Analysis of Census Bureau's April 2021 Differential Privacy Demonstration Product: Implications for Data on Children

By

Dr. William P. O'Hare

May 2021

Contents

Analysis of Census Bureau's April 2021 Differential Privacy Demonstration Product: Implications for Data on Children

By

Dr. William P. O'Hare

Executive Summary

The U.S. Census Bureau is planning to use a new method called differential privacy (DP) when it releases data from the 2020 Census to help protect confidentiality and privacy of respondents. This paper provides some information on how DP is likely to impact the accuracy of data for children (population ages 0 to 17) in the 2020 Census. The study is based on analysis of the most recent DP demonstration product released by the Census Bureau in April 2021, which applied DP to 2010 Census data. The DP demonstration product issued in April 2021 supersedes four earlier DP demonstration products.

This paper is meant to provide stakeholders and child advocates with some fundamental information about the level of errors DP will inject into the 2020 Census data for the population ages 0 to 17. It is meant to help stakeholders gain a better understanding of the implications of DP for children, and to enable data users to provide constructive feedback to the Census Bureau on their use of DP. In June of 2021 Census Bureau leadership will determine the final accuracy parameters the redistricting data (P.L. 94-171) to be released by September 30, 2021.

According to the Census Bureau, the demonstration file released by the Census Bureau on April 28, 2021 has been optimized for the redistricting application. However, 2020 Census data files that come out from the Census Bureau after the redistricting

data are released, for example the Demographic Profiles and the Demographic and Housing Characteristics files, will have more detailed data on children and the data in later files are likely to be made consistent with the total number of 0 to 17-year-olds reported in the redistricting data. So, errors in the data for 0 to 17-year-olds published in the redistricting data will have implications for child data in 2020 Census files that come out later. In that sense the analysis of the redistricting data can provide some understanding of the likely accuracy of later 2020 Census data products with data on children. The Census Bureau has indicated it hopes to engage stakeholders in decisions about what data to include and privacy parameters for those subsequent files.

To its credit, the Census Bureau has quantified its accuracy target for the redistricting data it will release next August/September "…we created an accuracy target to ensure that the largest racial or ethnic group in any geographic entity with a total population of at least 500 people is accurate to within 5 percentage points of their enumerated value at least 95% of the time." This leaves open what will happen to geographic units of less than 500 people, and it leaves open how large the errors will be for the 5 percent of the data that are more than 5 percent off.

This paper presents analysis of the error introduced by DP by comparing the data as reported in the 2010 Census Summary File and the same data after DP has been injected as released in the April 2021 Census demonstration file. Analysis presented in this paper found little impact of DP for large (highly aggregated) geographic units like states or large counties. However, the story is different for smaller places. Many smaller areas have high levels of error. For example, the count of children would exhibit absolute percent error of 5 percent or more in about 8 percent

of Unified School Districts after DP is applied. Bigger absolute error percentages are evident for several minority child populations. Also, the data show that 66 percent of Unified School Districts had absolute numeric errors of 10 or more children. Errors of this magnitude could have implications for federal and state funding received by schools and for educational planning. Data also show that 44 percent of places (cities, village, and towns) had absolute percent errors of five percent or more and 56 percent of places had absolute numeric errors of 10 or more children.

Moreover, after the injection of DP in the 2010 Census data included in the April demonstration product, there are over 91,000 blocks nationwide that had population ages 0 to 17, but no population ages 18 or over. Blocks with children and no adults is a highly implausible situation and the large number of blocks with children, but no adults may undermine confidence in the overall Census results. These implausible results are likely due to children being separated from their parents in DP processing. This separation is an ongoing concern for data on children.

Based on the errors for child population from the level of DP used in the April 2021 DP demonstration product, and the lack of clarity about privacy protection from DP, I recommend the Census Bureau reduce the size of errors injected into the 2020 Census data

There are a couple of reasons for sharing this information with child advocates now. First, when the 2020 Census results are published there may be some localities 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. As stated earlier, the U.S. Census Bureau is still looking for feedback on the use of DP in the 2020 Census. They are looking for cases where census data are used to make decisions. The Census Bureau is asking data users to examine the April 2021 DP demonstration product to see if the error injected by DP make the data unfit for use. After reading this report, we 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 data so feedback from census data users is important. If many users feel the current level of accuracy for data on children is not accurate enough for some uses, there is a chance the Census Bureau could make the data more accurate.

The demonstration product released on April 28, 2021 is the last demonstration product the Census Bureau will release before they Census Bureau Data Stewardship Executive Policy Committee decides on the DP parameters for the redistricting data (P.L 94-1717 file) that will be released by September 30, 2021 (a version of this file may be made available in August 2021).

Stakeholders, child advocates, and data users should take advantage of this opportunity to communicate their thoughts to the Census Bureau before a final decision is made. Let the Census Bureau know how the errors injected by DP are likely to impact your work and effect of lives of children in your state or community.

Thoughts and reactions to the data based on the DP file issued in April 2021 are due by May 28, 2021. Comments and responses can be sent to 2020DAS@census.gov. It would help if you put "April 2021 Demonstration Data" in the subject line of the email.

Analysis of Census Bureau's April 2021 Differential Privacy Demonstration Product: Implications for Data on Children

By

Dr. William P. O'Hare

Introduction

The U.S. Census Bureau is planning to use a new method called differential privacy (DP) in releasing data from the 2020 Census to help protect confidentiality and privacy of Census respondents.[1] This paper focuses on metrics for assessing the accuracy of census data for children (population ages 0 to 17) after DP is injected by reporting on the level of errors injected into the Census data for children based on the most recent demonstration product data available from the Census Bureau.

In short, DP injects errors in the data provided by respondents to make it more difficult for someone to be identified in the Census data. Adding or subtracting random numbers to the census results makes it more difficult to identify data for specific respondents. The U.S. Census Bureau (2020e) provides more information on the use

[1] The terminology in this arena can be confusing. Differential privacy is sometimes called "formal privacy." The system developed for the 2020 Census 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.

of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). For an independent look at differential privacy see Boyd (2020) and Bouk and Boyd (2021). More information about the DP issue and recent developments are provided in Appendix B.

The Census Bureau provided some suggested accuracy metrics with the April 28, 2021 release but as far as I can tell, none of the metrics provides data for the population age 0 to 17 (U.S. Census Bureau 2021b and c). This report tries to fill that gap.

The lack of Census Bureau-supplied metrics on children is probably because the April 2021 release did not contain data for the population ages 0 to 17. The population ages 0 to 17 had to be derived as explained later in the report.

I focus first on 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 and then census blocks. Places include big cities and small villages. They typically have policymaking authority, and they often provide programs for children. Blocks are the most basic building block for census data. Examination of block data show how DP is likely to impact the smallest areas. There is wide agreement that DP injects substantial errors into block-level data but there is less agreement on how important that is.

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 (Census Bureau provided data can be examined at https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html) .

For simplicity I only look at a few key measures here, but I believe they provide sufficient information to reach some conclusions. The measures used here, (mean absolute numeric error, mean absolute percent error, and outliers) are a subset of those featured in a Census Bureau webinar on this topic held May 14, 2021. Like the Census Bureau’s assessment of DP-infused data, I provide data for both numerical errors and percent errors because either can be important in some contexts and combining both provides a more complete picture of the error profiles for geographic units.

Errors are defined here as the difference between the data as reported in the 2010 Census Summary File and the same data after DP has been injected.

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 children) may be trivial in large places but very significant in smaller places. For example, a numeric error of 10 children in a school district of 1,000 children is only a 1 percent error, but a numeric error of 10 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 numeric errors and one for percent errors.

I believe the number and percent of large errors, sometimes called outliers, 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 they are accompanied by large swings in funding that are not connected to changes in population size. Such errors 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 DP demonstration file released by the Census Bureau on April 28, 2021 provides DP-infused data from the 2010 Census which can be compared to the 2010 Census data without DP to understand the impact DP has on data accuracy. The April 2021 DP demonstration file provides data for all census geographic levels including the smallest unit (census blocks).

As stated earlier in this paper, the Census Bureau file released in April 2021 does not provide data for the population age 0 to 17 directly, but it does provide the total population (all ages) and the population age 18 and older. By subtracting the population age 18 and older from the total population, one can derive the population ages 0 to 17. I call the population ages 0 to 17, children.

The data used in my analysis were originally provided by the Census Bureau in a huge (about 308 million records) Privacy Protected Microdata File (PPMF). Since many people do not have the computer power to analyze such a large file, the IPUMS- NHGIS unit at the University of Minnesota processed the PPMF and put the data into user-friendly tables. I analyze the data produced by IPUMS-NHGIS unit. The data used for this study are available at https://nhgis.org/privacy-protected-demonstration-data

Results

Table 1 provides several accuracy measures for the population ages 0 to 17 for four kinds of geographic units. The results shown in Table 1 indicate that DP is unlikely to have much of an impact on the child data for states (and the District of Columbia). Also, it is unlikely to have much impact on county child data (percentage wise) since most counties are relatively large. However, of the 3,141 counties examined here, about one-fifth have populations less than 10,000 total population, where DP may inject enough error to be problematic. For this subset of counties, DP may distort the data to a problematic degree (O'Hare 2019).

The situation is different for Unified School Districts and places (shown in Table 1), where DP is likely to cause substantial distortions for the child population. For census blocks, which are examined later in the report, the problematic situations are magnified because most blocks have very small populations.

Table 1 Key Statistics for Absolute Numeric and Absolute Percent Errors* for All Children Ages 0 to 17 for Selected Geographic Units

	States	Counties	School Districts	Places
Number of Units in the Analysis	51	3,142	10,822	29,364
Mean Size of District (Total Population)	6,053,834	98,264	28,372	7,851
Mean Absolute Numeric Error**	57	28	30	26
Mean Absolute Percent Error	0.01	0.79	2.1	12.8

Source: Author's analysis of data released by the Census Bureau on April 28, 2021.

Does not include Puerto Rico or geographic units with zero population age 0 to 17 in 2010 Summary 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.

I show the most important data in the text, but more detailed data are provided in Appendix A.

Application 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). 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. At the CNSTAT DP workshop held in December 2019 there were several

presentations reflecting implications of DP-infused data for children and school districts (Vink 2019; Nagle and Kuhn 2019: Sojourner 2019).

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 individual attendance boundaries to keep classrooms from becoming overcrowded. 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.

As noted earlier, DP has a bigger impact, percentage wise, in smaller places and the majority of Unified School Districts are relatively small. Out of 10,880 school districts, more than half have less than 10,000 total population. Many of the 10,822 Unified School Districts are very small; 266 of the Unified School Districts had a total population ages 0 to 17 of less than 100, and 1,910 districts had population ages 0 to 17 of less than 500 in the 2010 Census The translation of small numeric errors into large percent errors is also more apparent in looking at data for race and Hispanic groups within school districts.

Table 2 shows several measures of accuracy/error for 10,822 Unified School Districts in the 2010 Census. The data are provided for all children (all races) as well as for Black children, Hispanic children, and Asian children.[2] Other race groups were not examined here because the numbers were small, and time was limited. For the remainder of this report when I use the term Black or Asian, it means Black alone or Asian alone.

Data in Table 2 show the vast majority of Unified School Districts have at least one Black child, one Hispanic child, and one Asian child. But many districts have very small numbers of minority children. The average number of Black children in school districts where there was at least one Black child was 1,096, for Hispanics it was 1,599 and for Asians it was 354. These numbers are well below the overall average of 6,817 children. The relatively small number of Black, Hispanic, and Asian children in many districts results in these groups having smaller absolute numeric errors but larger absolute percent errors.

[2] . I use race alone rather than alone or in combination because the data for race alone was more easily available from the source file using that definition of race and I didn't have the time to pull together data on race defined as alone or in combination.

Table 2. Key Measures of Errors* for Children (ages 0 to 17) for Unified School Districts by Race and Hispanic Origin				
	All Children	Black Alone Children	Hispanic Children	Asian Alone Children
Number of Units Included in Tabulation	10,882	9,891	10,714	9,198
Mean Number of Children for Districts Used in Tabulation	6,817	1,096 Black Children	1,599 Hispanic Children	354 Asian Children
Mean Absolute Numeric Error	30	8	18	6
Mean Absolute Percent Error	2.1	37.9	23.2	48.8
Source: Authors analysis of data released by the Census Bureau on April 28, 2021.				
*in this paper error reflect the difference between the 2010 Census data without and with DP injected.				

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

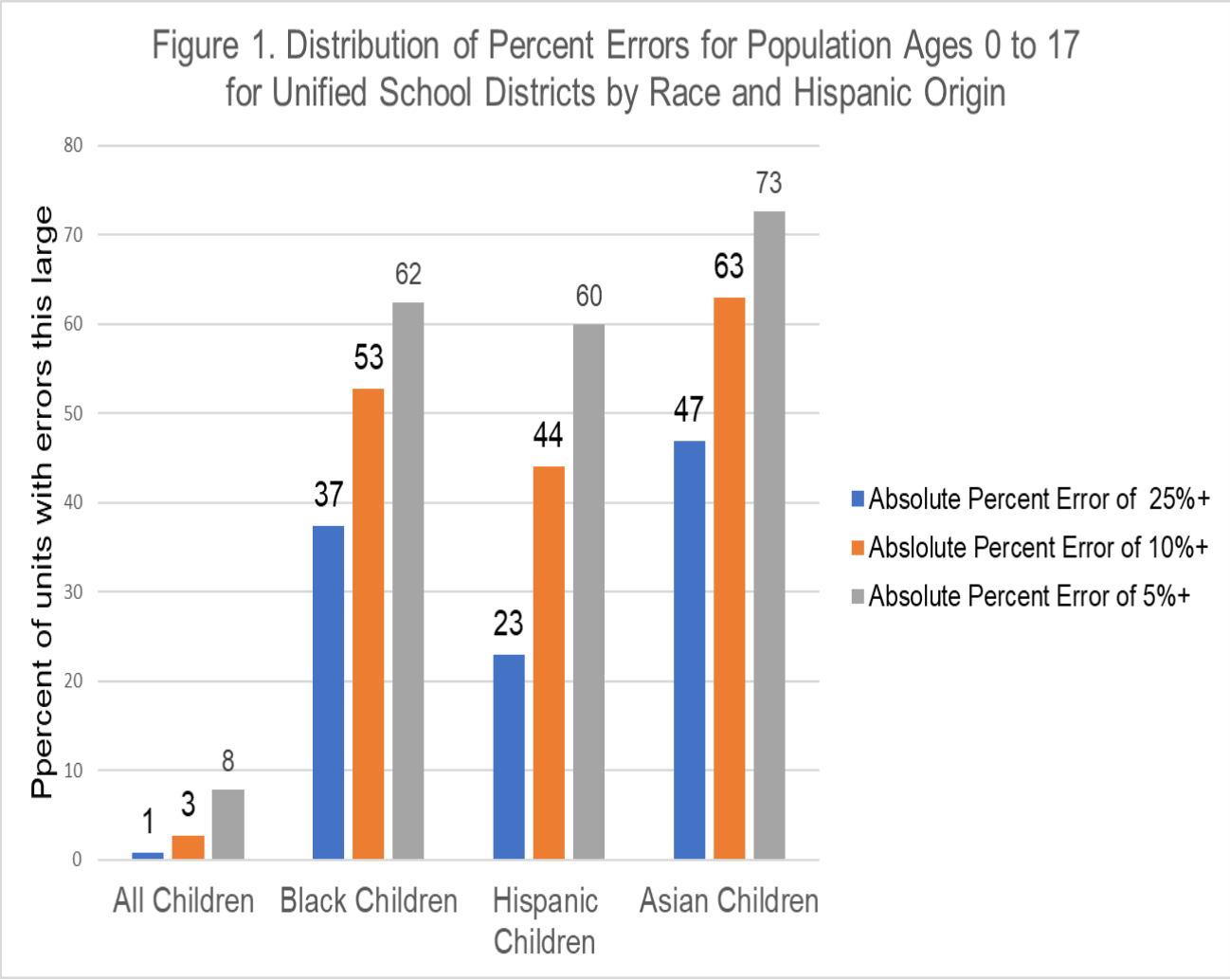
Table 2 shows the mean absolute numeric error for all children (all races) in Unified School Districts is 30 children. In other words, for the average unified school district the DP-infused data differs from the data without DP by 30 children. The mean absolute numeric errors for Black (8), Hispanic (18), and Asian (6) children are smaller than that for all children (30).

The mean absolute percent error shown in Table 2 for all children is 2.1 percent. For Black children, the mean absolute percent error was 37.9, for Hispanic children it was 23.2, and for Asian children was 48.8.

Means or averages are helpful, but they do not reveal the full story. An examination of the distribution error size can provide more information on the relative

accuracy of the DP-infused data. Large errors can be problematic even if the overall mean is relatively low.

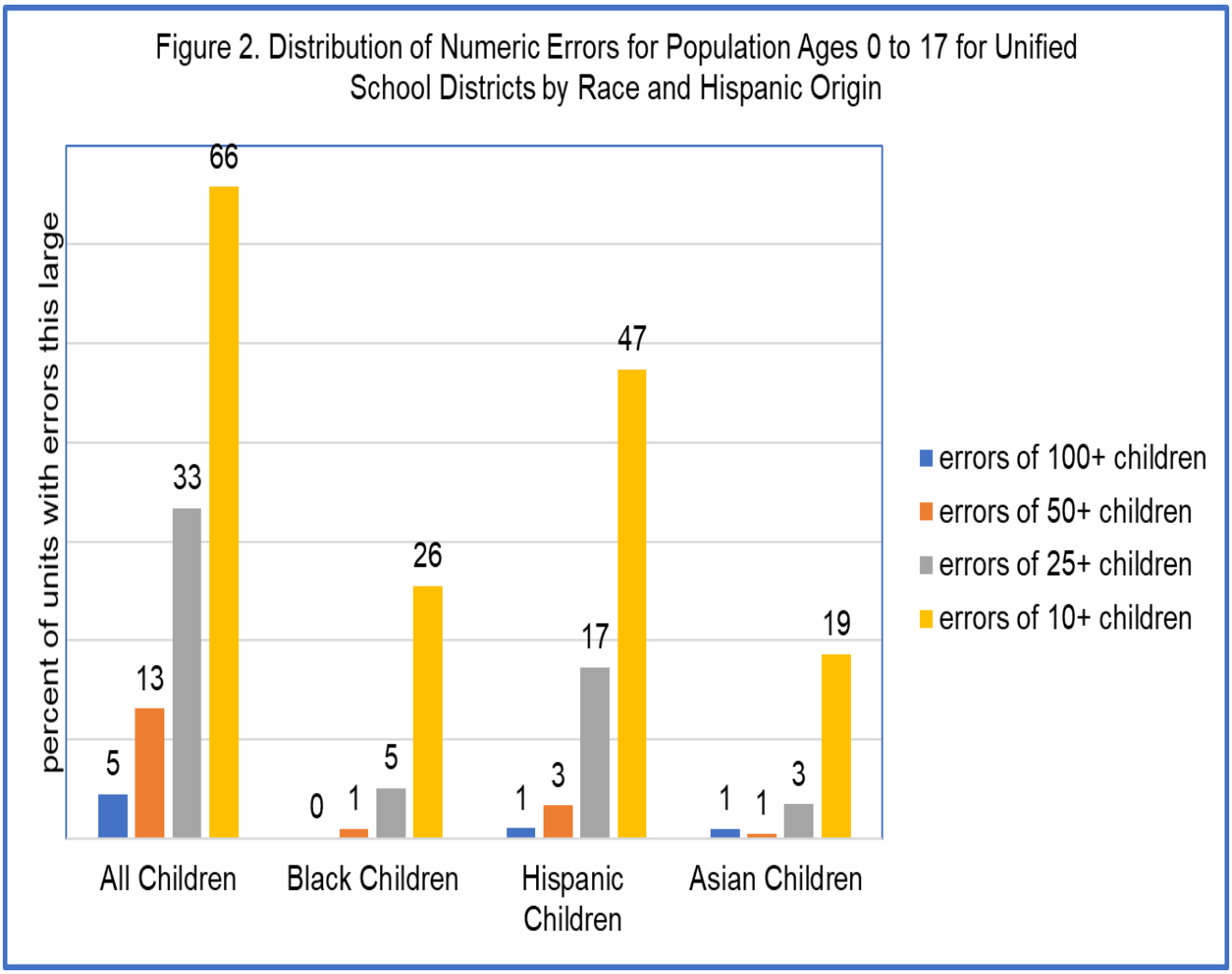
The absolute percent errors for Unified School Districts are put into three categories (more than 5 percent, more than 10 percent, and more than 25 percent). To be clear, the districts with more than 25 percent errors are also counted in the categories for more than 10 percent error and more than 5 percent error. These thresholds are judgmental, but I think they provide a reasonable range of errors.



The five percent and 10 percent categories are used by the Census Bureau 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.

Distributions of absolute percent errors are shown in Figure 1 which shows that for all children, 8 percent of districts had absolute percent errors of 5 percent or more, compared to 62 percent for Black children, 60 percent for Hispanic children, and 73 percent for Asian children. Since minority groups are smaller in population size, it is not surprising that there are more extreme absolute percent errors. Figure 1 shows that for minority children, absolute percent errors of 25 percent or more are relatively common,

The absolute numeric errors for Unified School Districts are put into four categories (more than 10 children, more than 25 children, more than 50 children, and more than 100 children). To be clear, the districts with errors of more than 100 children are also counted in the categories with errors of more than 50 children, more than 25 children and more than 10 children These thresholds are judgmental, but I think they provide a reasonable range of errors.



In each category of errors (10 children, 25 children, 50 children, and 100 children), there are many fewer districts that have this level of error for Black, Hispanic, and Asian children than there are districts that have this level of error for all children. Since the minority populations are smaller, the absolute numeric errors are also smaller even though the absolute percentage errors are larger.

Figure 2 shows for all children (all races) 66 percent of the Unified School Districts had errors of 10 children or more, but the figures for minority groups are

smaller (26 percent for Black children, 47 percent for Hispanic children and 19 percent for Asian children).

On the other hand, there are relatively few Unified School Districts with very large numeric errors. Only 5 percent of Unified School Districts has errors of 100 child or more, compared to 0 percent for Black children, 1 percent of Hispanic children and 1 percent for Asian children.

The national numbers shown above mask a lot of variation across states. Table 3 provides two key measures of accuracy (mean absolute numeric error and mean absolute percent error) for Unified School Districts in each state. The mean absolute numeric error for states ranges from a low of 10 for Vermont to a high of 114 for California. In other words, this error measure in California is more than ten times what it is in Vermont. The mean absolute percent error ranges from a low of 0 for Hawaii and DC to a high of 10.1 in Maine.

Table 3 States Ranked by Mean Absolute Numeric Error and Mean Absolute Percent Error for Children by School Districts

States Ranked by Mean Absolute Numeric Error				States Ranked by Mean Absolute Percent Error		
Rank	State	Mean Absolute Numeric Error		Rank	State	Mean Absolute Percent Error
1	California	114		1	Maine	10.1
2	Arizona	74		2	Vermont	8.4
3	Connecticut	51		3	New Mexico	5.4
4	Rhode Island	50		4	Oregon	5.0
5	Delaware	49		5	Idaho	4.6
6	DC	47		6	Montana	4.6
7	Utah	44		7	North Dakota	4.4
8	New Jersey	44		8	Washington	3.6
9	Washington	42		9	New Hampshire	3.2
10	Massachusetts	42		10	Colorado	3.1
11	South Carolina	41		11	Nebraska	3.0
12	Texas	36		12	Oklahoma	3.0
13	New Mexico	35		13	South Dakota	2.9
14	Colorado	34		14	Alaska	2.9
15	Michigan	33		15	New York	2.8
16	Illinois	33		16	Texas	2.6
17	Oregon	32		17	Kansas	2.1
18	New York	31		18	Arizona	2.1
19	Tennessee	29		19	New Jersey	2.1
20	Pennsylvania	28		20	Wyoming	1.9
21	Hawaii	28		21	Iowa	1.8
22	Indiana	28		22	South Carolina	1.8
23	Ohio	28		23	Missouri	1.8
24	Georgia	27		24	California	1.6
25	Alabama	26		25	Minnesota	1.5
26	Minnesota	24		26	Illinois	1.5
27	Arkansas	23		27	Arkansas	1.5
28	Oklahoma	23		28	Ohio	1.5
29	Missouri	22		29	Michigan	1.2
30	North Carolina	22		30	Wisconsin	1.2
31	Wisconsin	21		31	Indiana	1.0
32	Idaho	20		32	Rhode Island	1.0
33	Mississippi	20		33	Massachusetts	1.0
34	Maryland	19		34	Connecticut	0.7
35	New Hampshire	18		35	Pennsylvania	0.7
36	Kansas	18		36	Kentucky	0.7
37	Nebraska	18		37	Delaware	0.7
38	Iowa	17		38	Mississippi	0.7
39	Kentucky	17		39	Utah	0.5
40	Florida	17		40	Nevada	0.5
41	Wyoming	16		41	Alabama	0.5
42	Louisiana	16		42	Tennessee	0.4
43	North Dakota	14		43	Virginia	0.4
44	Virginia	14		44	Georgia	0.4
45	Nevada	14		45	North Carolina	0.3
46	South Dakota	14		46	West Virginia	0.3
47	Alaska	11		47	Louisiana	0.2
48	West Virginia	11		48	Florida	0.1
49	Maine	10		49	Maryland	0.1
50	Montana	10		50	DC	0.0
51	Vermont	9		51	Hawaii	0.0
	U.S. Total	30			U.S. Total	2.1

Source: Authors analysis of DP demonstration product released by the Census Bureau on April 26, 2021

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 April 28, 2021 (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 children in their area for things like planning youth activities, child facilities, and day care centers.

Data in Appendix A show the absolute mean numeric error for places is 26 children and the mean absolute percent error is 12.9.

Many of these places are small. There were 8,761 places where the number of children was less than 100 and 18,705 places where the number of children was less than 500, based on the 2010 Summary File. The fact that many places are small (in

population size) means they are likely to have relatively large absolute percent errors, and this is reflected in Figure 3.

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 44 percent of places had errors of 5 percent or more for the child population and more than one out of ten (11 percent) had 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 than for Unified School Districts.

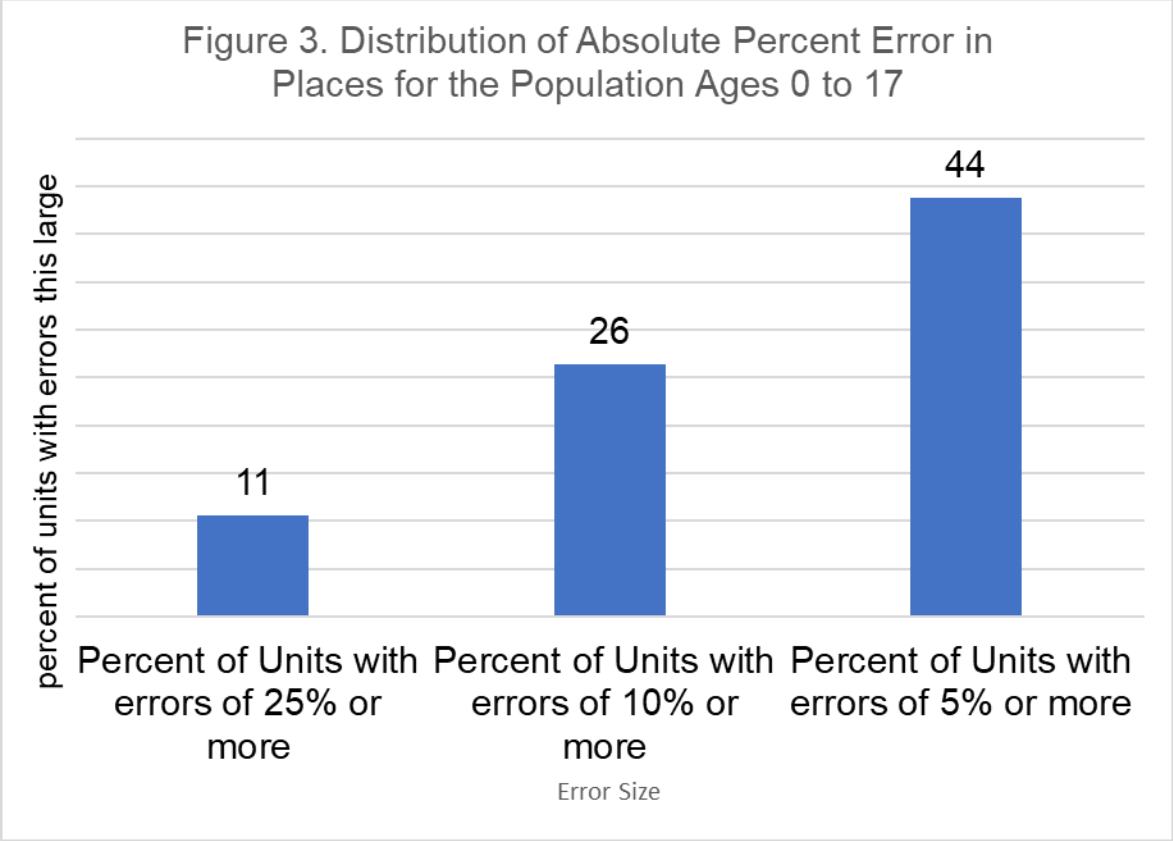


Figure 4 show the distribution of places by absolute numeric errors using the same categories as Figure 2. More than half (56 percent) of the places had errors of 10 or more children and one-quarter of places had absolute numeric errors of 25 or more children.

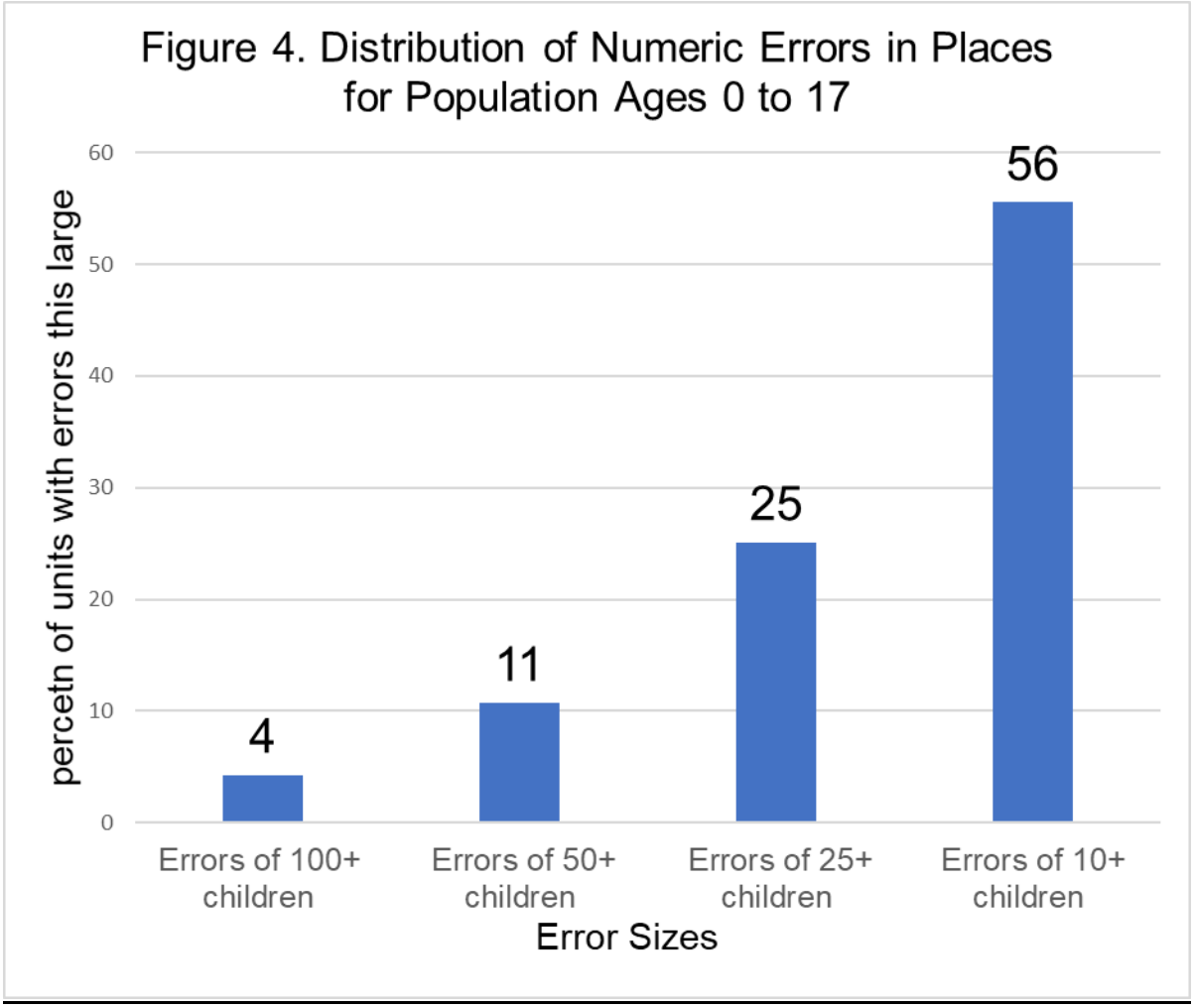


Table 4 provdies the distribution of absolute percent errors for each state using the same categories as used in Figure 3. There is a lot of variation across the states. For example, 35 percent of the places in North Dakota had absolute percent errors of 25 percent or more, compared to just 2 percent in Maine.

Table 5 provdes the distribution of absolute numeric errors for each state using the same categories as used in Figure 4. There is a lot of variation among the states. For example, 16 percent of places in California had errors of 100 or more children, compared to zero in DC, Maine, Montana, North Dakota, and South Dakota.

Table 4 Distribution of Places by Absolute Percent Error* for Population Ages 0 to 17 by State

		Number of places with errors of this size					Percent distribution within the State			
Row Labels	State Total	less than 5 percent	5 to 9.9 percent	10 to 24.9 percent	25 percent or more		less than 5 percent	5 to 9.9 percent	10 to 24.9 percent	25 percent or more
Alabama	577	344	102	80	51		60	18	14	9
Alaska	338	117	66	64	91		35	20	19	27
Arizona	445	232	78	65	70		52	18	15	16
Arkansas	541	272	104	106	59		50	19	20	11
California	1,505	969	201	185	150		64	13	12	10
Colorado	453	234	57	88	74		52	13	19	16
Connecticut	142	99	22	15	6		70	15	11	4
Delaware	76	46	8	9	13		61	11	12	17
DC	1	1	0	0	0		100	0	0	0
Florida	915	623	121	108	63		68	13	12	7
Georgia	623	395	94	94	40		63	15	15	6
Hawaii	151	84	40	17	10		56	26	11	7
Idaho	224	121	40	29	34		54	18	13	15
Illinois	1,367	861	225	197	84		63	16	14	6
Indiana	681	440	113	90	38		65	17	13	6
Iowa	1,008	447	202	198	161		44	20	20	16
Kansas	668	283	119	134	132		42	18	20	20
Kentucky	523	312	102	81	28		60	20	15	5
Louisiana	473	328	77	48	20		69	16	10	4
Maine	131	79	37	13	2		60	28	10	2
Maryland	518	328	69	67	54		63	13	13	10
Massachusetts	243	169	45	18	11		70	19	7	5
Michigan	691	417	112	109	53		60	16	16	8
Minnesota	902	461	166	159	116		51	18	18	13
Mississippi	362	217	82	48	15		60	23	13	4
Missouri	1,021	482	190	192	157		47	19	19	15
Montana	361	141	77	63	80		39	21	17	22
Nebraska	577	193	120	133	131		33	21	23	23
Nevada	128	61	17	21	29		48	13	16	23
New Hampshire	96	54	18	19	5		56	19	20	5
New Jersey	542	371	86	50	35		68	16	9	6
New Mexico	440	174	71	87	108		40	16	20	25
New York	1,187	676	233	219	59		57	20	18	5
North Carolina	738	439	140	111	48		59	19	15	7
North Dakota	391	117	48	90	136		30	12	23	35
Ohio	1,204	779	219	153	53		65	18	13	4
Oklahoma	727	300	140	150	137		41	19	21	19
Oregon	375	209	59	57	50		56	16	15	13
Pennsylvania	1,759	950	358	328	123		54	20	19	7
Rhode Island	34	23	2	3	6		68	6	9	18
South Carolina	395	233	76	49	37		59	19	12	9
South Dakota	376	146	62	81	87		39	16	22	23
Tennessee	428	293	67	54	14		68	16	13	3
Texas	1,745	1,059	299	254	133		61	17	15	8
Utah	323	219	51	39	14		68	16	12	4
Vermont	119	47	32	24	16		39	27	20	13
Virginia	591	376	97	80	38		64	16	14	6
Washington	627	389	105	83	50		62	17	13	8
West Virginia	400	208	78	82	32		52	20	21	8
Wisconsin	772	467	122	115	68		60	16	15	9
Wyoming	197	86	25	37	49		44	13	19	25
Total	29,111	16,371	5,074	4,596	3,070		56	17	16	11

Source: Authors analysis of data released by the Census Bureau on April 28, 2021.

Does not include Puerto Rico or geographic units with zero population age 0 to 17 in 2010 Summary File

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

Table 5 Distribution of Places by Absolute Numeric Error for Population Ages 0 to 17 by State												
		Number of places with errors of this size						Percent distribution within state				
State	State Total	Less Than 10	10 to 24	25 to 49	50 to 99	100 or more		Less Than 10	10 to 24	25 to 49	50 to 99	100 or more
Alabama	577	260	206	80	22	9		45	36	14	4	2
Alaska	338	208	89	28	9	4		62	26	8	3	1
Arizona	445	204	120	59	31	31		46	27	13	7	7
Arkansas	541	282	163	65	19	12		52	30	12	4	2
California	1505	439	398	236	188	244		29	26	16	12	16
Colorado	453	213	128	63	26	23		47	28	14	6	5
Connecticut	142	25	39	41	20	17		18	27	29	14	12
Delaware	76	26	27	15	7	1		34	36	20	9	1
DC	1		1					0	100	0	0	0
Florida	915	224	236	209	126	120		24	26	23	14	13
Georgia	623	261	200	96	46	20		42	32	15	7	3
Hawaii	151	26	40	40	26	19		17	26	26	17	13
Idaho	224	113	67	34	7	3		50	30	15	3	1
Illinois	1367	651	410	152	100	54		48	30	11	7	4
Indiana	681	350	209	82	25	15		51	31	12	4	2
Iowa	1008	624	303	52	17	12		62	30	5	2	1
Kansas	668	359	220	59	15	15		54	33	9	2	2
Kentucky	523	275	168	58	18	4		53	32	11	3	1
Louisiana	473	208	163	64	26	12		44	34	14	5	3
Maine	131	42	43	32	14			32	33	24	11	0
Maryland	518	192	146	83	54	43		37	28	16	10	8
Massachusetts	243	45	55	71	34	38		19	23	29	14	16
Michigan	691	237	217	142	54	41		34	31	21	8	6
Minnesota	902	473	279	88	34	28		52	31	10	4	3
Mississippi	362	167	130	44	17	4		46	36	12	5	1
Missouri	1021	577	300	97	32	15		57	29	10	3	1
Montana	361	220	108	28	4	1		61	30	8	1	0
Nebraska	577	368	162	37	6	4		64	28	6	1	1
Nevada	128	49	40	21	5	13		38	31	16	4	10
New Hampshire	96	32	32	22	8	2		33	33	23	8	2
New Jersey	542	115	134	148	102	43		21	25	27	19	8
New Mexico	440	206	152	63	15	4		47	35	14	3	1
New York	1187	305	370	315	123	74		26	31	27	10	6
North Carolina	738	317	235	111	57	18		43	32	15	8	2
North Dakota	391	297	81	8	4	1		76	21	2	1	0
Ohio	1204	565	358	172	69	40		47	30	14	6	3
Oklahoma	727	347	245	97	30	8		48	34	13	4	1
Oregon	375	149	128	62	20	16		40	34	17	5	4
Pennsylvania	1759	747	586	298	97	31		42	33	17	6	2
Rhode Island	34	6	10	10	5	3		18	29	29	15	9
South Carolina	395	158	120	72	33	12		40	30	18	8	3
South Dakota	376	256	105	11	3	1		68	28	3	1	0
Tennessee	428	168	161	65	22	12		39	38	15	5	3
Texas	1745	719	553	274	129	70		41	32	16	7	4
Utah	323	164	88	31	22	18		51	27	10	7	6
Vermont	119	56	35	20	7	1		47	29	17	6	1
Virginia	591	246	189	98	41	17		42	32	17	7	3
Washington	627	229	189	108	52	49		37	30	17	8	8
West Virginia	400	245	127	21	5	2		61	32	5	1	1
Wisconsin	772	373	240	103	36	20		48	31	13	5	3
Wyoming	197	119	59	12	6	1		60	30	6	3	1
Total	29111	11280	7693	3610	1546	904		39	26	12	5	3
Source: Author's Analysis of d released by the Census Bureau on April 28, 2021												
* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.												

Data for Census Blocks

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. For one thing, block-level census data are used for redistricting and this is one of the most important uses of census data in the public policy arena. One of the primary purposes of the Decennial Census is to provide comparable population figures for small areas across the country. Consequently, census accuracy for blocks is especially important.

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 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.

I don't 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 block level data are for making decisions.

Blocks are the smallest geographic unit used in the Census and there are about 8 million blocks in the 2020 Census. The average block has a total population of about 41 people and about 9 children. The small population size of blocks makes them susceptible to large percent errors when random numbers are injected with DP.

In terms of the distributions shown here, it should be noted that many blocks have zero children ages 0 to 17 and this can skew some of the data reported here. For example, in this analysis, a block with zero children shows up as a zero error and all the zeros impact the average error. It also impacts the proportion of blocks with extreme values because so much of the distribution is at zero.

Table 6 shows state ranked by key metrics for census blocks. Looking across all states, the mean absolute numeric error is 0.9 (i.e., about one child per block) and the absolute percent error is 27.6. The state mean for blocks with absolute percent errors of more than 5 percent is 38.1 percent.

The data in Table 6 indicates significant variation across states. For example, 53.2 percent of blocks in Rhode Island have absolute percent errors of 5 percent or more compared to 18.9 percent in North Dakota. Understanding why the share of census blocks in Rhode Island with errors of 5 percent or more is so much higher than North Dakota would involve examination of detailed data for those states. Other metrics show similar variation across states.

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

.

Table 6. States Ranked by Percent of Blocks with Errors* of 5 Percent or More for Children (ages 0 to 17)

Rank		Mean absolute numeric error	Mean absolute percent error	Percent of Blocks with Errors of More than 5%
1	Rhode Island	1.4	29.4	53.2
2	Connecticut	1.5	29.3	51.7
3	New Jersey	1.5	26.0	51.3
4	Delaware	1.3	33.8	49.6
5	District Of Columbia	1.6	21.6	49.3
6	New York	1.3	28.7	48.5
7	Pennsylvania	1.0	33.3	48.4
8	North Carolina	1.2	34.7	47.2
9	Indiana	1.0	33.7	47.0
10	Florida	1.2	31.4	46.2
11	Washington	1.3	31.1	45.9
12	Illinois	1.0	30.6	45.7
13	Ohio	1.0	29.9	45.5
14	Massachusetts	1.2	25.6	44.9
15	Michigan	1.0	28.5	44.4
16	South Carolina	1.0	33.6	43.2
17	California	1.5	24.0	42.2
18	Georgia	1.1	30.3	41.4
19	Wisconsin	0.9	29.5	41.3
20	Tennessee	0.9	30.7	41.1
21	New Hampshire	0.9	30.2	40.5
22	Maryland	1.1	26.6	40.1
23	Iowa	0.7	35.4	40.0
24	Minnesota	0.8	30.1	38.7
25	Kentucky	0.8	27.4	38.3
26	Colorado	1.0	27.7	37.9
27	Alabama	0.8	30.2	37.3
28	Hawaii	1.8	23.6	37.3
29	Oklahoma	0.8	33.1	36.8
30	Vermont	0.7	30.2	36.2
31	Texas	1.0	24.5	36.1
32	Virginia	0.9	27.9	36.0
33	Missouri	0.7	29.1	35.7
34	Arkansas	0.7	29.9	35.6
35	Louisiana	0.8	23.7	35.0
36	Arizona	1.0	25.1	34.2
37	Maine	0.7	27.9	34.0
38	Mississippi	0.7	26.4	33.8
39	Kansas	0.7	30.4	33.0
40	Oregon	0.8	24.5	31.7
41	West Virginia	0.6	28.6	31.6
42	Nevada	1.1	21.4	31.4
43	South Dakota	0.6	30.3	30.4
44	Nebraska	0.6	29.4	30.2
45	Utah	0.8	15.4	27.8
46	New Mexico	0.6	22.9	26.6
47	Idaho	0.5	21.4	25.1
48	Montana	0.4	22.3	22.8
49	Alaska	0.7	16.9	21.9
50	Wyoming	0.4	18.4	19.8
51	North Dakota	0.3	21.4	18.9
	Mean of states	0.9	27.6	38.1

Source: Authors analysis of data provided by David Van Riper of IPUMS at the University of Minnesota.

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

Impossible or Improbable Results

Another aspect of differential privacy is production of impossible or implausible results. One such result involves children. After differential privacy is applied to the 2010 Census data, many blocks have children (ages 0 to 17) but no adults (ages 18 or older). Realistically, a state or community could have a few cases such as this, but states or communities with many such cases are highly unlikely and raise questions about who these children are living with if there are no adults in their household.

Table 7 shows the states ranked by the percent of blocks in their state where there are children (ages 0 to 17) but no adults (ages 18 or older). Table 7 shows that for each state the number of such improbable blocks are relatively high (the mean is 1,785) but the share is relatively low (0.84%). Most states have at least one thousand such blocks. The total of such blocks from data used to compile Table 7 is over 91,000.

South Dakota, where 1.71 percent of all blocks have this condition, is the state with the highest rate. On the other hand, the District of Columbia (0.08) has the lowest percent of blocks where there are children but not adults. The state with the most census blocks that meet this condition is Texas with 6,250

To assess the impact of DP, I look at the difference between the number of blocks with children and no adults for a couple of states where data without DP were readily available. It appears the injection of DP into the 2010 Census data increased the number of such blocks dramatically from what was reported in the 2010 Census. In published data from the 2010 Census, there were 21 blocks in New York state where there was a population age 0 to 17, but no population ages 18 or older, but in the April

28th DP demonstration file there were 2,691 such blocks. For Alabama, the increase was from 5 to 2,263, for Alaska the number of such blocks increased from 3 to 304, and for Washington State it went from 11 to 1,112. It appears that DP greatly expands the number of blocks were there are children, but no adults compared to the processing used in the 2010 Census.

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 DP is used. If the processing retained the link between 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 is an on-going concern and is likely to have important impacts in later Census products which have more detailed data on children.

Other kinds of impossible or improbable results are shown in Appendix C. Including:

- 241,299 blocks where the total population changed from greater than 50 percent Non-Hispanic White to less than 50 percent Non-Hispanic White after DP was applied.
- 56,661 blocks with a population before DP was applied but no population after DP was applied.
- 674,588 blocks with population in households but no occupied housing units after DP was applied.
- 76,892 blocks with occupied housing units but no population after DP was applied

- 715,929 blocks with more than 15 persons per household after DP was applied.

These types of improbable result may indicate other improbable results also exist but are harder to detect.

It is not clear to me exactly what statistical problems might be caused by results such as those shown in Table 7, but they undermine the veracity of the census data broadly. The high number of improbable results reflected in Table 6 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.

Table 7. States Ranked by Percent of Blocks with Population Ages 0 to 17, but No Population Ages 18 or Older

State	Number of Blocks with population ages 0 to 17 but no population ages 18 or older	Percent of Blocks with population ages 0 to 17 but no population ages 18 or older
South Dakota	1,507	1.71
Nebraska	2,973	1.54
North Dakota	2,009	1.5
Kansas	3,524	1.48
Iowa	3,115	1.44
Maine	979	1.41
Vermont	437	1.34
Oklahoma	3,295	1.22
West Virginia	1,606	1.19
Minnesota	3,057	1.18
Montana	1,566	1.18
New Hampshire	570	1.17
Missouri	3,845	1.12
Arkansas	1,939	1.04
Wisconsin	2,577	1.02
Idaho	1,507	1.01
Mississippi	1,609	0.94
Alabama	2,263	0.9
Colorado	1,767	0.88
Indiana	2,363	0.88
Tennessee	2,120	0.88
Virginia	2,452	0.86
South Carolina	1,550	0.85
Kentucky	1,351	0.84
Michigan	2,771	0.84
New Mexico	1,364	0.81
Wyoming	701	0.81
North Carolina	2,296	0.79
New York	2,691	0.77
Illinois	3,415	0.76
Ohio	2,791	0.76
Pennsylvania	3,162	0.75
Arizona	1,768	0.73
Georgia	2,110	0.72
Utah	811	0.7
Texas	6,250	0.68
Alaska	304	0.67
Louisiana	1,282	0.63
Oregon	1,193	0.61
Maryland	878	0.6
Washington	1,112	0.57
Delaware	123	0.51
Florida	2,031	0.42
Massachusetts	669	0.42
Nevada	356	0.42
Connecticut	266	0.39
California	2,247	0.32
Rhode Island	67	0.27
New Jersey	362	0.21
Hawaii	41	0.16
District Of Columbia	5	0.08
State Means	1,785	0.84

Source: Authors analysis of data provided by David Van Riper of IPUMS at the University of Minnesota.

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

Summary

The previous section provides information on accuracy of DP-infused data and provides a profile of the likely errors for children that will be seen in data for in the 2020 Census if the Census Bureau uses DP as reflected in the April 2021 demonstration product.

The question that is not addressed in the previous section is whether the level of error reflected in this analysis would make 2020 Census for data on children "unacceptable." Each person will probably have a different answer to how much error in census data for children is too much error. Below I provide my personal perspective on that question.

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. The Census Bureau has control over the level of accuracy and level of privacy protection in the 2020 Census largely by changing a parameter called "Epsilon." Increasing the level of Epsilon will increase accuracy in most cases, but an increase in Epsilon will also lower the level of privacy protection. The DP demonstration product issued by the Census Bureau on April 28, 2021 used an Epsilon of 12.3.

However, the level of privacy protection afforded by the parameters (mostly Epsilon) used in the DP demonstration product released by the Census Bureau on April 28th, 2021, is unclear to me. I don't know how an Epsilon of 12.3 translates to a measure of privacy that I understand (by that I mean, something like 1 percent or

respondents are at risk of re-identification with an Epsilon of 12.3). Moreover, it is not clear how much privacy protection would be lost if epsilon were increased by several points from the 12.3 used in the April 2021 DP demonstration product. I have searched the Census Bureau website and I have not found any information about someone who has been harmed by re-identification in the 2010 Census. In assessing the tradeoff between privacy and accuracy, my decision might be different if lower privacy protection meant hundreds of innocent people would go to jail versus people getting annoying phone calls. But I have not seen any evidence on this question.

On the other hand, the problems that are likely to be caused by inaccurate census data on children are relatively clear. The data in this paper, and many other analyses, provide a rich set of metrics on the extent to which DP injects error in the Census data.

When the number of children in a school districts is under-reported by 5 or 10 percent, that could have big implications for their funding and when the number of 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 children. If the number of 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 I believe there is a strong link between under reporting the number of children and the loss of money in a general sense.

Table 8 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data. These programs totaled almost $39 billion in FY 2017. In addition, other programs like the school lunch

program run by the U.S. Department of Agriculture, childcare funds given out by the Department of Health and Human Resources, and many other government programs, also use census-derived data to distribute funds. Reamer (2020) identified 316 federal programs that use census-derived data to distribute about $1.5 billion to states and localities in Fiscal Year 2017.

Table 8. 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-

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). Localities also use Census data for decision-making and distribution of resources.

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, the less likely those decision will be correct ones.

The most recent data available from the U.S. Census Bureau regarding the likely impact of DP on 2020 Census data for children suggests that the error introduced will result in a high level of errors for some small geographic units. The data shown here also underscores the point that DP-infused data are most problematic for smaller (less populated) units of geography. This is important because there are a large number of small geographic units for which census data are produced. This point is illustrated here based on places, school districts, and blocks.

Given the level of errors in Unified School Districts, places and census blocks using the Epsilon level in the most recent DP demonstration product, and the lack of clear evidence about the level or impact of privacy loss, I recommend that Census Bureau increase the level of Epsilon used in the redistricting data and subsequent data products to provide more accuracy small area data for children.

Author Note

It should be noted that this analysis is not as full and complete as it should be because time did not allow such an analysis. The Census Bureau released the latest DP demonstration product on April 28, 2021 and requires feedback by May 28, 2021. Since stakeholders need time to read and absorb this paper it needed to be available well before May 28th. If there had been more time for analysis there is a lot more that could have been done. The data used here could be developed to provide a more granular picture of DP's impact. For example, one could calculate the measures shown here for all counties or all places within a state, or one could develop the measures for all census tracts within a county.

If more time had been available, it would have been useful to explore data for race and Hispanic groups more thoroughly. Also, it would have been useful to examine accuracy measures for geographic units of different population sizes. If I had more time, I would have used race alone or in combination rather than race alone. There is a good deal more that could be done to provide state-specific data.

It is unfortunate that the time limitations mean the Census Bureau will not receive the quality of feedback they seek.

Appendix A Detailed data

Table A1. Distribution of Numerical Errors for Unified School Districts by Race and Hispanic Origin

	All Children	Black Alone Children	Hispanic Children	Asian Alone Children
Number of units in analysis	10,882	9,891	10,714	9,198
Number of units with absolute numeric errors of 100+ persons	494	7	207	89
Percent of units with absolute numeric errors of 100+ persons	4.5	0.1	18.8	1.0
Number of units with absolute numeric errors of 50+ persons	1,429	86	646	47
Percent of units with absolute numeric errors of 50+ persons	13.1	0.9	3.3	0.5
Number of units with absolute numeric errors of 25+ persons	3,638	508	1,857	315
Percent of units with absolute numeric errors of 25+ persons	33.3	5.1	17.3	3.4
Number of units with absolute numeric errors of 10+ persons	7,159	2,526	5,081	1,715
Percent of units with absolute numeric errors of 10+ persons	65.8	25.5	47.4	18.6
Source: Authors analysis of data released by the Census Bureau on April 28, 2021				

Table A2 Distribution of Absolute Percent Errors for Unified School Districts by Race and Hispanic Origin

	All Children	Black Alone Children	Hispanic Children	Asian Alone Children
Number of units in analysis	10,882	9,891	10,714	9,198
Number of units with absolute numeric errors of 25% or more	90	3704	2463	4311
Percent of units with absolute numeric errors of 25% or more	0.8	37.4	23	46.9
Number of units with absolute numeric errors of 10% or more	299	5,222	4,732	5,797
Percent of units with absolute numeric errors of 10% or more	2.7	52.8	44	63.0
Number of units with absolute numeric errors of 5% or more	844	6,170	6,402	6,678
Percent of units with absolute numeric errors of 5% or more	7.8	62.4	60	72.6
Source: Authors analysis of data released by the Census Bureau on April 28, 2021				

Table A3. Summary Table Absolute Numeric Errors * for All Children for Places

	Ages 0 to 17
Number of Units in the Analysis	29,111
Mean Size of District (Total Population)	7,847
Mean Absolute Numeric Error**	25.9
Mean Absolute Percent Error	12.9
Number of Units with Absolute Numerical Errors of 100+ persons	1,245
Percent of Units with Absolute Numerical Errors of 100+ persons	4.3
Number of Units with Absolute Numerical errors of 50+ persons	3,113
Percent of Units with Absolute Numerical errors of 50+ persons	10.7
Number of Units with Absolute Numerical errors of 25+ persons	7,311
Percent of Units with Absolute Numerical errors of 25+ persons	25.1
Number of Units with Absolute Numerical errors of 10+ persons	16,175
Percent of Units with Absolute Numerical errors of 10+ persons	55.6

Source: Author's analysis of data released by the Census Bureau on April 28, 2021.

Does not include Puerto Rico or geographic units with zero population age 0 to 17 in 2010 Summary File

* Error is defined here as the difference between the data as reported by the Census respondents and the data after the application of DP.

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

Table A4. Summary Table Absolute Percent Error* for All Children Ages 0 to 17 for Places

	Ages 0 to 17
Number of Units in the Analysis	29,111
Mean Size of Hispanic Population in Districts (all ages)	7,847
Mean Absolute Numeric Error**	26
Mean Absolute Percent Error	12.9
Number of Units with errors of 25% or more	3,070
Percent of Units with errors of 25% or more	10.5
Number of Units with errors of 10% or more	7,666
Percent of Units with errors of 10% or more	26.3
Number of Units with errors of 5% or more	12,741
Percent of Units with errors of 5% or more	43.8

Source: Authors analysis of data released by the Census Bureau on April 26, 2021.

Does not include Puerto Rico or geographic units with zero population age 0 to 17 in 2010 Summary File

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

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

Appendix B Background

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 that 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. For a short review of methods used in recent census see Abowd (2021). 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.

Analysis of earlier releases of DP Demonstration Products found that DP injected unacceptably large errors in the Census data. The Census Burau has released four previous DP Demonstration Products (October 2019, May 2020, September 2020, and November 2020). While many analysts have looked at the fitness for use of data from those files, their results are not relevant. Somewhat belatedly, the Census Bureau announced that they purposefully used a low level of epsilon for those files, which would lead to a high level of privacy protection and a poor level of accuracy. This paper focuses on the DP demonstration product released by the Census Bureau on April 28, 2021, which the Bureau says is set for the level of epsilon needed for the redistricting data.

The application of differential privacy allows the Census Bureau to control the amount of error injected into the data which is largely controlled by something called "Epsilon." A higher-level epsilon means less error and more risk of violating confidentiality and a lower epsilon means more error and less risk of violating confidentiality. The previous demonstration files produced by the Census Bureau had Epsilons of 4 for population and 2 for housing.

The epsilon used in the April 2021 DP demonstration product file (12.3) is much higher than that used in previous DP demonstration products (6.0) and similar to the level they expect to use for the PL 94-171 redistricting files. That means the DP infused data released in April 2021 should be more accurate than previous DP Demonstration files.

In October 2019, the U.S. Census Bureau (2019) released what they call a "Demonstration Product" which applied DP to 2010 Census data to produce a new file

or set of tables. This file was released to the public so researchers could assess the impact of DP on census accuracy.

The National Academy of Sciences, Committee on National Statistics Workshop held December 11-12, 2019, titled. “Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations” provides a lot of data related to the accuracy of the Census Bureau’s October 2019 Demonstration Product including several presentations focused on children (Committee on National Statistics 2019). A written summary of the workshop is available by two of the CNSTAT Workshop organizers (Hotz and Salvo 2020).

Based on the evidence presented at the CNSTAT workshop and their own internal analysis the U.S. Census Bureau (2020b) concluded, “The October Vintage of the DAS falls short of ensuring ‘Fitness for use’ for several priority use cases.” This led to subsequent versions of DP-infused data being released by the Census Bureau.

Analysis of more recent data released by the U.S. Census Bureau continue to indicate the implementation of DP is likely to produce unacceptable results for young children. On May 27, 2020, the U.S. Census Bureau provided a revised application of differential privacy to the 2010 Census data on young children. Based on perusal of the U.S. Census Bureau website related to DP (Census Burearu 2020e) it appears that there will be no more demonstration files released to the public allowing assessments of the potential impact of DP on data for young children before DP is implemented in the 2020 Census. Thus, the demonstration file released by the U.S. Census Bureau on May 27, 2020, is the best data available to understand the implications of DP for data on young children in the 2020 Census.

The DP demonstration file released on April 28, 2021, supersedes the previous files. The DP demonstration product released by the Census Burau on April 28th, 2021, has parameters that are similar to what the Census Bureau plans to use in the redistricting data the Census Bureau will produce for the public. The file released on April 28, 2021, has an Epsilon of 10.3 for population and 2 for housing for a total of 12.3. This is a much higher epsilon value than has been used in previous DP demonstration products, meaning the data should be more accurate than previous DP demonstration files.

Appendix C Impossible or improbable results

Table C. Impossible or Implausibe Block Level Data after DP is Applied to 2010 Census Data

State	Blocks changed from greater than 50% Non-White Hispanic to less than 50% Non-White Hispanic		Blocks with population in Summary File 1 but no popualtion in DP file		Blocks with population in households but no occupied housing units		Blocks with occupied housing units but no population		Blocks with more than 15 persons per household	
Alabama	5,908	2.3	1,117	0.4	17,684	7.0	1,338	0.5	18,492	7.3
Alaska	1,004	2.2	116	0.3	2,299	5.1	207	0.5	2,389	5.3
Arizona	6,024	2.5	929	0.4	16,261	6.7	1,492	0.6	17,471	7.2
Arkansas	3,948	2.1	1,050	0.6	13,404	7.2	1,386	0.7	13,925	7.5
California	24,216	3.4	899	0.1	29,662	4.2	1,210	0.2	33,434	4.7
Colorado	5,596	2.8	1,055	0.5	12,148	6.0	1,588	0.8	12,709	6.3
Connecticut	1,962	2.9	72	0.1	3,253	4.8	90	0.1	3,606	5.3
Delaware	795	3.3	41	0.2	1,694	7.0	51	0.2	1,816	7.5
District of Columbia	101	1.6	-	-	197	3.0	-	-	235	3.6
Florida	14,441	3.0	802	0.2	34,911	7.2	996	0.2	38,061	7.9
Georgia	8,878	3.1	826	0.3	21,097	7.3	1,142	0.4	22,521	7.7
Hawaii	935	3.7	7	0.0	977	3.9	5	0.0	1,167	4.7
Idaho	2,499	1.7	838	0.6	7,235	4.8	1,219	0.8	7,488	5.0
Illinois	7,736	1.7	2,780	0.6	21,188	4.7	3,361	0.7	23,010	5.1
Indiana	4,198	1.6	1,420	0.5	16,859	6.3	1,722	0.6	17,839	6.7
Iowa	2,085	1.0	2,977	1.4	11,670	5.4	3,635	1.7	12,094	5.6
Kansas	3,751	1.6	3,524	1.5	14,156	5.9	4,741	2.0	14,588	6.1
Kentucky	2,313	1.4	718	0.4	10,545	6.5	909	0.6	11,132	6.9
Louisiana	4,139	2.0	532	0.3	11,254	5.5	716	0.4	12,017	5.9
Maine	633	0.9	388	0.6	7,149	10.3	850	1.2	7,415	10.7
Maryland	4,091	2.8	352	0.2	7,096	4.9	449	0.3	7,702	5.3
Massachusetts	2,973	1.9	217	0.1	7,895	5.0	278	0.2	8,542	5.4
Michigan	4,899	1.5	1,319	0.4	26,715	8.1	2,104	0.6	28,381	8.6
Minnesota	2,802	1.1	2,571	1.0	16,265	6.3	3,278	1.3	16,899	6.5
Mississippi	3,772	2.2	849	0.5	10,732	6.3	1,073	0.6	11,284	6.6
Missouri	4,722	1.4	2,810	0.8	23,140	6.7	3,738	1.1	24,070	7.0
Montana	1,336	1.0	1,230	0.9	7,946	6.0	1,878	1.4	8,109	6.1
Nebraska	2,056	1.1	3,455	1.8	10,647	5.5	4,350	2.3	10,875	5.6
Nevada	2,369	2.8	183	0.2	3,574	4.2	209	0.3	3,953	4.7
New Hampshire	623	1.3	238	0.5	4,189	8.6	414	0.9	4,384	9.0
New Jersey	4,174	2.5	127	0.1	7,471	4.4	143	0.1	8,384	4.9
New Mexico	4,964	2.9	733	0.4	9,281	5.5	1,045	0.6	9,635	5.7
New York	6,450	1.8	1,150	0.3	23,582	6.7	1,890	0.5	25,215	7.2
North Carolina	9,839	3.4	815	0.3	23,919	8.3	1,139	0.4	25,262	8.7
North Dakota	796	0.6	2,693	2.0	7,547	5.6	3,772	2.8	7,607	5.7
Ohio	5,443	1.5	1,665	0.5	20,159	5.5	2,038	0.6	21,691	5.9
Oklahoma	10,132	3.8	1,739	0.7	18,519	6.9	2,327	0.9	19,194	7.1
Oregon	4,176	2.1	625	0.3	8,592	4.4	862	0.4	9,082	4.6
Pennsylvania	6,186	1.5	1,785	0.4	26,037	6.2	2,579	0.6	27,792	6.6
Rhode Island	404	1.6	19	0.1	1,353	5.4	32	0.1	1,495	5.9
South Carolina	5,712	3.1	594	0.3	14,134	7.8	721	0.4	14,852	8.2
South Dakota	959	1.1	1,668	1.9	5,805	6.6	2,164	2.5	5,914	6.7
Tennessee	4,617	1.9	1,131	0.5	16,392	6.8	1,483	0.6	17,355	7.2
Texas	25,404	2.8	3,003	0.3	51,816	5.7	4,296	0.5	55,309	6.1
Utah	1,434	1.2	290	0.3	4,816	4.2	570	0.5	5,168	4.5
Vermont	373	1.1	230	0.7	3,281	10.1	335	1.0	3,381	10.4
Virginia	7,785	2.7	1,255	0.4	16,789	5.9	1,682	0.6	17,722	6.2
Washington	6,442	3.3	553	0.3	10,949	5.6	710	0.4	11,773	6.0
West Virginia	1,387	1.0	1,161	0.9	10,241	7.6	1,577	1.2	10,541	7.8
Wisconsin	2,926	1.2	1,530	0.6	18,221	7.2	2,223	0.9	19,014	7.5
Wyoming	891	1.0	580	0.7	3,852	4.5	875	1.0	3,935	4.6
Total	241,299		56,661		674,598		76,892		715,929	

Source: Authors analysis of data provided by David Van Riper of IPUMS at the University of Minnesota.

References

Abowd. J. (2021) declaration in suit brought by Alabama.

Bouk, D. and Boyd, D. (2021). "Democracy's Data Infrastructure.; The technologies of the U.S. Census."

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/ .

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 and Stojakovic, Z. (2021). The Importance of Small Area Census Data for School Demographics, Count all Kid website

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/ ,

McKibben and Cropper (2021)

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.

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 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) .

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

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., March 18, 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, March 26.

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 - 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. 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 .