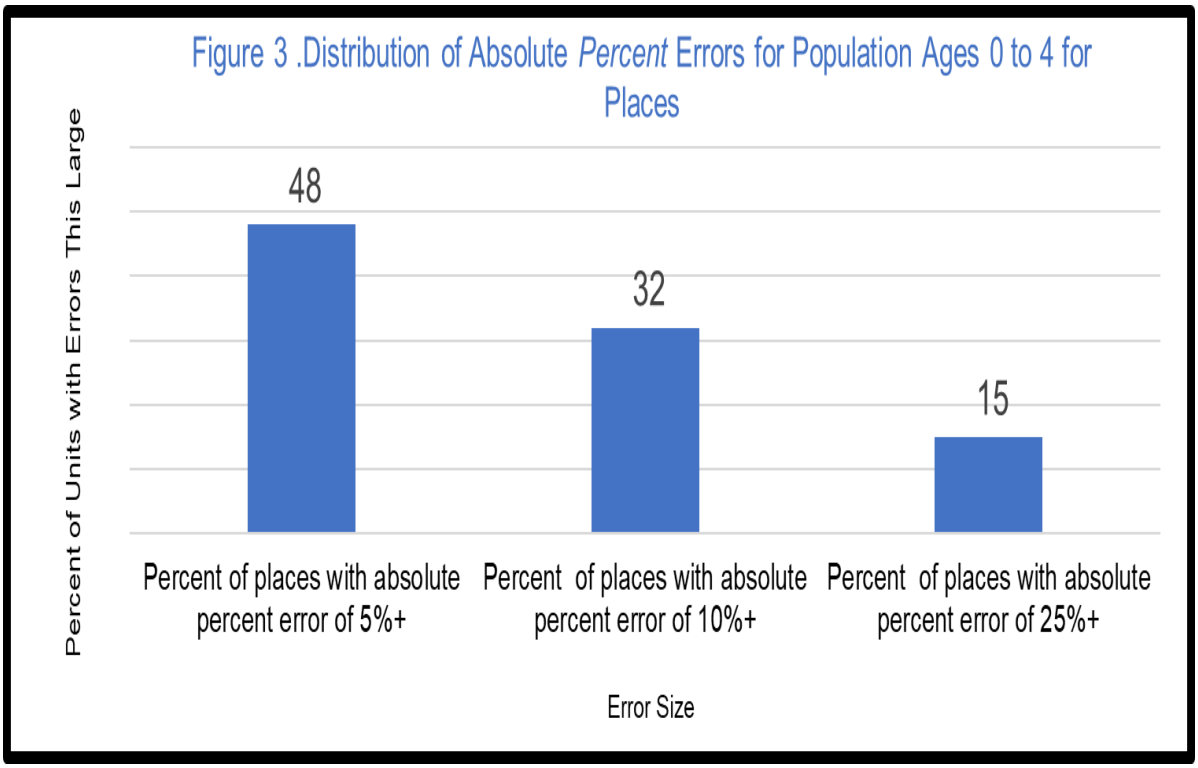


Figure 4 shows the distribution of Places by absolute *numeric* errors using the same categories as Figure 2. Data show 43 percent of the Places had absolute *numeric* errors of 5 or more young children, and only 2 percent had absolute *percent* errors of 25 or more young children.

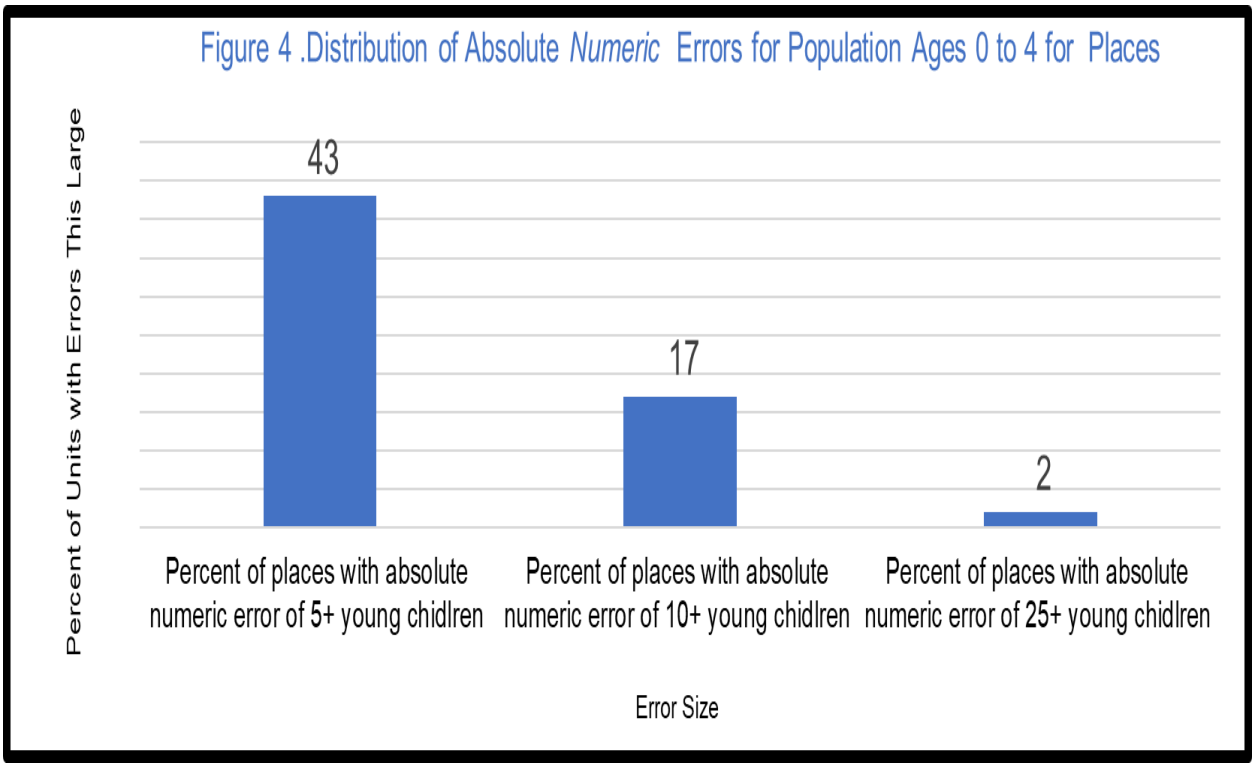


Table 6 shows states ranked on the percent of places in a state with absolute *percent* errors of 5 percent or more. Data for errors of 10 percent or more and 25 percent or more are also provided in the Table 6.

Table 6. States Ranked by Percent of Places in State with Absolute Percent Errsors of 5 percent or more population ages 0 to 4					
			Percent Distribution with state		
Rank	State	Number of Places in State	Absolute Percent error of 5%+	Absolute Percent error of 10%+	Absolute percent error of 25%+
1	VERMONT	118	72	57	30
2	MONTANA	336	67	56	27
3	NORTH DAKOTA	345	67	56	25
4	NEW MEXICO	413	63	49	21
5	ALASKA	308	62	46	18
6	NEBRASKA	541	60	45	19
7	WYOMING	185	60	44	17
8	WEST VIRGINIA	394	60	43	23
9	KANSAS	646	59	45	21
10	SOUTH DAKOTA	356	59	45	19
11	OKLAHOMA	706	59	44	21
12	NEW HAMPSHIRE	96	58	42	24
13	ARIZONA	427	58	41	19
14	NEVADA	122	57	42	19
15	IOWA	980	56	40	21
16	COLORADO	434	54	39	18
17	ARKANSAS	529	53	38	22
18	MISSOURI	992	53	38	19
19	PENNSYLVANIA	1,742	52	34	20
20	IDAHO	215	52	36	26
21	MAINE	130	50	31	22
22	SOUTH CAROLINA	393	48	29	15
23	KENTUCKY	520	48	29	16
24	MINNESOTA	886	48	33	18
25	VIRGINIA	587	48	32	18
26	DELAWARE	76	47	28	12
27	RHODE ISLAND	34	47	41	24
28	ALABAMA	574	47	31	17
29	WASHINGTON	616	46	32	16
30	INDIANA	676	46	29	17
31	NORTH CAROLINA	732	46	31	19
32	TEXAS	1,714	46	30	16
33	OREGON	367	46	34	14
34	NEW YORK	1,180	44	27	17
35	MICHIGAN	686	44	27	15
36	OHIO	1,196	44	28	18
37	WISCONSIN	762	44	27	14
38	ILLINOIS	1,359	43	27	17
39	LOUISIANA	472	43	25	17
40	UTAH	320	43	24	14
41	MARYLAND	511	41	32	13
42	TENNESSEE	427	41	23	14
43	MISSISSIPPI	362	41	24	15
44	GEORGIA	621	40	25	14
45	HAWAII	150	39	22	12
46	CALIFORNIA	1,460	37	27	13
47	MASSACHUSETTS	242	36	18	11
48	CONNECTICUT	141	35	18	11
49	FLORIDA	910	32	18	10
50	NEW JERSEY	536	30	19	11
	U.S. Total	28,527	48	32	17

There is a lot of variation across the states. For example, 72 percent of the Places in Vermont had absolute *percent* errors of 5 percent or more, compared to 30 percent of Places in New Jersey. Table 6 shows a large percentage (10 percent or more in all states) of Places had absolutes *percent* errors of 25 percent or more.

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 7.

There is a lot of variation among the states. For example, 66 percent of places in Massachusetts have absoltue *numeric* errors of 5 or more young children compared to 18 percent of North Dakota.

Table 7. States Ranked by Percent of Places with Absolute Numeric Error of 5 or more children age 0 to 4					
			Distribution within state		
Rank*	State	total	errors 5+	errors 10+	errors 25+
1	MASSACHUSETTS	242	66	38	7
2	HAWAII	150	65	41	5
3	CONNECTICUT	141	62	31	4
4	NEW HAMPSHIRE	96	61	29	3
5	FLORIDA	910	61	35	6
6	CALIFORNIA	1,460	60	33	6
7	MAINE	130	56	31	3
8	VIRGINIA	587	55	28	5
9	NEW JERSEY	536	55	29	4
10	WASHINGTON	616	55	27	6
11	NEW YORK	1,180	54	25	2
12	RHODE ISLAND	34	53	35	12
13	NEVADA	122	52	22	3
14	ARIZONA	427	52	28	4
15	MARYLAND	511	51	29	5
16	MICHIGAN	686	51	18	2
17	SOUTH CAROLINA	393	50	24	3
18	TEXAS	1,714	49	20	3
19	LOUISIANA	472	48	21	2
20	NORTH CAROLINA	732	47	18	1
21	UTAH	320	46	18	2
22	PENNSYLVANIA	1,742	45	16	2
23	COLORADO	434	45	16	3
24	GEORGIA	621	45	17	2
25	VERMONT	118	44	20	1
26	DELAWARE	76	42	21	3
27	OHIO	1,196	42	14	1
28	OREGON	367	42	16	3
29	TENNESSEE	427	42	15	0
30	NEW MEXICO	413	42	17	2
31	ALABAMA	574	41	14	1
32	MISSISSIPPI	362	40	15	0
33	INDIANA	676	39	14	1
34	WEST VIRGINIA	394	39	12	1
35	MONTANA	336	38	10	0
36	ILLINOIS	1,359	38	13	1
37	IDAHO	215	37	13	0
38	OKLAHOMA	706	36	10	1
39	WYOMING	185	36	9	1
40	WISCONSIN	762	36	11	0
41	KENTUCKY	520	35	12	0
42	MINNESOTA	886	33	9	0
43	ARKANSAS	529	32	9	0
44	MISSOURI	992	31	8	0
45	ALASKA	308	30	8	1
46	KANSAS	646	29	6	0
47	IOWA	980	23	4	0
48	SOUTH DAKOTA	356	22	4	2
49	NEBRASKA	541	20	4	0
50	NORTH DAKOTA	345	18	3	0
	U.S, Total	28,527	43	17	2

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. Many of the concerns are captured in the in this passage from the National Academy of Sciences report on the 2020 Census (2022, page __),

> "The adoption of the differential privacy-based solution was made with unusual hast relative to other sweeping changes in census methodology that were researched and tested much more extensively… the Census Bureau committed to overhauling its disclosure avoidance approach in 2018 without testing it or prototyping, much less have a working system in place."

Below I review a couple of issues regarding DP that were not addressed in my analysis but may impact census 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) is not available for census blocks or any the sub-state level. 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 most important 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 from the 2020 Census, 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. Readers are likely to have their own opinion about this.

One problem with use of DP for small areas is the implausible or impossible results produced. For example, there were 162,743 blocks that 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 2023c). 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 (2023c) 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 numerous blocks with children but no adults and other impossible or implausible 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 may wonder what other errors are in the data that are not so apparent.

Despite the caution 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 often because it is the best (or only) data they have at that level of geography.

The mean absolute percent error for all urban blocks was 25 percent and for all rural blocks it was 29 percent (U.S. Census Bureau 2023c). 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 with Differential Privacy (DP) was applied to the DHC file. DP is administered to the young children (population age 0 to 4) 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.

The Census Bureau (2023b, page 10) describes the situation this way,

"Finally, regarding (iv) the Disclosure Avoidance System (DAS) TopDown Algorithm does break the connection between people and households, however, we recognize the importance of accurate data on children and families, so we have included some ''lagged variable' tables that restructure the tables so users can still get those links as count queries. It is important to note that differential privacy itself doesn't eliminate these links. The DHC data uses the TopDown Algorithm so those links are broken, but the Supplemental DHC (S-DHC) uses a different implementation of differential privacy that preserves links between children and families."

This statistical disconnection of children and parents is an on-going concern and may have important impacts in later Census products which have more detailed data on young children. Also, 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.

After the Demographic and Housing Characteristics file is released, the next files from the 2020 Census are a set of Detailed-Demographic and Housing Characteristics files. 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. The method is called PHSafe (U.S. Census Bureau 2023d, slide 35). Hopefully, that will alleviate concerns. Data that links children and adults in the Supplemental - Detailed Demographic and Housing file will be available in late 2023 or 2024. When this file is released, it will provide an opportunity to assess how well it handles retaining the relationships between children and parents.

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 Census results should be examined through the lens of equitable data.

In terms of equity, Table 3 shows substantial differential accuracy in terms of absolute *percent* errors for Unified School Districts by race and Hispanic Origin Status after DP is applied. For Hispanic young children, the mean absolute *percent* error was 28, for Black young children the mean absolute *percent* error was 37, and for Asian young children was 47, compared to 6 for non-Hispanic white children. What does this say about the equity of using the DP method? Undercounts for young Black and Hispanic children are not yet available from the 2020 Census, but 2010 Census shows these groups were undercounted at a high rate (O'Hare 2015). To be clear the higher percent error rates for Black, Hispanic, and Asian young children is due to their smaller population sizes, not something the Census Bureau does specifically for or to those groups in processing the data. Nonetheless, there are concerns about data equity.

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?

Measuring Privacy Protection

A question that is not addressed in the previous sections of this report 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 and the answer to that question probably depends on the application.

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 Census Bureau‘s Demonstration Products. 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.

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 easily 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 the quality services 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.

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-State Cooperative 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) acknowledge the problems the implementation of DP has caused data users and 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."

The implementation of DP in the 2020 Census also caused significant delays in releasing the 2020 Census data. As this paper is being written, the only data from the 2020 Census that has been released is the re-apportionment data and the redistricting data. This delay damages the Census Bureau's reputation. I see no evidence that the Census Bureau took into consideration the likely delays in getting data from the 2020 Census out to data users and the damage use of DP would do to the Census Bureau's reputation in their decision to implement DP in the 2020 Census. Delays in releasing data and harm to the reputation of the Census Bureau should be taken into consideration in any future applications of DP.

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. 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 young child implications of using a DAS method with DP in the Demographic and Housing Characteristics file.

Big errors are most problematic. Errors of a few young children or a few percent are not likely to cause big problems, but when data is off by as much as 10 or 25 percent, it will cause problems and the analysis presented here indicates errors of such magnitude are likely to be prevalent in the 2020 Census for Unified School Districts and for Places.

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 continue to have concerns about the application of DP in the 2020 Census.

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

Gordon, N. and Reber, S. (2023). "Title I of ESEA: How The Formulas Work, " Brookings https://all4ed.org/publication/title-i-of-esea-how-the-formulas-work/

Gutierrez, F. (2022).American Community Survey and Child Well-Being, presentation at Population Association of America annual Conference,

Harvard Data Science Review (2022). "Differential privacy for the 2020 U.S. Census: Can We Make Data Both Private and Useful?" Special Issue, Il I
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/ ,

IPUMS-NHGIS (2023). “Privacy protected 2010 Census Demonstration Data,” https://www.nhgis.org/privacy-protected-2010-census-demonstration-data#v20230403-coverage

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. (2015). *The Undercount of Young Children in the U.S. Decennial Census. Springer Publishers* https://www.springer.com/gp/book/9783319189161#

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. (2021d). “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 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. (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. (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/

O'Hare. W. P. (2022d). Analysis of Census Bureau's August 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children, (Sept) . https://secureservercdn.net/198.71.233.229/2hj.858.myftpupload.com/wp-content/uploads/2022/09/Implications-of-Differential-Privacy-for-kids-9-21-2022-FINAL-.pdf

O'Hare, W. P. (2022e), Presentation at National Academy of Science Workshop on 2020 Census Data Products – Public Workshop on the Demographic and Housing Characteristics Files, June 21- 22 https://www.nationalacademies.org/documents/embed/link/LF2255DA3DD1C41C0A42D3BEF0989ACAECE3053A6A9B/file/D15F9C49F2F01F604B6440B90766E66EA9CE7E3B133A?noSaveAs=1

O'Hare, W.P. (2022f). 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. (2019). *Counting for Dollars 2020: The Role of the Decennial Census in the Geographic Distribution of Federal Funds. Brief 7: Comprehensive Accounting of Census-Guided Federal Spending (FY2017)*. https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds

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

Ruggles, S. and Van Riper, D. (2022). The Role of Chance in the Census Bureau Database Reconstruction Experiment, *Population and Policy Review*, Vol. 41. pp 781-788.
.
Swanson, D.A., Cossman, R. (2021)The Effect of the Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Mississippi. The Journal of the Mississippi Academy of Sciences, Vol 66, No 3, pp210-218.

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. Census Bureau (2022h). "Disclosure Avoidance Protections for the American Community Survey," December 14, , Donna Daily, https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html

U.S. Census Bureau (2023a) U.S. Census Bureau Workshop on Assessing Fitness-for-Use of Differential Privacy Adjusted Census Data, Workshop at the Population Association of America conference, New Orleans, April 12-15

U.S. Census Bureau (2023b) Letter from Robert L. Santos to Cherokee Bradley on Recommendations and Comments to the U.S. Census Bureau from the National Advisory Committee, 2022 Special Session on Differential Privacy. Dated April 7, 2023.

U.S. Census Bureau (2023c) Detailed Summary metrics

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

Winkler, R.L, Butler, J.L, Curtis, K. J, , Egan-Robertson, D. (2022) Differential Privacy and the Accuracy of County-Level Net Migration, Estimates, *Population and Policy Research*, Vol 41, pp 417-435.