November 23,2020

Implications of Differential Privacy for Reported Data on Children in 2020 U.S. Census [1]

by

William P. O'Hare

President, O'Hare Data and Demographic Services LLC

Consultant to the Count All Kids 2020 Census Complete Count Committee

Executive Summary

The U.S. Census Bureau is planning to use 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 young children (ages 0 to 4) from the 2020 Census. The analysis also examines other age groups of children in the context of school districts.

The U.S. Census Bureau is still refining its effort to implement DP, but analysis of the most recent demonstration data available for young children shows that for several kinds of geographic units (counties, State legislative districts, school districts, places, and census tracts) the distortions injected by DP to help protect privacy, foster large errors for the population ages 0 to 4. For example, the Census Bureau's May 2020 demonstration file showed that the 2010 Census count of children ages 0 to 4 would exhibit errors of 10 percent or more in about two-thirds (64 percent) of all census tracts

[1] The Funders Census Initiative 2020 provided support for this report. The views expressed are those of the authors and should not be attributed to our advisors or funders.

after the application of DP. And more than a quarter of the tracts (28 percent) had errors of 25 percent or more for children age 0 to 4.[2]

Data for school districts were also examined. For smaller populations (i.e., age 4 or ages 0 to 4) there were substantial errors for school districts. For example, DP methods introduced errors of 10 percent or more for counts of children age 4 in 68 percent of school districts. DP introduced errors of 10 percent or more for counts of children ages 0 to 4 in 44 percent of school districts. For the population ages 5 to 17 and for ages 0 to 17 the error rates are lower.

Smaller geographic areas in terms of population size tend to have higher levels of error injected by DP. This is important because the census is designed to produce data for a lot of small geographic units These errors are likely to cause problems in many use cases such as the amount of state and federal funds received by school districts. For a small school district to get 10 percent less money than it deserves will cause serious problems. It will be difficult for child advocates to support the use of DP in the 2020 Census if it produces errors like those identified in this paper.

The final decision about the use of DP in the 2020 Census is likely to be made in December 2020 or January 2021, and the U.S. Census Bureau is still looking for feedback from data users. Comments can be sent to 2020DAS@census.gov.

[2] In this analysis, errors are the difference between DP-infused 2010 Census data and the 2010 Census data without DP.

Introduction

The U.S. Census Bureau is planning to use a new method called differential privacy (DP) to help protect confidentiality and privacy in the 2020 Census.[3] This paper provides some information on how DP is likely to impact the quality of data for young children from the 2020 Census. 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 DP here to distinguish the version of DAS that includes DP from other versions of DAS.

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. "

Basically, DP injects error into the census tabulations that are based on the true responses to the census by adding or subtracting random numbers from table cells that reflect the true responses. 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

[3] Note that differential privacy is sometimes called "formal privacy,"

with regular updates of their work (U.S. Census Bureau 2020c). For an independent look at differential privacy see Boyd (2020).

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."

This paper focuses on metrics for assessing the accuracy side of that tradeoff with respect to young children by reporting empirical evidence about the likely level of errors injected into the Census data for children.

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

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

However, it should be noted that there was another Demonstration Product released by the U.S. Census Bureau on September 17, 2020.[4],[5] Since this product only provides data for the population ages 0 to 17 without any data provided for smaller age groups of children one cannot use that file to examine data for the population ages 0 to 4.

It is important to acknowledge that the data produced in the September 17th release appears to reduce the error for children ages 0 to 17. Table 1 shows several accuracy measures for census tracts for ages 0 to 17 from the May 27, 2020, and the September 17, 2020, releases. For example, the Mean Absolute Percent Error of the count of the population ages 0 to 17 decreased from 8.6 percent to 4.0 percent for tracts (mean population for census tracts is about 4,000 total population).

[4] Technically the U.S. Census Bureau released a Privacy Protected Microdata File (PPMF) which NHGIS-IPUMS converted into tables that could be compared to the unperturbed 2010 Census results.
[5] It should be noted that there was an error in the file released by the U.S. Census Bureau on September 17, 2020 that was identified after its release. However, it is not clear that the error identified would have any impact on the data shown here (U.S. U.S. Census Bureau 2020f).

Table 1. Data Comparing the Accuracy for the Population Ages 0 to 17 from 5-27-2020 Census Bureau Release to 9-17-2020 Census Bureau Release for Census Tracts		
	Data from May 27, 2020, Census Bureau Release	Data from September 17, 2020, Census Bureau Release
Number of Units in the Analysis	**72,328**	**72,301**
Mean Size of District (Total Population)	4,267	4,267
Mean Absolute Numeric Error*	28	18
Mean Absolute Percent Error	8.6	4.0
Number of Units with errors of **5%** or more	16,659	8,616
Percent of Units with errors of **5%** or more	23.0	11.9
Number of Units with errors of **10%** or more	5,146	2,356
Percent of Units with errors of **10%** or mor**e**	7.1	3.3
Number of Units with errors of **25%** or more	1,332	709
Percent of Units with errors of **25%** or more	1.8	1.0
Source: Authors analysis of data released by the Census Bureau on May 27, 2020.		
Does not include Puerto Rico or geographic units with zero population age 0 to 4 in either 2010 Summary File or DP-infused data.		
* The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.		

The evidence in Table 1 suggests that if the more recent version of DP were applied to data for young children, the results might be more encouraging. But until that evidence is provided, the data from May 27th is the best we have available to judge the likely impact of DP on the data for young children.

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 data based on the application of DP to 2010 Census data. A brief explanation of each measure of accuracy or bias provided by the U.S. Census Bureau is shown in Appendix A.

While there are a large number of accuracy measures that could be calculated from the May 27, 2020 U.S. Census Bureau release, I only look at a few key measures here. I include the Mean Absolute Error in the tables shown here (I label this Mean Absolute Numerical Error in the tables to distinguish it from the Mean Absolute Percent Error) but I also include the Mean Absolute Percent Error which I believe is a more important measure. An absolute error reflects the magnitude of the error regardless of direction. This approach is 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. A geographic unit with an absolute error of 10 percent or more could be 10 percent too high or 10 percent too low. I focus on percent error because it reflects the size of the error relative to the size of the population. An error of a given magnitude (say 1000 people) may be trivial in large places but very significant in smaller places.

While the Mean Absolute Percent Error is informative, I focus this analysis on the number and percent of geographic units that have relatively large errors. I believe the

number and percent of large errors or outliers are the most important measures of accuracy. I use three benchmarks to identify large errors. The 5 percent benchmark shown in Table 1 is one used by the U.S. Census Bureau in their most recent set of metrics and the 10 percent benchmark in one used in the metrics proposed by the U.S. Census Bureau in March 2020. I added the benchmark of 25 percent to provide some indications of extremely large errors.

I believe it is these extreme errors that will be the biggest problem caused by DP. The fact that the biggest errors (percentage wise) happen in smaller places is likely to generate concerns in many places across the county.

Evaluation of Data for Age 0 to 4

I focus first on the population ages 0 to 4 because this age group is special in terms of having experienced the highest net undercount of any age group in the 2010 Census. Also, in their March 2020 release, the U.S. Census Bureau (2020a) provided data related to several “Use Cases” and the population ages 0 to 4 was one of those. However, it is important to recognize that there is no evidence to indicate that DP will affect the population age 0 to 4 differently than any other five-year age group.

Table 2 provides several accuracy measures for the population ages 0 to 4 for several different types of geographic units used in the census. The results shown in Table 2 indicate that DP is unlikely to have much of an impact on the data for states (and the District of Columbia). However, the situation is quite different for most types of

substate geographic units. For all the substate geographic units shown in Table 2, DP is likely to cause severe distortions.

Table 2. Summary Table Showing the Likely Impact of Differential Privacy on Census Reported Data for Children Ages 0 to 4

Level of Analysis	States + DC	Counties (or County equivalent)	State Legislative Districts (Upper Chamber)	State Legislative Districts (Lower Chamber)	Places	Census Tracts
Number of Units in the Analysis	**51**	**3,141**	**1,942**	**4,720**	**27,560**	**72,056**
Mean Size of District (Total Population)	6,053,834	98,295	158,983	64,898	8,282	4,274
Mean Absolute Numeric Error *	217	83.7	216.3	144	52	44
Mean Absolute Percent Error	0.1	7.4	3.1	5.3	43.6	21.3
Number of Units with errors of 5% or more	0	1,206	374	1,788	23,609	58,716
Percent of Units with errors of 5% or more	0	37	19	38	86	81
Number of Units with errors of 10% or more	0	640	83	655	20,316	46,339
Percent of Units with errors of 10% or more	0	20	4	14	74	64
Number of Units with errors of 25% or more	0	180	2	64	13,215	19,996
Percent of Units with errors of 25% or more	0	6	0	1	48	28

Source: Authors analysis of data released by the Census Bureau on May 27, 2020.

Does not include Puerto Rico or geographic units with zero population age 0 to 4 in either summary File or DP-infused data.

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

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 small areas is especially important. In addition, Reamer (2020) shows that about two-thirds of federal funding formulas which use census-derived data use substate data.

Large errors are more prevalent in smaller geographic units. The percent of substate units with errors of 10 percent or more ranges from a high of 74 percent for places[6] (mean population of about 8,000) to a low of 4 percent of the Upper Chamber of State Legislative Districts (mean population of 159,000). In Table 2, the Mean Absolute

[6] Places are defined by the U.S. Census Bureau as Incorporated Places and Census Designated Places (CDPs).

Percent Error is negatively associated with the mean population size of the district. Districts with larger mean population size have smaller Mean Absolute Percent Error.

In general, types of geographic units that tend to be smaller in population size, such as Places and Census Tracts have more large errors than types of geographic units that are large like counties. However, even among counties, smaller counties have larger percentage errors than larger counties (O'Hare 2019).

Application to School District data

It is important to recognize that the young children (ages 0 to 4) examined in the previous section will grow older in the decade after the 2020 Census and they will become a large part of the school-age population before the 2030 Census. Schools are probably the most widespread public institution closely related to children. They exist in every corner of the country. 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. At the CNSTAT workshop a couple presentations reflected implications for school districts (Vink 2019; Nagle and Kuhn 2019). So this section of the paper examines data related to school districts. I examine data for age 4 alone, ages 0 to 4, and the school-age population (ages 5 to 17). In the use case data scenarios provided by the U.S. Census Bureau (2020a) in March 2020, age 4 and ages 0 to 4 are both included. Ages 5 to 17 reflect the school-age population which has a lot of funding implications for school districts. To provide a complete picture I also examine the data for all children (ages 0 to 17).

Ages 0 to 4

I look at ages 0 to 4 first, because this follows logically from the data in Table 2. Table 3 shows that for ages 0 to 4 many school districts have large errors. More than four out of ten (44 percent) have errors of 10 percent or more and nearly one-fifth (17 percent) have errors of 25 percent or more.

Table 3. Summary Table Showing Impact of Differential Privacy for Ages Most Relevant for Unified School Districts

	Age 4 Only	Ages 0 to 4	Ages 5 to 17	Ages 0 to 17
Number of Units in the Analysis	10,529	10,840	10,875	10,880
Mean Size of District (Total Population)	29,299	28,482	28,390	28,377
Mean Absolute Numeric Error*	33	70	84	56
Mean Absolute Percent Error	32	16	7	5
Number of Units with errors of **5%** or more	8,659	7,083	4,231	2,590
Percent of Units with errors of **5%** or more	82	65	39	24
Number of Units with errors of **10%** or more	7,188	4,808	2,015	1,185
Percent of Units with errors of **10%** or more	68	44	19	11
Number of Units with errors of **25%** or more	4,315	1,872	469	307
Percent of Units with errors of **25%** or more	41	17	4	3

Source: Authors analysis of data released by the Census Bureau on May 27, 2020.

Does not include Puerto Rico or geographic units with zero population age 0 to 4 in either 2010 Summary File or DP-infused data.

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

Age 4 alone

Table 3 shows accuracy measures for age 4 for school districts. Age 4 is important because it is used to project the number of new kindergarten students a school is likely to receive in the following year. The measures in Table 3 indicate DP is likely to inject severe distortions into the Census reported data for children age 4. More

than two-thirds of school districts (68 percent) have errors of 10 percent or more and 41 percent of school districts have errors of 25 percent or more.

School-Age Population (Ages 5 to 17)

Table 3 provides accuracy measures for the school-age population (ages 5 to 17) for school districts. Almost one-fifth (19 percent) of the School Districts exhibited errors of 10 percent or more, and 4 percent experienced errors of 25 percent or more for the population ages 5 to 17. It is important to understand the school-age population often drives funding for a school system. A school system that receives 10 percent less money than it deserves will face difficult choices and is likely to end up with larger class sizes and possible lower teacher pay.

In summary, Census school district data for age 4, ages 0 to 4, and ages 5 to 17 based on DP-infused statistics show an extremely high level of errors. Such data would be unusable in many cases. This indicates the U.S. Census Bureau still has improvements to make before the data for young children is acceptably accurate.

Unit size and Accuracy

The data shown here also underscore 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 counties and school districts. The majority of School Districts are relatively small. Out of 10,880 school districts, more than half have less than 10,000 total population and the Mean Absolute Percent Error for the

population ages 0 to 4 for School Districts with less than 10,000 people is 24 percent. Of the 3,141 counties examined here, about one-fifth have populations less than 10,000 total population and the Mean Absolute Percent Error for such counties for ages 0 to 4 is 20 percent. Census tracts and Places are also relatively small geographic units. The average census tract has 4,274 total population and the average Place has a total population of 8,282. The U.S. Census Bureau produces data for many small geographic areas and the errors DP injects into the data for young children in these small areas is quite large.

Conclusion

The most recent data available from the U.S. Census Bureau regarding the likely impact of DP on 2020 Census data for young children suggests that the level of error introduced will result in a high level of errors for some geographic units.

It is worth noting that 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.

There are a couple of reasons for sharing this information with child advocates at this point in time. First, 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. One can provide feedback on this issue at this email address 2020DAS@census.gov. Second, 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 which may result in some odd statistics.

Appendix A The following metrics for accuracy are proposed by the U.S. Census Bureau:

1. Mean/Median Absolute Error (MAE): This is a measure of the “average” absolute value of the count difference for a statistic.

2. Mean/Median Numeric Error (ME): This is a measure of the magnitude and direction of the average difference for a statistic.

3. Root Mean Squared Error (RMSE): This is a measure of the square root of the average squared error for a statistic. It is the traditional measure of error for U.S. Census Bureau sample survey statistics.

4. Mean/Median Absolute Percent Error (MAPE): This is a measure of the “average” relative difference for a statistic.

5. Coefficient of Variation (CV): This is the relative error counterpart to RMSE. It is another traditional measure of error in U.S. Census Bureau sample survey statistics

6. Total Absolute Error of Shares (TAES): This measure finds the proportion of each MDF value to the total MDF value for the summary geography and subtracts the proportion of the CEF value to the total CEF value for the summary geography. The absolute value of these proportional differences across evaluation geographies is then summed to the summary geography level. The goal is to provide a measure of the distributional error in the MDF shares.

References

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

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

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 Burearu (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. 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