Analysis of Census Bureau's August 2021 Differential Privacy Demonstration Product: Implications for Data on Children[1]

By

Dr. William P. O'Hare

November 2021

[1] This research was supported by the Annie E. Casey Foundation. The findings and conclusions in this report are those of the author(s) alone and do not necessarily reflect the opinions of the Casey Foundation.

Contents

Tables and Figures

Analysis of Census Bureau's August 2021 Differential Privacy Demonstration Product: Implications for Data on 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 confidentiality and privacy of respondents in the 2020 Census. 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, which was released by the Census Bureau on August 12, 2021. The DP demonstration product issued in August 2021 supersedes earlier DP demonstration products and uses the same DP parameters as those used in the production of the redistricting data (this file is sometimes called the Public Law - PL - 94-171 file) also issued by the Census Bureau on August 12, 2021.

To be clear, decisions about the use of DP in the redistricting data have been made so this analysis will have no impact on those decisions or that data. In June of 2021, Census Bureau leadership determined the final accuracy parameters for the redistricting data (P.L. 94-171) after a long process which included a series of demonstration products and user feedback.

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 adding additional error 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. After release of the April 2021 demonstration product, many stakeholders, including the Count All Kids Campaign, urged the Census Bureau to make data more accurate. Subsequently the Census Bureau responded to those requests and set epsilon at 19.61 for the redistricting file to increase the accuracy of the figures.

The August 2021, demonstration product file applied DP to 2010 Census data thus allowing comparisons of data with and without DP. 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 related to the number of children. For example, based on analysis of the August 2021 DP file, the count of children would exhibit absolute percent error of 5 percent or more in about 7 percent of Unified School Districts after DP is applied. Bigger absolute error percentages are evident for several minority child populations. The data also show than 61 percent of Unified School Districts had absolute numeric errors of 10 or more children and 7 percent of Unified School Districts have errors of 50 or more children. Errors of this magnitude could have implications for federal and state funding received by schools and for educational planning. More analysis is needed on this point.

The injection of DP, in the 2010 Census data included in the August Demonstration Product, resulted in there being are over 160,000 blocks nationwide that had population ages 0 to 17, but no population ages 18 or over. In the data without DP injected, there

were only a few hundred such blocks nationwide. Blocks with children and no adults is a highly implausible situation and the large number of such blocks in the 2020 Census 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 of children from parents in the data processing is an ongoing concern for data on children.

The negative implications of DP for small areas and small populations are important because DP will be used in the remaining 2020 Census data files including Demographic Profiles file, the Demographic and Housing Characteristic (DHC) file, and the Detailed-Demographic and Housing Characteristics (D-DHC) file. Those files will provide data for smaller population groups, such as children ages 0 to 4 by race. Given the larger impact of DP for smaller groups, it is important to monitor the quality of data for children in future 2020 Census files.

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 intended to help stakeholders gain a better understanding of the implications of DP for 2020 Census data on children and enable stakeholders to use 2020 Census data responsibly.

There are a couple of reasons for sharing this information with child advocates now. First, 2020 Census results for some localities may include situations where the number of young children reported looks suspect. It is important to make sure child advocates are aware of the potential impact of DP so they can explain odd child statistics to local leaders.

There is a second reason for sharing this information with state and local child advocates. The U.S. Census Bureau is still looking for feedback on the use of DP in the 2020 Census. In particular, they are looking for cases where census data are used to make decisions. The Census Bureau is asking data users to examine the DP demonstration products to see if the error injected by DP make the data unfit for their use case.

There is some latitude in how much error the Census Bureau injects into the data for future products, 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 in future 2020 Census products.

If readers know of situations where census data are used for decision-making, they should notify the Census Bureau. General information is fine, but information about what specific demographic characteristic(s) are used at what geographic level is even better. Thoughts and reactions to the use of DP in 2020 Census can be sent to 2020DAS@census.gov.

Analysis of Census Bureau's August 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.[2] This paper uses key metrics to assess the accuracy of census data for children (population ages 0 to 17) after DP is injected. The report focuses on the level of errors injected into the Census data for children based on the most recent DP demonstration product data available from the Census Bureau.

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 intended to help stakeholders gain a better understanding of the implications of DP for 2020 Census data on children and enable stakeholders to use 2020 Census data responsibly.

[2] The terminology in this arena can be confusing. Differential privacy is sometimes called "formal privacy." The specific 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.

In short, DP injects random 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 of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). More information about the DP issue and recent developments are provided in Appendix A. The Census Bureau (2021) recently released a report to help data users understand the impact of differential privacy on the 2020 Census data. For an independent look at differential privacy see Boyd (2020) as well as Bouk and Boyd (2021).

The Census Bureau provided some suggested accuracy metrics with the demonstration product release on the April 28, 2021, 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.

There may be some hesitation to use the analysis of the redistricting data to infer how accurate later 2020 Census products will be. However, it is still unclear if the data in later files will 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 could potentially 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. The Census Bureau has indicated it hopes to engage stakeholders in decisions about what data to include and privacy parameters for those subsequent files. In the current plan, the

Census Bureau plans to start the production process for the DHC files in the summer of 2022, after releasing a couple of demonstration products and getting user feedback.

This paper updates a similar one which analyzed the Demonstration Product released by the Census Bureau in April 2021 (O'Hare 2021). Comparing the results of this report to the previous one is mostly good news, there was a lot of improvement between the April 2021 Demonstration Product and the August 2021 Demonstration Product, but there are also a few areas of concern that are addressed in this report. The improvement is related to a large increase in the epsilon value used for the redistricting data compared to the April 2021 DP file. Following release of the April 2021 DP file, the Census Bureau was responsive to many stakeholders, including the Count All Kids Campaign, calling for more accuracy in the data. The Census Bureau set the final epsilon level at 19.61 compared to 12 for the April 2021 file.

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 often have policymaking authority, and they often provide programs for children. Blocks are the most basic building block for census data. 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.” There is some consensus that accuracy must incorporate a variety of measures. 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 metrics can be examined at (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?#%20)

For simplicity, I only look at a few key measures of accuracy here, but 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 (https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/2020-disclosure-avoidance-system-update.html).

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, which are often called “outliers.” I use two sets of benchmarks to identify large errors: one set for absolute numeric errors and one set for absolute percent errors. Absolute percent error categories are:

- 5 percent or more,
- 10 percent or more, and
- 25 percent or more.

For absolute numeric errors, the categories are:

- 5 or more children,
- 10 or more children,
- 25 or more children, and
- 100 or more children.

I believe the number and percent of large errors are likely to be the most important measures of accuracy in the 2020 Census. Large errors are likely to be a statistical problem and a public relationship problem for the Census Bureau, particularly if they are accompanied by large swings in census-related funding.

Data Used in This Study

As stated earlier in this paper, the DP demonstration file released by the Census Bureau on August 12, 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 August 2021 DP demonstration file provides data for all census geographic levels including the smallest unit (census blocks). The Census Bureau has released five previous demonstration products (October 2019, May 2020, September 2020, November 2020, and April 2021). The DP demonstration file released in August 2021, supersedes the previous files.

The Census Bureau file released in August 2021 does not provide data for the population age 0 to 17 directly, but it does provide data for 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.

For the files analyzed here, the data for Puerto Rico was removed and any geographic units with no children in the 2010 Summary File were removed from the files used for analyses (except for block analysis where I had to use data that was already tabulated).

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 less than 5,000 children, where DP may inject enough error to be problematic. For this subset of counties, DP may distort the data to a problematic degree as shown by O'Hare (2019). Based on analysis of the August 2021 Demonstration File, for 1,377 counties with less than 5,000 children, the mean absolute numeric error was 8 and the mean absolute percent error was 0.6.

The situation is different for Unified School Districts and Places, where DP is likely to cause substantial distortions for the child population for some areas. 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	3,143	10,882	29,111
Mean Size of District (Child Population based on Summary File)	1,454,539	23,602	6,817	1,893
Mean Absolute Numeric Error**	15	10	20	7
Mean Absolute Percent Error	0.003	0.32	2	4.3
Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 12, 2021. Data from IPUMS NHGIS, University of Minnesota www.nhgis.org				
Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 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.				

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). At the Committee on National Statistics of the National Academy of Sciences 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). However, keep in mind that these 2019 analyses are based on a very low level of epsilon that was used in the Demonstration Product released in October 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). When school district boundaries change during the decade, the Census Bureau must go back to individual block data to re-constitute the new district lines for the Small Area Income and Poverty Estimates (SAIPE).

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, on small populations and the majority of Unified School Districts are relatively small. Out of 10,882 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 for ages 0 to 17 of less than 100, and 1,910 districts had population for 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. Race/ethnic groups are smaller than the total and they typically have higher percentage errors.

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 White

children, Black children, Hispanic children, and Asian children.[3] For the remainder of this report when I use the term White, Black, or Asian, it means White alone, Black alone or Asian alone. Other race groups were not examined here because the numbers were small and might produce unreliable estimates.

Data in Table 2 show the vast majority of Unified School Districts have at least one White children, one Black child, one Hispanic child, and one Asian child. But many districts have very small numbers of minority 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.

Table 2. Distribution of Numerical Errors* for Children (population ages 0 to 17) in Unified School Districts by Race and Hispanic Origin)

	All Children	White**	Hispanic	Black**	Asian**
Number of units in analysis	10,882	10,880	10,713	9,890	9,197
Mean Number of Children in Districts (In group at column heading)	6,817	4,450	1,599.0	1,044	322
Mean Absolute Numeric Error	20	14	11	7	5
Mean Absolute Percent Error	2	2	17	31	40

Source: Authors analysis of data released by the Census Bureau on August 12, 2021

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

** These categories use the race alone definition .

School Districts with zero children no included in analysis

[3] 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.

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 20 children. In other words, for the average unified school district the DP-infused data differs from the data without DP by 20 children. The mean absolute numeric errors for White (14), Hispanic (11), Black (7), and Asian (5) children are smaller than that for all children (20).

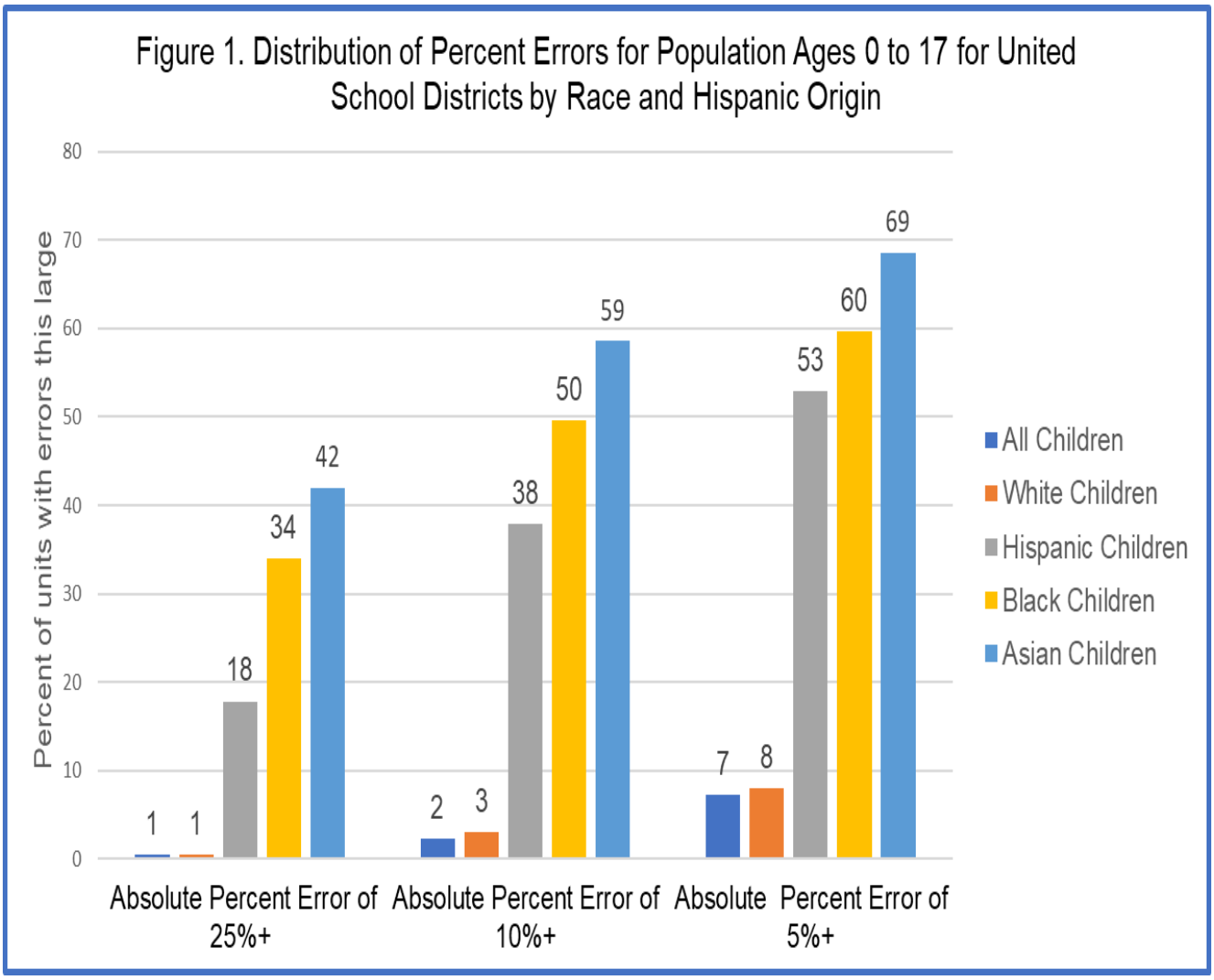
The mean absolute percent error shown in Table 2 for all children is 2 percent. For White children the mean absolute percent error is 2, for Hispanic children it is 17, for Black children, the mean absolute percent error is 31, and for Asian children it is 40. It is worth noting that for many groups there were large numbers of school districts with very small numbers of minority children and that is likely to lead to a lot of high percent errors.

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 error 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

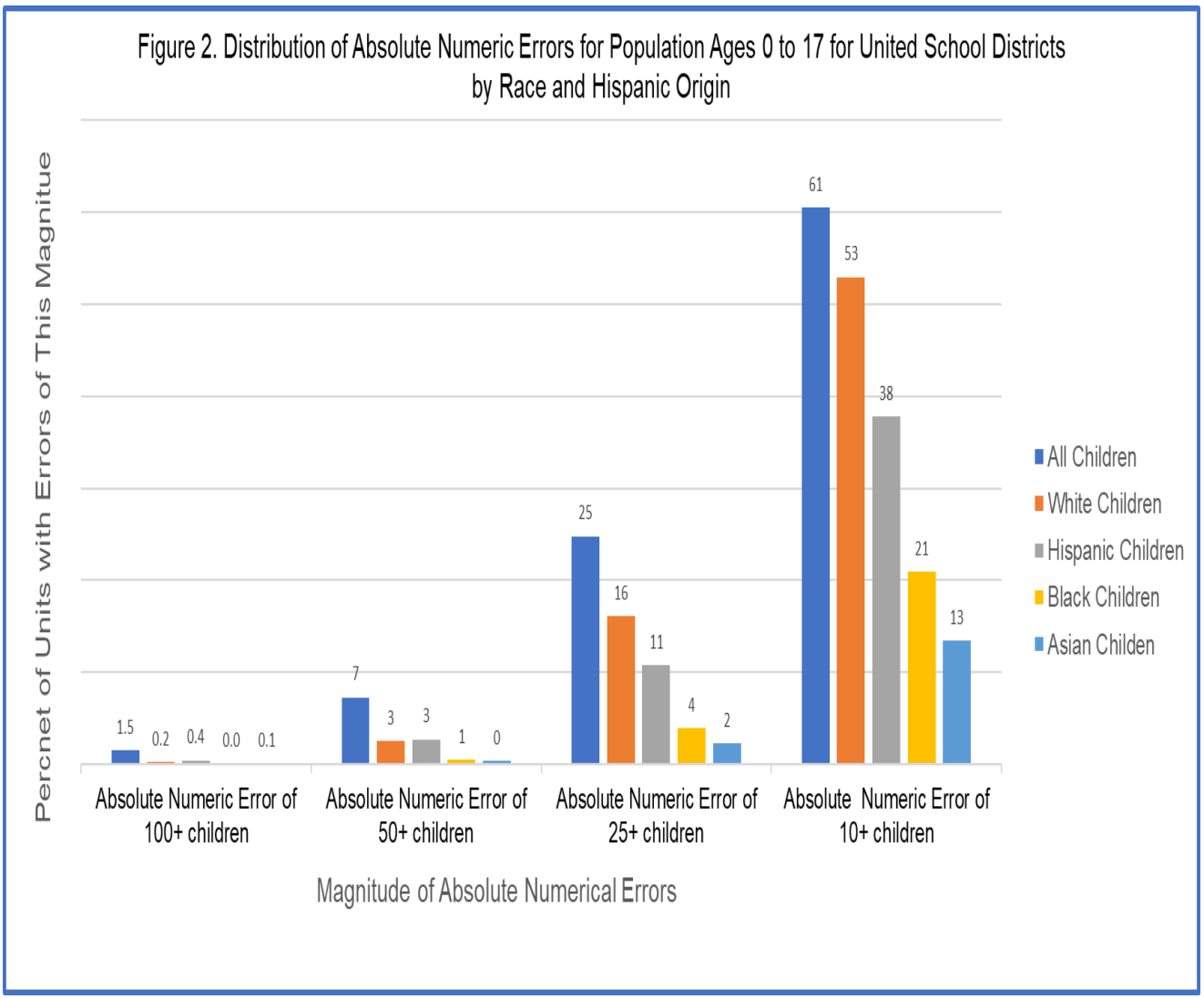
more than 10 percent error and more than 5 percent error. These thresholds are judgmental, but they provide a reasonable range of errors.

The 5 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.



The distributions of absolute percent errors are shown in Figure 1. For all children, 7 percent of School Districts had absolute percent errors of 5 percent or more compared to 8 percent of White children, 60 percent for Black children, 53 percent for Hispanic children, and 69 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 also shows that for minority children, absolute percent errors of 25 percent or more are relatively common,

Figure 2 shows 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 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.

On the other hand, there are relatively few Unified School Districts with very large numeric errors. Only 2 percent of Unified School Districts has errors of 100 child or

more, compared to 0.2 percent for Black children, 1 percent of Hispanic children and 1 percent for Asian children. (Zero may include rounds to zero.)

Figure 2 shows for all children (all races) 61 percent of the Unified School Districts had errors of 10 children or more, and the figure for White children is 53 percent but the figures for minority groups are smaller (21 percent for Black children, 38 percent for Hispanic children and 13 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 6 for Maine and Vermont to a high of 54 for California. In other words, this error measure in California is almost ten times what it is in Vermont and Maine. The mean absolute percent error ranges from a low of 0 for Hawaii and DC to a high of 4.7 for Vermont.

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

Rank*		Mean Absolute Numeric Error		Rank *		Mean Absolute Percent Error
1	California	54		1	Vermont	4.7
2	Hawaii	44		2	Maine	4.5
3	Arizona	34		3	Montana	4.5
4	Delaware	31		4	North Dakota	4.1
5	Texas	25		5	Idaho	4.0
6	DC	25		6	Oregon	3.6
7	Washington	25		7	South Dakota	3.4
8	South Carolina	24		8	New Mexico	3.3
9	Michigan	24		9	Oklahoma	2.9
10	New York	23		10	Washington	2.9
11	Illinois	22		11	Nebraska	2.7
12	Utah	21		12	Texas	2.7
13	Connecticut	21		13	Colorado	2.7
14	Ohio	21		14	Kansas	2.2
15	Oregon	20		15	New Hampshire	2.2
16	Oklahoma	20		16	Alaska	2.1
17	Missouri	20		17	Missouri	1.9
18	New Jersey	19		18	Wyoming	1.9
19	Indiana	19		19	Iowa	1.8
20	Arkansas	19		20	South Carolina	1.5
21	Colorado	19		21	Minnesota	1.5
22	Wisconsin	19		22	New Jersey	1.5
23	Idaho	18		23	New York	1.5
24	New Mexico	18		24	Ohio	1.4
25	Mississippi	18		25	Illinois	1.4
26	Rhode Island	17		26	Arkansas	1.3
27	Minnesota	17		27	Indiana	1.2
28	North Carolina	17		28	Wisconsin	1.2
29	Tennessee	17		29	Arizona	1.2
30	Massachusetts	16		30	California	1.0
31	Kansas	16		31	Michigan	1.0
32	Alabama	16		32	Nevada	0.6
33	Iowa	15		33	Mississippi	0.6
34	Wyoming	15		34	Kentucky	0.6
35	Pennsylvania	14		35	Rhode Island	0.4
36	Georgia	14		36	Pennsylvania	0.4
37	South Dakota	14		37	Massachusetts	0.4
38	Nebraska	14		38	Tennessee	0.4
39	Florida	13		39	Alabama	0.3
40	North Dakota	12		40	Connecticut	0.3
41	Louisiana	12		41	Utah	0.3
42	Nevada	11		42	Virginia	0.3
43	Kentucky	11		43	Georgia	0.3
44	New Hampshire	10		44	Delaware	0.3
45	Alaska	10		45	North Carolina	0.3
46	Montana	9		46	West Virginia	0.2
47	Virginia	9		47	Louisiana	0.2
48	West Virginia	9		48	Florida	0.1
49	Maryland	9		49	Maryland	0.1
50	Vermont	6		50	DC	0.0
51	Maine	6		51	Hawaii	0.0
	U. S. Total	20			U.S. Total	1.7

Source: Authors analysis of DP Demonstration Product Released by the Census Bureau on August 12, 2021

* Ranks based on unrounded data.

The problems that are likely to be caused by inaccurate census data on children are relatively clear. When the number of 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 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.

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

Table 4. 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-	

Table 4 only reflects federal spending, but 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.

Data for Places

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

Places include both incorporated Places and Census Designated Places (CDPs). There are a little more than 29,000 Places for which the infusion of DP data was produced in the August 12, 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, childcare facilities, and health care centers.

The absolute mean numeric error for Places is 7 children and the mean absolute percent error is 4.3. Many of these Places are relatively 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 18 percent of Places had errors of 5 percent or more for the child population and nearly one out of ten (9 percent) had errors of 10 percent or more. Only 3 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 percentage errors are larger than for Unified School Districts.

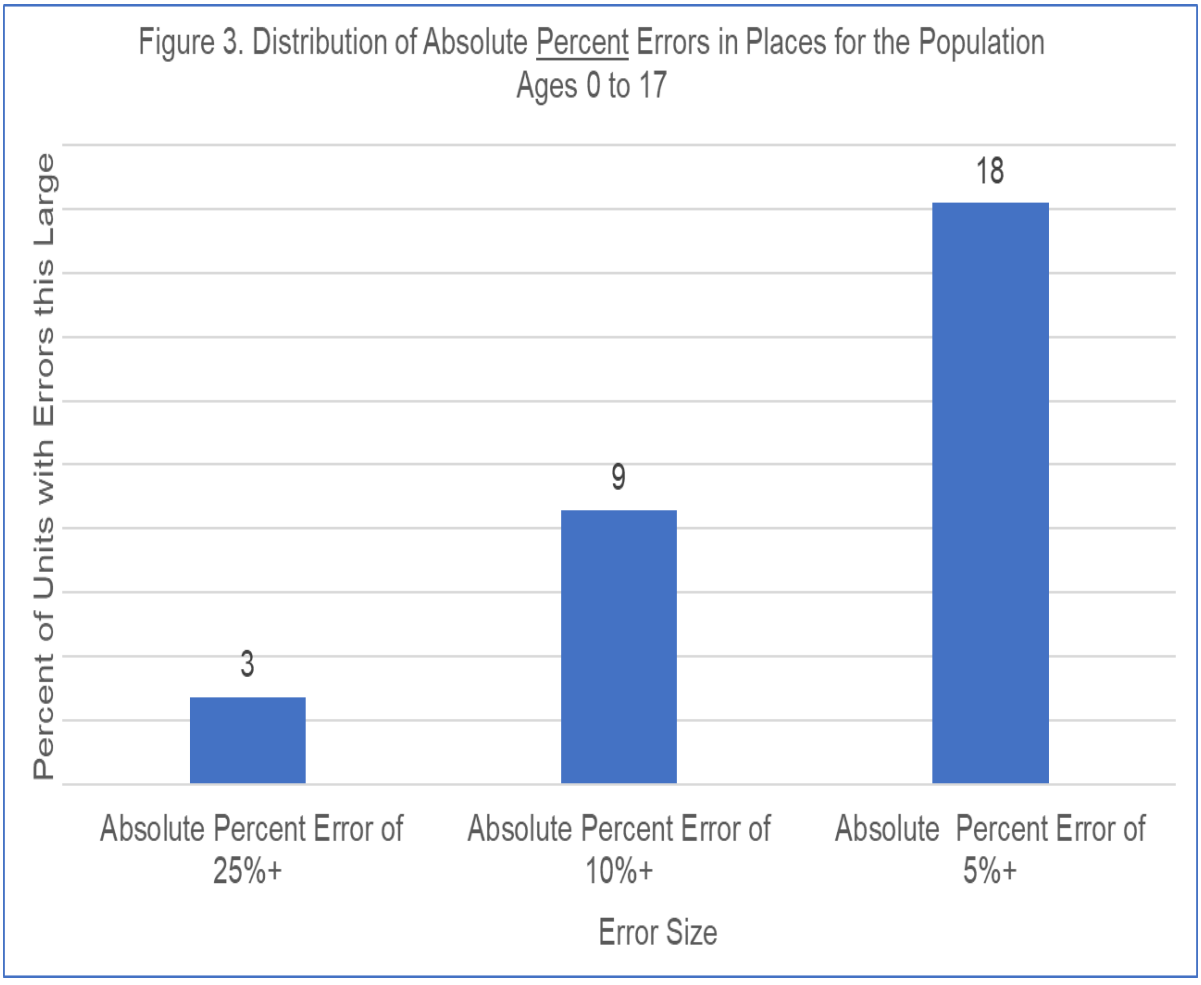


Figure 4 shows the distribution of Places by absolute numeric errors using the same categories as Figure 2. Over one-fifth (21 percent) of the Places had errors of 10 or more children but only 1 percent of Places had absolute numeric errors of 50 or more children. Very few places have numerical errors of 100+ children

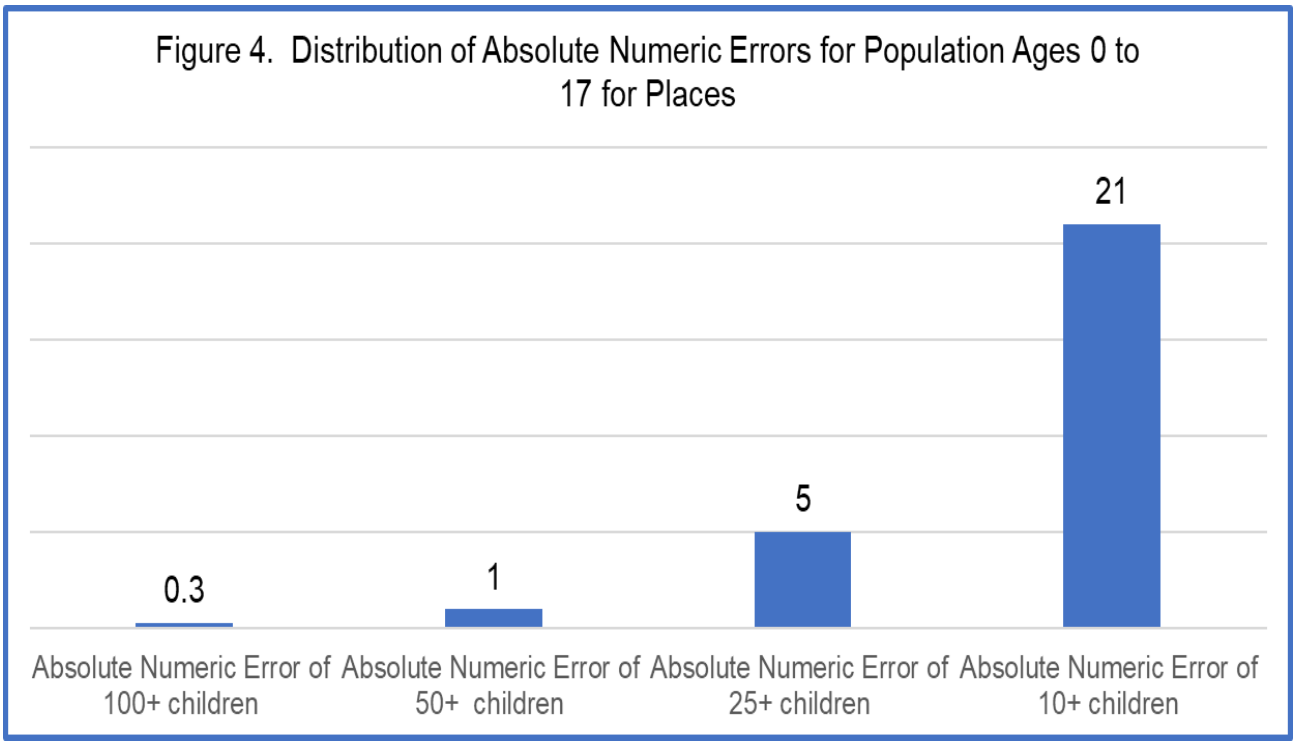


Table 5 provides the distribution of absolute numeric errors for each state using the same categories as used in Figure 4. For the categories of error of 100+ children almost all the states are zero or round to zero.

Table 5. Distribution of 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	494	69	12	2	0		86	12	2	0	0
Alaska	338	309	22	7	0	0		91	7	2	0	0
Arizona	445	355	55	20	10	5		80	12	4	2	1
Arkansas	541	485	47	5	3	1		90	9	1	1	0
California	1,505	942	327	149	52	35		63	22	10	3	2
Colorado	453	364	72	9	7	1		80	16	2	2	0
Connecticut	142	56	56	22	6	2		39	39	15	4	1
Delaware	76	60	14	2	0			79	18	3	0	0
DC	1	-	0	1	0	0		0	0	100	0	0
Florida	915	570	234	82	27	2		62	26	9	3	0
Georgia	623	502	97	19	4	1		81	16	3	1	0
Hawaii	151	84	54	9	2	2		56	36	6	1	1
Idaho	224	189	29	5	1	0		84	13	2	0	0
Illinois	1,367	1,121	191	39	15	1		82	14	3	1	0
Indiana	681	582	76	17	6	0		85	11	2	1	0
Iowa	1,008	948	50	10	0	0		94	5	1	0	0
Kansas	668	610	48	9	1	0		91	7	1	0	0
Kentucky	523	474	40	7	2	0		91	8	1	0	0
Louisiana	473	390	71	11	1	0		82	15	2	0	0
Maine	131	68	52	11	0	0		52	40	8	0	0
Maryland	518	374	108	30	6	0		72	21	6	1	0
Massachusetts	243	112	80	41	8	2		46	33	17	3	1
Michigan	691	418	218	38	12	5		60	32	5	2	1
Minnesota	902	773	97	22	10	0		86	11	2	1	0
Mississippi	362	318	38	5	1	0		88	10	1	0	0
Missouri	1,021	907	92	18	3	1		89	9	2	0	0
Montana	361	341	19	1	0	0		94	5	0	0	0
Nebraska	577	535	35	6	1	0		93	6	1	0	0
Nevada	128	101	19	4	4	0		79	15	3	3	0
New Hampshire	96	49	38	9	0	0		51	40	9	0	0
New Jersey	542	285	167	74	14	2		53	31	14	3	0
New Mexico	440	387	45	7	0	1		88	10	2	0	0
New York	1,187	600	422	138	26	1		51	36	12	2	0
North Carolina	738	593	127	13	4	1		80	17	2	1	0
North Dakota	391	379	11	1	0	0		97	3	0	0	0
Ohio	1,204	993	165	37	7	2		82	14	3	1	0
Oklahoma	727	630	80	15	1	1		87	11	2	0	0
Oregon	375	290	63	17	5	0		77	17	5	1	0
Pennsylvania	1,759	1,292	384	73	9	1		73	22	4	1	0
Rhode Island	34	13	14	5	2	0		38	41	15	6	0
South Carolina	395	327	53	13	2	0		83	13	3	1	0
South Dakota	376	357	17	2	0	0		95	5	1	0	0
Tennessee	428	345	63	15	4	1		81	15	4	1	0
Texas	1,745	1,425	247	49	18	6		82	14	3	1	0
Utah	323	254	45	19	3	2		79	14	6	1	1
Vermont	119	76	38	5	0	0		64	32	4	0	0
Virginia	591	469	95	26	1			79	16	4	0	0
Washington	627	467	119	33	4	4		74	19	5	1	1
West Virginia	400	379	20	1	0	0		95	5	0	0	0
Wisconsin	772	617	123	22	9	1		80	16	3	1	0
Wyoming	197	181	12	2	2	0		92	6	1	1	0
U.S. Total	29,111	22,890	4,658	1,187	295	81		79	16	4	1	0

Source: Authors analysis of data released by the Census Bureau on August 12, 2021

Does not Include Puerto Rico

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

Table 6 provides the distribution of Places by 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, 12 percent of the Places in North Dakota had absolute percent errors of 25 percent or more, compared to 0 in several states (in some of these states the rate rounds to zero).

Table 6. Distribution of Percent Errors for Places by State 11-9-2021										
		Number of Places with Errors of this size					Percent Distribution within the State			
	Number of Places	less than 5 percent	5 to 9.9 percent	10 to 24.9 percent	25% or more		less than 5 percent	5 to 9.9 percent	10 to 24.9 percent	more than 25%
Alabama	577	492	51	24	10		85	9	4	2
Alaska	338	223	43	36	36		66	13	11	11
Arizona	445	354	38	25	28		80	9	6	6
Arkansas	541	445	60	28	8		82	11	5	1
California	1505	1290	100	67	48		86	7	4	3
Colorado	453	349	43	38	23		77	9	8	5
Connecticut	142	118	17	4	3		83	12	3	2
Delaware	76	64	7	3	2		84	9	4	3
DC	1	1	0	0	0		100	0	0	0
Florida	915	839	35	26	15		92	4	3	2
Georgia	623	552	50	18	3		89	8	3	0
Hawaii	151	137	6	4	4		91	4	3	3
Idaho	224	190	16	9	9		85	7	4	4
Illinois	1367	1218	101	36	12		89	7	3	1
Indiana	681	612	44	15	10		90	6	2	1
Iowa	1008	781	117	79	31		77	12	8	3
Kansas	668	487	92	65	24		73	14	10	4
Kentucky	523	448	52	19	4		86	10	4	1
Louisiana	473	426	31	15	1		90	7	3	0
Maine	131	97	27	7			74	21	5	0
Maryland	518	442	38	27	11		85	7	5	2
Massachusetts	243	215	16	8	4		88	7	3	2
Michigan	691	506	94	70	21		73	14	10	3
Minnesota	902	721	87	67	27		80	10	7	3
Mississippi	362	333	21	7	1		92	6	2	0
Missouri	1021	798	107	82	34		78	10	8	3
Montana	361	245	55	42	19		68	15	12	5
Nebraska	577	390	96	55	36		68	17	10	6
Nevada	128	92	16	11	9		72	13	9	7
New Hampshire	96	68	13	11	4		71	14	11	4
New Jersey	542	460	33	25	24		85	6	5	4
New Mexico	440	299	65	39	37		68	15	9	8
New York	1187	868	165	123	31		73	14	10	3
North Carolina	738	641	63	25	9		87	9	3	1
North Dakota	391	222	69	55	45		57	18	14	12
Ohio	1204	1097	76	26	5		91	6	2	0
Oklahoma	727	525	112	68	22		72	15	9	3
Oregon	375	292	47	25	11		78	13	7	3
Pennsylvania	1759	1396	184	135	44		79	10	8	3
Rhode Island	34	27	2	4	1		79	6	12	3
South Carolina	395	340	29	22	4		86	7	6	1
South Dakota	376	252	59	41	24		67	16	11	6
Tennessee	428	406	18	3	1		95	4	1	0
Texas	1745	1507	140	65	33		86	8	4	2
Utah	323	295	18	7	3		91	6	2	1
Vermont	119	67	32	17	3		56	27	14	3
Virginia	591	512	51	25	3		87	9	4	1
Washington	627	538	45	27	17		86	7	4	3
West Virginia	400	334	43	21	2		84	11	5	1
Wisconsin	772	626	68	46	32		81	9	6	4
Wyoming	197	145	25	14	13		74	13	7	7
U.S. Total	29345	24013		1712	801		82	10	6	3
Authors analysis of data released by the Census Bureau on August 12, 2021										
Does not Include Puerto Rico										
* In this paper, errors reflect the difference between the 2010 Census data with and without DP.										

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. One of the primary purposes of the Decennial Census is to provide comparable population figures for small areas across the country. 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. 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 geographic units. From this perspective, errors at the block level are not so important.

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 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 a large share of the distribution is at zero.

Three different means are presented in Table 7 which shows states ranked on the percent of blocks with absolute errors of 5 percent or more. Looking across all states, the mean absolute numeric error is 1.5 and the mean absolute percent error is 36. The state mean for blocks with absolute percent errors of more than 5 percent is 43 percent.

The data in Table 7 indicates significant variation across states. For example, 59 percent of blocks in Rhode Island have absolute percent errors of 5 percent or more compared to 21 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 7. 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 absolute percent errors of 5% or more
1	Rhode Island	2.3	40	59
2	Connecticut	2.5	39	58
3	New Jersey	2.5	35	58
4	New York	2.2	39	55
5	Pennsylvania	1.8	46	55
6	District Of Columbia	2.3	29	55
7	Delaware	2.1	43	55
8	Indiana	1.7	47	54
9	North Carolina	2.0	45	53
10	Ohio	1.7	42	52
11	Illinois	1.8	43	52
12	Florida	2.0	40	51
13	Michigan	1.7	39	50
14	Massachusetts	2.0	34	50
15	Washington	2.0	40	50
16	South Carolina	1.6	44	48
17	Wisconsin	1.5	41	47
18	Tennessee	1.5	41	47
19	Georgia	1.8	40	47
20	California	2.4	30	47
21	New Hampshire	1.5	41	46
22	Iowa	1.3	49	45
23	Maryland	1.8	34	45
24	Kentucky	1.4	38	44
25	Minnesota	1.4	41	44
26	Alabama	1.3	40	42
27	Colorado	1.6	36	42
28	Vermont	1.2	40	41
29	Missouri	1.2	39	40
30	Texas	1.6	32	40
31	Virginia	1.5	36	40
32	Arkansas	1.2	40	40
33	Oklahoma	1.3	42	40
34	Louisiana	1.3	32	40
35	Hawaii	2.7	28	39
36	Mississippi	1.2	36	39
37	Maine	1.1	37	39
38	Arizona	1.6	30	37
39	Kansas	1.1	40	37
40	West Virginia	1.0	39	36
41	Oregon	1.2	31	35
42	Nevada	1.7	26	34
43	South Dakota	0.9	40	34
44	Nebraska	1.0	39	34
45	Utah	1.3	20	32
46	New Mexico	1.0	28	29
47	Idaho	0.9	28	28
48	Montana	0.7	29	25
49	Alaska	1.0	20	24
50	Wyoming	0.6	24	22
51	North Dakota	0.5	27	21
	Mean of States	1.5	36	43
Source: Analysis 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				
** Ranking is based on unrounded data.				

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 8 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 8 shows that for each state the number of such improbable blocks are relatively high (the mean is 3,208) but the share is relatively low at 1.5 percent. Most states have at least one thousand such blocks. The total of such blocks from data used to compile Table 8 is over 160,000.

South Dakota, where 3.16 percent of all blocks have this condition, is the state with the highest rate. On the other hand, the District of Columbia (0.03) has the lowest percent of blocks where there are children but no adults. The state with the most census blocks that meet this condition is Texas with 10,588.

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 August 2021 DP demonstration file there were 4,651 such blocks. For Alabama, the increase was from 5 to 4,195, in Alaska the number of such blocks increased from 3 to 402, and for Washington State, it went from 11 to 2,041. 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.

Table 8. States Ranked by the Percent of Blocks with Children (ages 0 to 17) but No Adults (ages 18 or older)			
Rank*		Number of Blocks with population age 0 to 17 but no population ages 18 or older	Percent of Blocks with population age 0 to 17 but no population ages 18 or older
1	South Dakota	2,795	3.2
2	Nebraska	5,612	2.9
3	Iowa	6,175	2.9
4	Kansas	6,419	2.7
5	North Dakota	3,473	2.6
6	Minnesota	5,732	2.2
7	Missouri	7,363	2.1
8	West Virginia	2,899	2.1
9	Vermont	693	2.1
10	Oklahoma	5,539	2.1
11	Maine	1,371	2.0
12	Montana	2,483	1.9
13	Arkansas	3,482	1.9
14	Indiana	5,003	1.9
15	Wisconsin	4,654	1.8
16	Mississippi	2,994	1.7
17	New Hampshire	834	1.7
18	Tennessee	4,071	1.7
19	Alabama	4,195	1.7
20	Idaho	2,493	1.7
21	Kentucky	2,646	1.6
22	Ohio	5,832	1.6
23	Illinois	7,107	1.6
24	South Carolina	2,720	1.5
25	Pennsylvania	6,191	1.5
26	Virginia	4,129	1.4
27	Colorado	2,830	1.4
28	Michigan	4,581	1.4
29	Wyoming	1,187	1.4
30	Georgia	3,897	1.3
31	New York	4,651	1.3
32	North Carolina	3,804	1.3
33	New Mexico	2,117	1.3
34	Texas	10,558	1.2
35	Louisiana	2,282	1.1
36	Arizona	2,675	1.1
37	Washington	2,041	1.0
38	Maryland	1,498	1.0
39	Utah	1,192	1.0
40	Oregon	1,967	1.0
41	Delaware	224	0.9
42	Alaska	402	0.9
43	Florida	3,852	0.8
44	Connecticut	479	0.7
45	Massachusetts	1,121	0.7
46	Nevada	544	0.6
47	Rhode Island	144	0.6
48	California	3,848	0.5
49	New Jersey	747	0.4
50	Hawaii	78	0.3
51	District Of Columbia	2	0.0
	State Means	3,208	1.5
Source: Analysis analysis of data provided by David Van Riper of IPUMS at the University of Minnesota			
* Ranking is based on unrounded data.			

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 DP 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. Keeping the link between parents and children within a household while employing DP is also a concern for data from the American Community Survey. The Census Bureau is discussing the use of Differential Privacy or Formal Privacy in the American Community Survey but has said it will not implement any change before 2025.

Other kinds of impossible or improbable results are shown below.[4]

- 347,975 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.
- 152,496 blocks with a population before DP was applied but no population after DP was applied.
- 504,325 blocks with population in households but no occupied housing units after DP was applied.
- 148,253 blocks with occupied housing units but no population after DP was applied
- 546,072 blocks with more than 15 persons per household after DP was applied.

[4] This information was provided by David Van Riper at IPUMS

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 8, but they undermine the veracity of the census data broadly. The high number of improbable results reflected in Table 8 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.

Conclusion

The previous sections of this report provide information on accuracy of DP-infused data and provide a profile of the likely errors for children that will be seen in data for in the 2020 Census. For large geographic areas such as states, DP is unlikely to have much impact on data for children. However, for small areas and smaller populations there are still some concerns about the level of inaccuracy resulting from the use of DP.
The data shown here underscores the point that DP-infused data are most problematic for smaller (less populated) units of geography or smaller populations. This point is illustrated here based on Unified School Districts, Places, and census blocks, and for minority populations. This is important because there are a large number of small geographic units for which census data are produced. It is also important because the next file that will be released by the Census Bureau with 2020 Census data is the Demographic and Housing Characteristics file where many smaller groups (such as the population age 0 to 4 by race) will be released.

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.” or “unfit for use.” Each person will probably have a different answer to how much error in census data for children is too much error.

Appendix A 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. 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.

DP is meant to balance the quality/accuracy of census data and the need to protect respondent confidentiality or privacy. The application of differential privacy allows the Census Bureau to control the amount of error injected into the data.

A measurement called epsilon is a key to how DP works. The Census Bureau (2021, page 32) defines epsilon as:

> "A measure of privacy loss. Higher values of epsilon results in more privacy loss, whereas lower values result in less privacy loss. Epsilon may also be referred to as the privacy-loss budget, although in the TopDown Algorithm, the privacy-loss budget is allocated using the parameter rho defined in Zero-Concentrated Difference Privacy"

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.

Over the past few years, the Census Bureau has released several Demonstration Products with different levels of epsilon to help users understand the implication of DP. In October 2019, the U.S. Census Bureau (2019) released what the first of what they called 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. The Census Bureau has released five demonstration products (October 2019, May 2020, September 2020, November 2020, and April 2021) prior to the August 2021 demonstration products

Somewhat belatedly, the Census Bureau announced that they purposefully used a low level of epsilon for earlier DP files, which lead to a high level of privacy protection and a poor level of accuracy. Many data users and analysts thought the level of epsilon in the early demonstration products was what the Census Bureau was planning to use for the 2020 Census data and the high level of inaccuracies elevated concern. Clearer communication from the Census Bureau on the point of the earlier demonstration products would have reduced the level of concern among users. .

References

Bouk, D. and Boyd, D. (2021). “Democracy’s Data Infrastructure; The technologies of the U.S. Census.” Knight First Amendment Institute at Columbia University, 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/ .

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 Kids website https://countallkids.org/the-importance-of-small-area-census-data-for-school-demographics/

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/ ,

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) .

O’Hare, W.P. (2021). Analysis of Census Bureau’s April 2021 Differential Privacy Demonstration Product: Implications for Data on Children, Count All Kids Website,

https://countallkids.org/resources/analysis-of-census-bureaus-april-2021-differential-privacy-demonstration-product-implications-for-data-on-children/

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 (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,” U.S. Census Bureau, Washington, DC. FEBRUARY 02, 2021.

U.S. Census Bureau (2021b. “Developing the DAS: Demonstration Data and Progress Metric” U.S Census Bureau, Washington, DC. , 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 (2021). “Disclosure Avoidance for the 2020 Census: An Introduction,” November, U.S. Census Bureau, Washington, DC.

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 .