Small Counts, Big Mistakes:

An Analysis of the Tradeoff Between Statistical Accuracy and Individual Privacy at Low Level Geographies in the 2020 Census

Claire Brockway

April 24th, 2020

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts in Public Policy *at* The University of Chicago

Preceptor: Rich Gauthier

Acknowledgements

I want to provide recognition to all those in my life who made this project possible. Without a network of selfless family, friends, and faculty by my side, this research process would not have been nearly as rewarding.

The research and methodology of this paper could not have been possible without the help of Quentin Brummett. Thank you for sparking my interest in differential privacy and always making time in your busy schedule for my countless questions. You will always exist as a counterfactual to my belief that no one else in the world finds algorithms used by the census in the least bit interesting.

Finally, I would like to dedicate this paper to all four of my parents for their unwavering support and guidance throughout the thesis process and the course of my undergraduate education. For this and more, I am forever grateful. Thank you for setting the bar so high.

Abstract

The census is a critical asset to the United States' system of government and private enterprise. The data collected and distributed by the Census Bureau informs the planning and stewardship of both public and private resources at all levels of geography. The census serves as the foundation for equitable political representation in the U.S. House of Representatives, state legislatures, and municipal governments. Without the census, our nation's promise of equal representation would stand hollow.

As Americans become increasingly concerned about the privacy of their individual data, the Census Bureau faces a tradeoff between protecting the identities of its survey respondents and disclosing the statistics it collects in its data projects. To address this tradeoff, the Census Bureau announced in 2018 that it will implement a new statistical algorithm on its 2020 decennial census called differential privacy to protect the identities of its respondents.

The impact of differentially private algorithms on a dataset with as many critical applications as the census remains relatively unknown. Some census stakeholders worry that the additional statistical noise this algorithm produces will undermine the accuracy of the 2020 census by distorting low level geographical census population estimates. They fear these distortions may lead to larger down-the-line errors in government allocations and legislative districting.

This study examines the impact of the new differentially private algorithm on federal fiscal allocations to United States' public school districts using 2010 census data. More specifically, this study quantifies the fiscal error associated with this new privacy technique in the federal dissemination of Title 1 Grants to low-income public school districts and Rural Education Achievement Program (REAP) Grants to rural public school districts. Through examining the errors in eligibility rates and federal authorization amounts between a standard 2010 census dataset and differentially private 2010 census dataset, this study sets out to understand whether this new privacy policy should be a concern of census data users given the upcoming 2020 census. Ultimately, this study finds that the use of 2010 differential privacy census data introduced eligibility and allocation inaccuracies for federal grants to public school districts that amounts to millions of dollars and thousands of left behind, in-need children.

Table of Contents

Introduction:

The Decennial United States census comprises the single most important data collection project in the nation. Every ten years, the U.S. census sets out to report an accurate count of every household and individual residing in the country at that given moment. In turn, its data is used by a variety of stakeholders, from social scientists to politicians and government officials. Census data serves not only as the foundation for a significant amount of research in the social sciences, but also as a key input in formulas that allocate government resources and representation to communities across the country. In order for the census to result in accurate and informative data, it requires universal participation from each of the nation's inhabitants.

Title 13 of the United States Code requires the Census Bureau to prioritize privacy in its collection of personal information from individuals and businesses. This Title guarantees respondents that their personally identifiable information will not be published or used against them by a government agency or court (Gauthier 2019). The harsh consequences that come with violating Title 13 highlight the government's emphasis on protecting the responses of individuals. If the title is violated, individuals can serve a federal prison sentence of up to five years along with a fine of up to $250,000 (Gauthier 2019).

Over the last century, privacy has increasingly become a priority for the U.S. Census Bureau and the internal policies through which the bureau guarantees respondents' confidentiality have evolved over time in response to technological advancements. Not only have computing power and algorithmic sophistication rapidly increased in the last twenty years, but so has the volume of auxiliary information available to non-governmental sources that could try to reidentify a census database. As the Fundamental Law of Information Recovery states, “overly accurate data that answers too many questions will destroy privacy in a spectacular way” (Dwork and Rothblum).

After the Census Bureau staged its own successful internal database reidentification attack in 2018, in which it was able to accurately reconstruct the personal information of 52 million Americans from the tabular data of the 2010 census, the bureau recognized the need to protect its institutional integrity and legitimacy (Marvin, Jeffrey, et al. 2019)[1]. In order to encourage census participation and foster privacy, the bureau announced its intention to use a new privacy technique called *differential privacy* to protect its 2020 data. Differential privacy protects an individual's identity through the insertion of additional statistical noise to a dataset. This paper will define statistical noise as unexplained variability within a dataset that typically results from the idiosyncrasies of a unique sample population. While differential privacy is a practice increasingly gaining popularity among large tech companies such as Google and Apple, social scientists and statisticians are concerned with its application to the 2020 census. Many census stakeholders worry that this additional noise will undermine the accuracy of the 2020 census in that it will dramatically distort population estimates in census geographical areas which may lead to larger down-the-line errors in government allocations and legislative districting (Ruggles 2018). These errors would result in the inequitable and undemocratic distribution of our government’s resources and representation. At this moment,

[1] This paper will define a database reidentification attack as an attempt to construct a record-by-record copy of a confidential database exclusively through the public aggregate statistics it publishes

it is unclear what the true effect this new privacy policy will be in its application to a dataset as large as the census.

While the Census Bureau announced that the total population of every state will be published “as enumerated” in 2020, smaller levels of geography, from congressional districts down to census blocks, will still have their statistics published under this differentially private treatment. Therefore, in these smaller geographic units, additional noise and statistical inaccuracies can be expected. As the National Congress of American Indians recently noted, “The implementation of differential privacy could introduce substantial amounts of noise into statistics for small populations living in remote areas, potentially diminishing the quality of statistics about tribal nations” (Underwood and Zamarripa 2019). Tribal nations, school districts, and counties are among the many small level geographies whose data is likely to be implicated by this new policy.

This study aims to understand the extent to which the census’s application of differential privacy could impact the statistical accuracy of population counts in low level geographies for the bureau’s upcoming 2020 data. Through a retrospective comparison of standard 2010 census dataset to a 2010 census dataset with differential privacy techniques applied to it, this study will attempt to isolate the impact of this privacy policy on the eligibility criteria and allocation amounts of federal programs to a specific low level U.S. geography: school districts. Ultimately, this study will strive to answer the question: to what extent does the U.S. Census Bureau’s policy of differential privacy undermine statistical accuracy in its determination of eligibility and allocation amounts for federal grant programs to public school districts?

Background:

In order to fully understand the impact of differential privacy on census data accuracy, it is critical that one understands the role of the census as both a facilitator in our interactions with the American federalist system and a holder of our critical private information. The following background section of this study will first outline the process and purpose of facilitating an accurate census in 2020. After, the background section will offer a brief overview of the history of privacy policies implemented by the Census Bureau to protect their respondents' identities to contextualize the bureau's decision to adopt differential privacy for the 2020 census cycle. Finally, the background section will offer a basic explanation for how differentially private algorithms function as disclosure avoidance mechanisms on the upcoming 2020 census dataset.

An Introduction to the 2020 Census:

Listed among the civic duties of each American inhabitant[2] is the requirement to participate in the decennial census. The census is mandated and enshrined in Article 1, Section 2 of the U.S. Constitution, which states:

"Representatives and direct Taxes shall be apportioned among the several States

[2] The census collects data on every resident of the United States, not just every citizen. This includes non-citizens who live within the country, regardless of if they entered the country legally or illegally.

> which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons" (U.S. Constitution).

While each United States resident will, on average, complete the census only six to seven times over the course of their lifetime, it is one of their most important interactions with the federal government. An accurate count of the population not only acts as the basis for fair political representation, but it guides many functions of public life:

1. *Apportionment*

The apportionment of the 435 seats in The House of Representatives is divided among the 50 states according to each state's share of the total U.S. population. It is predicted that after the 2020 Census, states in the south and west will gain additional seats — and political clout — at the expense of northeastern and midwestern states (Population Reference Bureau 2019). On December 31st, 2020, the Census Bureau is anticipated to release apportionment population counts that will determine the size of state delegations to the 2022 U.S. House elections as well as state votes in the U.S. Electoral College in the 2024 presidential election.

2. *Redistricting*

The results of the decennial census not only guide the number of representatives each state receives, but also the geographical boundaries that each represents. Exactly one year from Census Day (April 1st, 2020), the Census Bureau publishes redistricting data that is used by state and local officials to inform the redrawing of congressional, state, and local district boundaries in accordance with the nation's democratic maxim that each individual's voting power is closely equivalent (i.e. one-person, one-vote rule) (Population Reference Bureau 2019).

3. *Allocation of Fiscal Resources to States and Localities*

One of the most important uses of census data is to determine the amount of federal allocations that state and local governments receive from the national government. It is estimated that during the 2015 Fiscal Year, Census Bureau data was used to distribute more than $675 billion in federal money to state and local communities for the purposes of education, housing, health, and infrastructure programs (Mellnik 2019). Universal participation in the census helps to ensure that accurate data is being used to equitably distribute funding for numerous government programs such as Head Start, Title I, highway construction and planning, and Medicaid (Population Reference Bureau 2019).

4. *Private and Public Sector Planning*

From an economic standpoint, data from the decennial census is critical to determine the demand for a wide range of government, business, and nonprofit services. A variety of public and private organizations rely on this data to measure the need for new products and services like roads, hospitals, schools, and other investments. The U.S. population is rapidly changing both demographically and geographically because of an increase in minority populations as well as a general migration of individuals and families out of midwestern and northern states to the west and the south among other reasons. The census helps to guide public and private business owners towards meeting the needs of a quickly changing market.

5. *Emergency Response*

In the wake of national emergencies and natural disasters, it is critical that first responders and disaster recovery personnel have access to accurate demographic counts to identify where and how much help is needed. Over the next decade, as climate-related disasters are predicted to become increasingly extreme, and globalization increases the probability of disease outbreak, an accurate census is necessary to assist public health personnel and epidemiologists in their responses to crises (Population Reference Bureau 2019). At the present moment, the United States federal government is relying on Census Bureau data to facilitate the efficient spread of medical resources during the COVID-19 pandemic. Population counts from the census are helping officials identify spatial clusters of at-risk populations across the nation such as the eldery and homeless (U.S. Census Bureau 2020).

6. *A Basis for Federal Survey Sampling Frames*

The decennial census is not the only mass survey conducted by the federal government, but it does inform the basis from the sampling frame that other important surveys utilize in conducting research relevant to the functioning of the U.S. government. The American Community Survey and the Current Population Survey, to name a couple, must rely on population estimates from the decennial census for 10 years for their population controls in building a representative probability sample (Population Reference Bureau 2019).

The six uses of census data described above exemplify why each American resident has a stake in a complete and accurate census. Before the census began on April 1st, 2020, local complete count commissions convened to strategize ways to encourage their communities to participate in the census to ensure their fair share of critical government resources and representation. The Census Bureau itself mobilized a mass marketing effort to draw attention to the 2020 census through social media and digital advertising. It is estimated that the Bureau spent $1.43 per U.S. resident on 2020 census marketing, up from $1.22 in 2010 (Mellnik 2019). State and local governments as well as nonprofits are specifically targeting their low response groups in these marketing campaigns and complete count commissions. Historically, geographic areas with high percentages of non-native English speakers, undocumented immigrants, young children, and racial and ethnic minorities are at a greater risk of being undercounted. Often, these individuals are difficult to count in the census because they are hard to contact, locate, persuade, and interview (Child Trends 2019). Locating and then contacting these individuals to ensure they receive and complete the census proves to be difficult since often they are highly mobile and potentially experiencing homelessness. Even if these individuals can be located by the Census Bureau, persuading them to comply with the census becomes arduous as many of these demographic groups tend to be suspicious of the government and demonstrate low levels of civic engagement. Finally, hard-to-count individuals are difficult to interview because language barriers, illiteracy, and a lack of access to the internet are disproportionately common among low income and minority regions of the country.

California is among many states with a large hard-to-count population, and is estimated to be the home of 10 million individuals who have been classified as potential census non-compliers. Despite its tight budget, the California state government has now committed $187 million to promoting the 2020 census. In 2010, it had spent a mere $2 million (Mellnik 2019). California's new governor, Gavin Newsom, initiated this 95-fold increase in census promotion to ensure that his ethnically and economically diverse set of constituents did not lose their right to equitable political representation and federal allocations. For each Californian who goes uncounted by the census, the state loses $2,000. In 2010, California's census participation was down three points from 2000, 73 to 76 respectively. According to a recent study by the Urban Institute, $187 million is a small amount to spend when states with large minority populations have around $300 million to lose as a result of an undercount (Child Trends 2019).

The 2020 census will pose new and notable challenges to both its enumerators and its participants. In its 2020 cycle, the census will move away from using paper forms as its primary medium of collecting data, choosing instead to encourage individuals to answer questions on the Internet. Census enumerators working in the field are now expected to ask participants to use mobile phone apps to complete the survey. While most households will be expected to answer the census online or by phone, the Census is still anticipating 30 percent of its surveys will be completed through the mail in paper form to accommodate households that do not have access to reliable broadband (Mellnik 2019). While it is not predicted that this shift in medium will impact census accuracy or response rates, some fear that the online website where individuals record their responses could be at risk of crashing while the census is in the field from March to June (Ruggles 2018). Since Healthcare.gov famously crashed in October 2013 when only a few thousand Americans tried to apply for health insurance, the

Census Bureau has to prepare its website for traffic that could be several orders of magnitude larger (Lapowsky 2019).

The 2020 census was in the media spotlight in the wake of President Donald Trump's administration's attempt to include a citizenship question of every respondent for the first time in the bureau's history. The Secretary of Commerce, and overseer of the Census Bureau, Wilbur Ross, argued that a citizenship question on the 2020 census would aid in the federal enforcement of the Voting Rights Act by providing accurate counts of citizen populations at different geographic levels. Opponents of the citizenship question argued that this proposal would only serve to weaponize the census by sparking fear in non-citizen respondents that their answers could be used to assist the federal government in finding their residences and removing them from the country. Ultimately, these opponents felt that the citizenship question would discourage minority and immigrant populations, who often lean left on the political spectrum, from participating in the census. Their concern was that the new U.S. House of Representative districts drawn from the resulting data would unjustly promote Republican interests. While Ross's proposal was eventually deemed unjustified by the Supreme Court in the case of *Department of Commerce vs. New York*, the betrayal and distrust many American residents felt towards the 2020 census did not disappear as quickly as the proposed question did. Many undocumented immigrants and members of Latinx communities remain suspicious of the census and fear that their participation in it will result in deportation or arrests from federal agencies such as Immigration and Customs Enforcement (ICE) (Child Trends 2019). Despite these justified concerns, section 8(c) of the

Census Act “prohibits federal, state, and local government agencies from using statistical datasets, including special tabulations produced by the Census Bureau to the ‘detriment’ of any individual who responded to a census or survey from which the dataset is built” (Gauthier 2019). As mentioned earlier, undocumented immigrants and ethnic minorities are critical low response demographic groups and their concern for the privacy and protection of their individual data should be recognized by the Census Bureau as a top priority in the 2020 census cycle.

This leads to the last notable new feature of the 2020 Census, and the topic of this analysis: differential privacy. In this new age of big data, people’s attitudes towards their personal data are shifting and the perceived risk of a data breach at the Census Bureau could be enough to lead to catastrophically high rates of nonresponse across the country. In December of 2018, the Census Bureau announced its plan to adopt a new emerging disclosure avoidance technology called differential privacy to protect its data against reidentification attacks. In the next section, this paper will offer a brief overview of the progression of the Census Bureau’s privacy protections to provide a context to readers as to how differential privacy emerged as a policy priority in 2020.

A History of Census Privacy Protections:

Today, in the wake of high-power computing and massive online repositories of personal information, the Census Bureau cannot ignore the issue of privacy in its data collection and dissemination methodologies, but not too long ago privacy procedures were far less of a policy priority to bureau officials. In fact, the first census in 1790 had no formal privacy protections. Instead of legally mandating that personal information disclosed in the census be guarded, census officials were required to post census lists in town squares for public review (Gauthier 2019). The astonishing increase in emphasis on individual privacy in the census over the past two centuries is mirrored by the change in public attitudes on disclosure of personal information. This was seen in a recent survey by Pew Research Center, which revealed that seven-in-ten U.S. adults felt that their personal information is less secure than it was five years ago (Auxier, Brooke, et al. 2019).

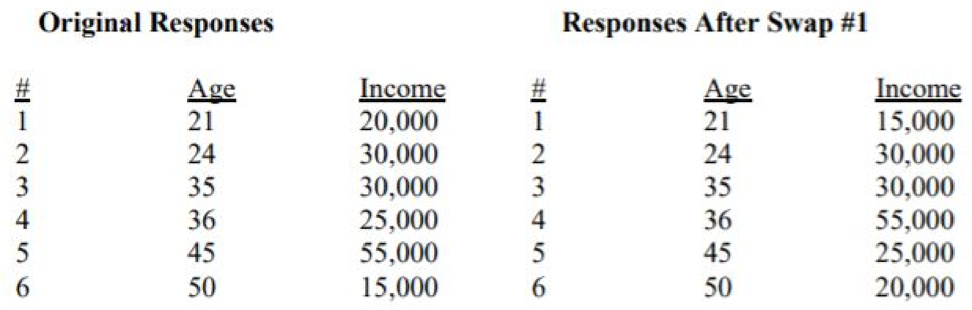
Most notable in the recent history of census privacy protections is the transition from preventing the *direct* disclosure of personal information to address the growing threat of *indirect* disclosure. While direct disclosures involve the intentional release of private information, indirect disclosures occur unintentionally when a set of statistics published in a dataset can be paired with other existing datasets to reveal the personal characteristics of respondents (US Census Bureau 2019). The twenty-first century is full of indirect disclosure threats and attacks, from the 2013 data breach at Adobe that revealed nearly three million individuals’ credit card information to the Equifax attack that exposed the social security numbers, driver license registrations, and addresses of close to half of the United States’ population. Fast and powerful computers coupled with advancements in the field of data science and the growth of personal data available online have forced the Census Bureau to transition privacy from a backburner issue to a frontline priority.

In the first censuses, bureau officials did not prioritize privacy. When the Census Bureau realized in 1840 that its response rates among businesses and manufacturing companies were dismal due to their reluctance to provide public information on their statistics, it was ordered that census marshalls provide assurances about the privacy of these responses. This was fully implemented in 1850, when for the first time, census responses were not posted publicly. Throughout the rest of the nineteenth century, privacy protections were augmented through stricter internal bureau policies. In 1880, new laws prohibited census takers from disclosing responses, and fined them if they did. Soon after, in 1890, census files were removed from state and local governments, unless paid for. By 1929, the sharing and sale of Census records ended, and a new law stated that an individual or his/her descendents could have access to that individual's responses (Gauthier 2019).

Both World Wars presented unique challenges to the Census Bureau's attempts to increase its privacy protections. Under President Taft's administration in 1916 and the Second War Powers Act in 1942, census data could be shared for the purposes of the military draft, which perpetrated a general distrust of the bureau among young men and women (The Leadership Conference Education Fund 2015). By 1954, this distrust was fully realized by the Census Bureau leading them to draft Title 13 of the U.S. Code. This code outlawed the publication and sharing of census results for non-statistical purposes. The code was upheld in the federal court decision of *United States v. Bethlehem Steel Corp.* (1958) and remains in effect today (Gauthier 2019).

By the end of the 1950s, the Census Bureau had automated computers to conduct their tabulations. As computers become an integral part of the bureau's processes, more sophisticated privacy measures were adopted (Gauthier 2019). The 1990 census introduced two important privacy protections that still exist within the bureau today: data swapping and "blank and impute" protections. Data swapping involves interchanging the values of sensitive variables among respondents in a similar geographical area. The swap of a variable does not provide any masking, but instead the probability that a swap masks a particular individual is inversely proportional to the frequency that the value of the variable appears in the file (Moore 2019). Figure 1 displays an example of a census microdata file containing the age and income of six respondents[3]. To protect the anonymity of the respondents, income values were randomly swapped.

Figure 1:



Original Responses			Responses After Swap #1		
#	Age	Income	#	Age	Income
1	21	20,000	1	21	15,000
2	24	30,000	2	24	30,000
3	35	30,000	3	35	30,000
4	36	25,000	4	36	55,000
5	45	55,000	5	45	25,000
6	50	15,000	6	50	20,000

Data swapping as a privacy technique has multiple advantages that help to explain its success through much of the digital age. When it is performed across multiple key variables, this technique can erode any relationship that earlier existed between the individual record

[3] When studying survey and census data, microdata can be defined as information at the level of the individual respondent. In the case of the decennial census, microdata would contain information on an individual's home address, educational level, age, employment status, etc.

and its corresponding respondent. Additionally, data swapping protects individuals with extreme or unique variable values that might make them vulnerable to identification in a microdata file (Moore 2019). For instance, if one Asian individual lived in a primarily white neighborhood, the other characteristics of the individual could be discovered quickly in a small geographic table had there not been swapping.

Another privacy mechanism applied by the Census Bureau in the 1990s is *blanking and imputing*. This method involves selecting a couple of records from a microdata file and replacing their real values with imputed values. In recent years, the Census Bureau has built on the blanking and imputing method in its use of synthetic data. Synthetic data can be created from a survey using posterior predictive models that generate data with many of the same properties as the original dataset. For each variable in the original dataset, a regression model is created. Then, for each record, the actual value of the variable is replaced by the model's new imputed value. A census researcher or data user can then draw random samples from these synthetic survey populations to answer database queries[4] (Gauthier 2019).

Over the past two years, compelling evidence has emerged indicating the Census Bureau's old privacy techniques are no longer completely effective. In 2018, a team of Census Bureau employees responded to internal privacy concerns through staging its own database reconstruction attack on its 2010 census data. The team ultimately was able to reconstruct the personal information of 100 million Americans. After accounting for technical errors, the team had correctly identified 52 million individuals, a little over a sixth

[4] A database query can be defined simply as a request for information or data from a database.

of the U.S. population (Hansen 2018). In response to this successful re-identification attack, the Census Bureau moved to adopt a more sophisticated version of privacy protection labeled differential privacy which will be outlined in further detail in the next section of this paper.

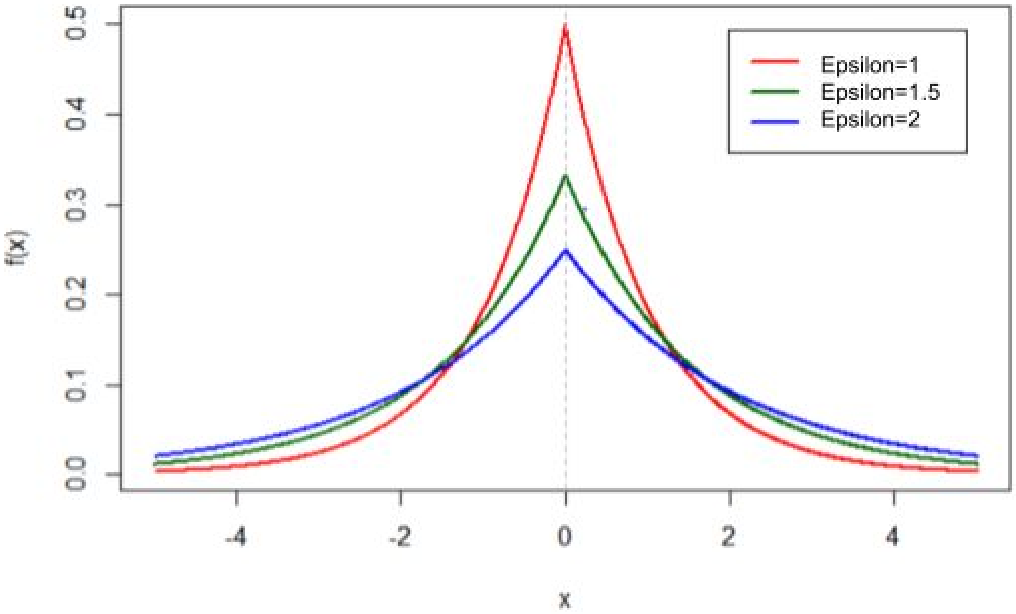
Defining Differential Privacy:

The main statistical mechanism of interest in this analysis is the differentially private algorithm. This algorithm protects individual variable responses through the insertion of statistical noise into a database. Statistical noise can be artificially generated through the Laplace Distribution, which will be explained in greater detail later in this section. In order for the algorithm to achieve its intended result of full individual privacy, the amount of noise inserted into the database should mathematically guarantee that the result of a differentially private analysis can make the same inference about a respondent's information, regardless of if that individual is included as an input in the database.

The differentially private technique was developed in 2006 by Cynthia Dwork, a computer science professor at Harvard University. Dwork asserted that privacy could actually be quantified through specific inputs into a differentially private algorithm (Abowd and Schmutte 2018). To explain what this means in context, consider the following scenario. A database containing sensitive information wants to protect its contents from being reverse-engineered by an adversary to reveal or learn its data inputs. In this context, an adversary can be defined as someone with the intent to learn or reveal the sensitive information in this database. In order to prevent this adversary from learning such sensitive information, the Census Bureau has two main options: 1.) publish less detailed statistics or 2.) publish noisier statistics. Differential privacy is the mechanism to achieve the second solution. Its algorithms incorporate statistical noise into the dataset so that the data an adversary sees is much too noisy and imprecise to result in a successful privacy breach.

There is a fine line between protecting privacy through statistical noise and ruining the integrity of the statistics the Census Bureau publishes after inserting the noise. Differential privacy allows statisticians to choose their ideal balance between database utility and individual privacy through the parameter *epsilon*. Epsilon can be thought of as the quantification of privacy loss in a database. Therefore, as epsilon gets larger, there is less privacy protection and less statistical noise. The chosen epsilon value is imputed into a Laplace probability distribution. The Laplace distribution is centered on an individual statistic's true value and its variance widens as epsilon decreases. So the probability a statistic deviates greatly from its mean increases as epsilon decreases (Georgian Partners 2018). A plot of the Laplace Noise distribution at different levels of epsilon is shown in Figure 2.

Figure 2:



The purpose of this paper will be to understand the tradeoff between privacy and statistical accuracy in a set of differentially private data products from the 2010 decennial census released by the Census Bureau in September of 2019. It is important to note that these data products have a fixed epsilon value of six (Manson et al. 2019). The Census Bureau has yet to release what epsilon value it will use for the 2020 decennial census data products, but after investigating the available 2010 products, this analysis will aim to quantify the statistical inaccuracies in terms of the error they produce in federal fiscal misallocations and, in turn, recommend whether the epsilon parameter be increased, decreased, or remain the same.

Literature Review:

Since differentially private algorithms are relatively new to both private and public sector data, the existing literature on their impact on statistical accuracy is not only sparse, but also quite speculative. It is important to note that the 2020 Census will be the first time that differential privacy will be applied comprehensively to a dataset that implicates every American inhabitant. Therefore, a variety of statistics and privacy experts have dedicated time and resources into examining the impact of differential privacy on the accuracy of Census data in a variety of common application scenarios. In the following literature review, this study will first synthesize multiple conflicting predictions on the extent to which differential privacy will undermine the accuracy of 2020 census data abstractly. Additionally, the literature review will explore several studies that attempt to quantify the impact of Census data inaccuracy on government allocations and congressional redistricting.

Differential Privacy and its Predicted Impacts on Statistical Accuracy:

Social scientists, computer scientists, and statisticians have yet to reach a consensus on the true impact of differential privacy on census data accuracy. Critics of differential privacy argue that the Census Bureau's previous disclosure avoidance technique of swapping was sufficient despite the recent, successful database reidentification attack. In his working paper on the *Differential Privacy and Census Data: Implications for Social and Economic Research*, Steven Ruggles, a data scientist with the Minnesota Population Institute, highlights that the *only* successful database reconstruction attack on census data was carried out by an inhouse census team who can be assumed to have high levels of familiarity with the data (Ruggles 2019).

Additionally, Ruggles and his team questioned the Census's claim after its own attack that, "the micro-data from the confidential 2010 Hundred-percent Detail File (HCF)[5] can be accurately reconstructed" using public use summary tabulations. In Ruggle's opinion, it should not have come as a great surprise to the Census Bureau that the individual-level characteristics of respondents could be predicted using tabular data. However, the only circumstances where this database reconstruction tactic is possible is when a particular census block or geographical area is heterogeneous, since the unique characteristics of individuals in tables are easy to parse out into individual-level microdata. But 47% of census blocks contain a single race and 60% contain a single ethnicity. Reconstructing these homogeneous blocks would be near impossible. In fact, Ruggles and his team highlight that only 50% of the census's reconstructed data accurately matched the source data. Even if the reconstructed data did match this source data, in order for it to be connected to the actual identities of respondents it would have to be matched with an external identified database (such as one with names or Social Security numbers). While the Census Bureau did attempt to externally validate its findings, Ruggles points out that they could only re-identify a fraction of the microdata. Since this is the case, Ruggles and other social science researchers remain skeptical of the need for differential privacy in the 2020 census (Ruggles 2019).

[5] The Hundred-percent Detail File (HCF) contains information collected by each decennial census on the age, sex, race, and Hispanic make up of the population on each inhabitated census block.

On the other side of this debate are privacy proponents who advocate that differential privacy is the best precaution the Census Bureau can take as other publicly available information such as credit reports, property records, and voter registration rolls augment the risk of a database reconstruction attack. John Abowd, the chief scientist and associate director for research at the Census Bureau, argues that, "the database reconstruction theorem is the death knell for traditional [data] publication systems from confidential sources… It exposes a vulnerability that we were not designing our systems to address". Abowd has led the movement within the Census Bureau to adopt differential privacy as a policy. He and other census employees argue that differential privacy has four key advantages that were critical in the bureau's decision to adopt it (Abowd and Schmutte 2018):

1. *Transparency* — The Census Bureau argues that differential privacy is a transparent method of disclosure avoidance. While this claim is widely contested in the social science community, the bureau counters with its assertion that data users in 2000 and 2010 did not really know the true error that was introduced through old privacy methods like swapping. Meanwhile, differentially private data from 2010 is directly comparable to standard 2010 census data on the census's website, so curious researchers can explore the error introduced by the method for themselves.
2. *Tunable privacy guarantees* — As previously acknowledged, the amount of privacy loss in a differentially private algorithm is quantified by the parameter epsilon. If the statistics released by the Census Bureau are wildly inaccurate in future data projects,

the epsilon parameter can be easily refined to a more socially efficient level, or a level that maximizes the utility of both privacy proponents and census data users.

3. *Protects against accurate database reconstruction* — Differential privacy inhibits an adversary from uncovering what individuals are included in a summary table released by the census. Therefore database reconstruction attacks cannot successfully reidentify any respondents if their presence in the data is unknown.
4. *Protections for every member of the population* — The Census Bureau insists that this method of disclosure avoidance is democratic and equitable in that every individual in the database will be protected from reidentification. Under earlier privacy techniques, individuals with unique or rare characteristics were often at a greater risk of reidentification than those with common characteristics.

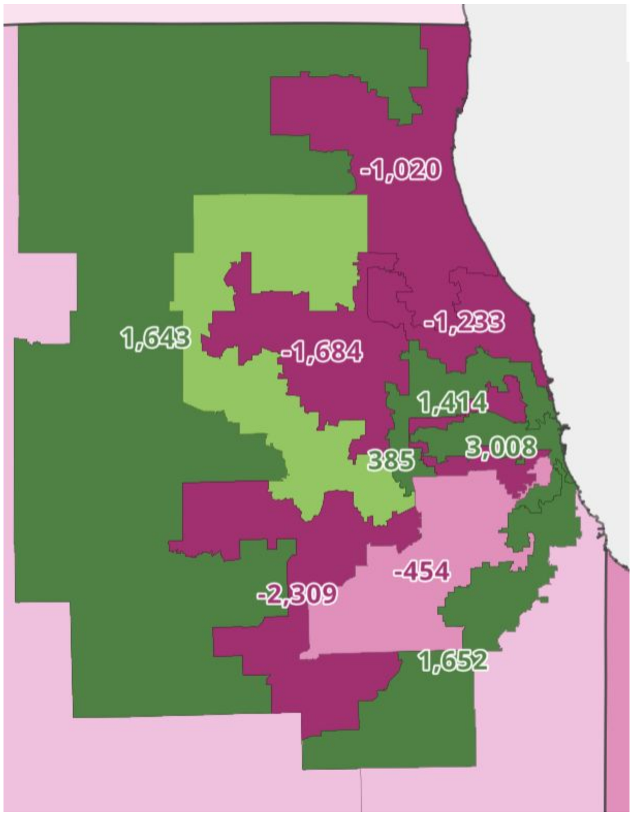
The Census Bureau's list of advantages for differential privacy are compelling to individuals concerned about database disclosure, but they do not answer much about the question of the tradeoff between statistical accuracy and privacy in 2020. While Ruggles and his colleagues indicate concerns about this new policy, they offer no statistical analysis in their paper to offer evidence to their claim that census-based social science and economic research will be negatively impacted by differential privacy. In the next section of this literature review, this paper will review two studies that test differentially private data's impact on statistical accuracy on legislative redistricting and Title 1 education funds.

Predicted Implications for Redistricting and Government Allocations:

The statistics the decennial census outputs inform the redistribution of federal resources and power in the form of federal fiscal allocations and congressional seats. It is fair to assert that census statistics either directly or indirectly facilitate almost every American's interactions with the federal government. Differential privacy opponents fear that this algorithm will degrade statistical accuracy to a large enough extent that the government will unjustly allocate taxpayer dollars and congressional seats (Ruggles 2019).

Recently, the Caliper Corporation conducted a GIS study on how the boundaries of the current 116th Congressional Districts populations would have been different if they had been formed under differentially private 2010 census data. They were ultimately able to conclude that most district populations were at risk of a statistical inaccuracy of one thousand to three thousand people. As an example, Illinois' seventh district, under differentially private data, has 3008 more constituents than were actually reported in 2010 (Caliper 2019). The results of this analysis for the Chicago region are shown in Figure 3.

Figure 3:



A statistical inaccuracy of 1,000 to 3,000 people in the grand scheme of all federal legislative redistricting is quite negligible considering that each district for the U.S. House of Representatives is supposed to contain, on average, 700,000 people (Caliper 2019). But this error cannot be completely ignored. In 2010, the state of Minnesota held onto its eighth district by only a couple thousand people. With the 2020 Census on the horizon, Minnesotans are invested in a complete census count, especially considering that an analysis of reapportionment done by Election Data Services, a political consulting firm in Virginia, predicts that Minnesota could lose a seat by only a few thousand people when the districts are next redrawn (Callaghan, Peter, et al. 2020). It is possible that the difference between eight and seven seats for Minnesota, and almost 15% more political power, could come down to the statistical inaccuracy introduced by differential privacy.

In another study, census employees, John M. Abowd and Ian M. Shmutte, explored the extent to which a differentially private version of the 2010 census would impact Title 1 school funding allocations. Each year, the Department of Education allocates Title 1 Grants to public schools with financially disadvantaged students using the formula:

$$A_d = E_d \text{ x } C_d$$

A_d is the allocation of money to a specific school district

E_d is the count of Title 1 eligible students

C_d is the adjusted State Per-Pupil Expenditure (SPPE)

Abowd and Shmutte hypothesized that if differential privacy actually produced large enough statistical inaccuracies in the 2010 census data products, then error could potentially be quantified by measuring the misallocation of Title 1 funds. Using an epsilon value of .1, they found that the average public school district's allocation inefficiency would be about $63,000, or roughly $18 per student (Abowd and Schmutte 2018).

This study was inspired by Abowd's and Shmutte's investigation of the impact of differential privacy at granular levels of geography. While it is generally predicted that this new disclosure method will have little impact on higher levels of U.S. geography like regions or states, many social scientists and statisticians are unsure of what will happen to lower level geographies like census blocks, tribal lands, and school districts. If a state's population is off by a couple hundred, there are little to no ramifications in the grand scheme of funding, but in such granular geographical tracts like school districts, the loss of a hundred students to do a privacy algorithm can have major consequences for grant eligibility and regional educational planning. Abowd and Shmutte's study will provide the basis of the methodology of this paper's analysis.

Methods:

In order to understand the extent to which differential privacy undermines data accuracy, this study assumed the typical role of the federal government by using census data to allocate fiscal sums to geographic regions based on their demographic characteristics as published by the census. More specifically, this study focused on two policy programs that are critical to the funding of the United States' Public School Districts: Title 1 Grants and Rural Education Achieve Program (REAP) Grants. Through the comparison of two 2010 census datasets, one differentially private and one with standard disclosure controls, this study attempted to quantify the fiscal misauthorizations that the Census Bureau is at risk of introducing in its upcoming 2020 cycle.

Data Collection:

In September 2019, the Census Bureau released a set of data products from the 2010 census which had differentially private algorithms with an epsilon value of six applied to it to allow social scientists and data stakeholders to explore the impact of this privacy measure on statistical accuracy. This study conducted identical analyses of federal allocations to school districts using two census datasets: the 2010 differentially private census dataset and the standard 2010 census dataset with standard privacy procedures applied to it.

Both datasets were downloaded through IPUMS, an open source database for longitudinal survey and census data. IPUMS provided the data in a format that made it compatible with statistical softwares like R and GIS which were later used for this analysis. It is important to note that the 2010 census datasets are in the format of cross tabulations at different geographic levels. That is, this data is not at the individual level, but rather at the level of a certain geographic entity such as a state, county, census tract, or school district. The data provided counts of certain demographic groups at each geographic level. For instance, the school district data files contained information on the number of children of each age or ethnicity that live within that school district.

An important limitation of this method of data collection is that the differentially private dataset that IPUMS provides excludes some relevant variables for each geographic unit that could increase the accuracy of this analysis. For instance, the urban/suburban/rural classification that the Census Bureau typically assigns to geographic units is missing from the differentially private dataset provided by IPUMS and these classifications are critical for determining school district eligibility for specific federal programs. This missing variable prevented this study from quantifying the exact misallocation of federal funds that require a specific population density maximum threshold for rural school districts. Additionally, the

IPUMS dataset only included unified public school districts (i.e. school districts that serve grades 9-12) in its differentially private dataset, thus excluding elementary and secondary school districts that only serve a subset of grades. Despite these limitations in the available data, this study was still able to prescribe eligibility error rates and fiscal misauthorization rates to the geographical units available in the IPUMS data which will serve to contextualize the implications of this policy for readers as the 2020 census approaches.

Data Analysis:

The federal allocation process requires census data in two stages: 1.) determining program eligibility and 2.) quantifying the formula grant authorization amount. This analysis investigated how both the standard and differentially private 2010 census dataset differ in determining which school districts qualified for four United States Department of Education grant programs and how much funding that school district received or lost as a result of their initial qualification and their population. The respective programs this analysis investigated are as follows:

- The three types of Title 1 LEA (Local Education Agency) Grants
    - Basic Grants
    - Concentration Grants
    - Targeted Grants
- One type of the Rural Education Achievement Program (REAP) Grants
    - Small, Rural School Achievement (SRSA) Grants

Program eligibility for these federal formula grants usually require that a geographic area has a certain subpopulation above or below some arbitrary threshold. For example, Title 1 Basic Grants require at least 10 children and 2% of a school district's student population to be at or below the national poverty line. Using R as my statistical software of choice, I gave a binary code of 1 to districts in the differentially private dataset that met each grant program's criteria and a code of 0 to districts that did not meet the program's criteria. I then did the same with the standard 2010 data. Through merging the two datasets, I was able to assess for each policy how many school districts qualified for the program under both datasets and how many districts were misclassified as a result of the inserted noise in the differentially private dataset.

To assess the discrepancies in program eligibility, this analysis constructed confusion matrices for each federal program investigated. A confusion matrix serves as a visually informative and intuitive tool from which a reader can discern the extent to which the differentially private and standard 2010 datasets agree with one another in their determination of program eligibility for school districts. The format these matrices follows is shown in Table 1:

Table 1: Illustration of a Confusion Matrix

	# of Eligible Geographic Areas in Standard Data	**# of Non Eligible Geographic Areas in Standard Data**
# of Eligible Geographic Areas in Differentially Private Data	True Positives	False Positives
# of Non Eligible Geographic Areas in Differentially Private Data	False Negatives	True Negatives

From these matrices, this analysis was able to calculate false positive and false negative rates. The false positive rates provides the conditional probability of the differentially private dataset falsely classifying a geographic area as eligible for a federal program given that the standard data, or the true data, tells us that it is ineligible. Meanwhile, the false negative rate offers the conditional probability that the differentially private dataset falsely determines that a geographic area is ineligible for a federal program given that the standard data already determined that the area was indeed eligible. The formulas for both of these rates are provided below:

False Negative Rate: $\frac{(False\ Negatives)}{(False\ Negatives + True\ Positives\ )} * 100$

False Positive Rate: $\frac{(False\ Positives)}{(False\ Positives + True\ Negatives)} * 100$

While ideally both the false positive and false negative ratio would be minimized between the two datasets, it is important to understand the tradeoffs between the two. The main implication that would stem from a false positive ratio being high would be that geographical areas just above the eligibility criteria for a program in the 2010 standard data will now receive funding under the 2010 differentially private data. It can be inferred that not much harm can come from providing another school district with more federal funding to augment their educational support programs, but it is important to consider that each program's total budget is limited. In some cases, one geographic area's increased funding under the algorithm could come at the expense of another equally or more deserving geographic area. That being said, high false negative ratios have much worse implications. A high false negative ratio implies that geographic areas that should truly qualify for federal funding are being falsely rejected in the differentially private data. This could result in deserving geographic areas being left behind in the federal allocation process, potentially harming at-risk youth or adult populations.

The second stage of the federal allocation process that differentially private data potentially puts at risk is the determination of the grant amount to each school district. Since all of the programs that this analysis investigated are formula grants, meaning that these grants are awarded based on some sort of statistical criteria or threshold provided in the census, noisy population estimates introduced by differential privacy can be predicted to alter the amount of money geographical areas receive. To determine how detrimental these differentially private statistics are to the accuracy of these formula grants, this analysis quantified the differences in the grant authorization and allocation amounts when using the two census datasets in the funding formulas.

This analysis uncovered the fiscal error introduced by differential privacy to federal grant programs through two distinct calculations: 1.) the difference in the two datasets' authorization amounts and 2.) the difference in the two datasets' allocation amounts. These calculations reflect the two stage process through which a formula grant program calculates funding amounts to geographic areas. Authorization amounts establish dollar ceilings on the amount that can be appropriated for a specific program according to the outputs of a federal grant formula, but do not necessarily guarantee that amount of financing for the program. Allocation amounts are determined during the appropriations process in which Congress and the President agree upon a fiscal year budget for a specific grant program. During the appropriations process it is common for the initial authorization amount for a program to be greatly reduced in accordance to the limits of the federal budget. In these cases, the percentage of the total authorization amount that a given geographical area qualifies for will be the same percentage of the decided allocation amount that the geographic area will receive. The two step calculation for translating authorization amounts to allocation amounts is shown below:

Step 1: Calculate the geographic area's authorization percentage

$$\frac{\textit{Individual Geographic Area's Authorization Amount}}{\textit{Total Authorization Amount of all Areas}} = \text{Geographic Area's Authorization Percentage (AP)}$$

Step 2: Determine what dollar amount of the program budget the authorization percentage corresponds to

(Individual Geographic Area's AP)*(Total Allocation Budget) = Geographic Area's Allocation Amount

Determining formula grant authorization and allocation amounts relies on some estimates that cannot be found in the two census datasets. For example, the Title 1 Grant formulas required the State Per Pupil Expenditure (SPPE) which had to be obtained from the National Center for Education Statistics (NCES). Once all components to each program's formula were collected and the data was cleaned, I analyzed the difference between the total standard 2010 census data's fiscal allocations for a specific grant program and the differentially private data's total fiscal allocations for the same program. This allowed my analysis to fiscally quantify the error this differential privacy contributes in the formula grant process which will be powerful in later determining whether or not it should be implemented in the 2020 census. The inputs and formulas for each program as well the results of this analysis are provided in the Discussion of Findings section of this paper.

Discussion of Findings:

The next section of this paper outlines the findings from my statistical comparison of the 2010 differentially private data to the standard 2010 data in two contexts: 1.) Title 1 Grants and 2.) Rural Education Achievement Program (REAP) Grants to unified public school districts. These findings ultimately allow readers to predict the impact of the differentially private algorithm on the future accuracy of government allocations to low level geographic units.

Title 1 Grants:

Before quantifying the actual impacts of differential privacy on government allocations to school funding programs, it is important to discuss why this spending is so critical. The federal grant program that this analysis investigates first is Title 1. Title 1 is currently the largest federal aid program for public school districts across the United States (U.S. Department of Education 2018). As of the 2015-2016 school year, 55,906 public schools received some sort of Title 1 funding for 26 million eligible children (United States Department of Education 2018). These grants provide financial assistance to local educational agencies (LEAs) that meet a certain threshold of children from low-income families in order to ensure that these students have the resources to allow for them to meet national and state academic standards. Federal funds to these school districts are allocated through formulas that take the census poverty estimates of school-aged children as well as the cost of education in each state as inputs. The table below briefly outlines the three main types of Title 1 Grants provided to school districts and the census-derived criteria that must be met for a school district to be eligible:

Table 2:

Type of Title 1 Grant	**Criteria**
Basic Grant	At least 10 children and 2% of the LEA's school-age population are at or below the poverty line
Concentration Grant	Eligible for the Basic Grants and have over 6,500 children or 15% of the LEA's school-age population are at or below the poverty line
Targeted Grant	At least 10 children and 5% of the LEA's school-age population are at or below the poverty line

Accurate census data is critical in ensuring that high-need districts receive the federal support they need to support low-income students' pursuit of a high quality education. Over the first section of this analysis, this paper will investigate whether the differentially private estimates of children living in poverty led to mistakes in determining school district eligibility for each of these Title 1 programs.

Disagreement/Agreement Rates

Basic Grants

Recall that Basic Title 1 Grants are allocated to school districts with at least 10 children and 2% of its student body population living at or below the current federal poverty line. Using both the differentially private and standard 2010 census data, I determined which districts were eligible for the Basic Title 1 Grants in each dataset and investigated the degree of agreement/disagreement between them. The two datasets yielded the following confusion matrix for the 10,858 school districts:

Table 3:

`	# of Eligible LEAs in Standard Data	# of Non Eligible LEAs in Standard Data
# of Eligible LEAs in Differentially Private Data	10,570	47
# of Non Eligible LEAs in Differentially Private Data	38	190

From this confusion matrix, I was able to calculate the false positive rate for Title 1 Grant eligibility under the differentially private data as 19.83%. The main implication of a high false positive rate is that more school districts that are on the brink of qualifying will receive additional funding to aid their students that they may not have received otherwise. While this does not necessarily seem harmful in the grand scheme of this program, it is important to remember that these federal programs exist under budget constraints set by Congress. If more schools are qualifying for Title 1 aid under the differentially private data, then there will be more LEAs sharing the same amount of funding, resulting in less funding per school district.

Another cause for concern is the calculated false negative rate for Basic Title 1 Grant eligibility, which in this case is .358%. To put this statistic in context: out of all the districts that truly are eligible for Basic Title 1 Grants in the standard dataset, .358% would not qualify for the funding they need to support their low-income populations as a result of the differentially private algorithm. While this percent seems almost negligible, when I aggregated all the children living in poverty in these false negative districts, I found that up to 551 children would not receive the proper educational funding needed for them to succeed in the classroom.

Concentration Grants

The criterion for a school district to qualify for a Title 1 Concentration Grant is more stringent than it is for Basic and Targeted Grants. To qualify, the district must have at least 6,500 or 15% of their children living at or below the poverty line. The program is designed to provide extra assistance to large, economically disadvantaged school districts. Since differential privacy inserts more noise to geographic areas with smaller populations, it is predicted that there should be less discrepancies in eligibility between the two datasets for a program that is designed for school districts with higher population counts. The actual results of two datasets agreements/disagreements are shown in Table 4 below.

Table 4:

	# of Eligible LEAs in Standard Data	# of Non Eligible LEAs in Standard Data
# of Eligible LEAs in Differentially Private Data	6620	1
# of Non Eligible LEAs in Differentially Private Data	0	4225

There are no false negatives found in this confusion matrix, indicating that the differentially private dataset performs quite well in determining eligibility for Concentration Grants. There is one instance of a false positive in the differentially private dataset, yielding a false positive rate of .024%. This false positive represents Hamilton County School District in Chattanooga, Tennessee. Since this school district reported to the census in 2010 that it had 6,494 school-aged children at or below the poverty line, it is conceivable that the differentially private algorithm increased this number up to 6,503 to qualify the district for a Concentration Grant. This is an instance in which differential privacy has positive implications on fiscal allocations. While it would be expected that a district with 6,494 impoverished students should face the same challenges as a school district with 6,500 students, because of the seemingly arbitrary eligibility cut off for this Title 1 program, the latter district would receive funding while the former would not. What is important to remember though is that the insertion of statistical noise in a differentially private algorithm is random. Therefore in theory, schools just above this eligibility cutoff in the standard 2010 census dataset could have just as likely been bumped below the cutoff and lost funding in the differentially private dataset.

Targeted Grants

Targeted Title 1 Grant eligibility functions very similarly to Basic Title 1 Grant eligibility, with the caveat that instead of 2% of the school-aged population being at or below the poverty, Targeted Grant eligibility requires 10%. The confusion matrix for Targeted Grant eligibility for the two datasets is shown below.

Table 5:

	# of Eligible LEAs in Standard Data	# of Non Eligible LEAs in Standard Data
# of Eligible LEAs in Differentially Private Data	8655	38
# of Non Eligible LEAs in Differentially Private Data	29	2123

The false positive rate for this program comes out to 1.758%, significantly less than that of the Basic Title 1 Grants (19.83%). The calculated false negative rate of .33% indicates that for every 300 Targeted Grant eligible school districts, one of these districts will not receive funding under the differentially private dataset. The total number of impoverished students that would be missed for Targeted Grant under as a result of this privacy measure totals to 431. It should be alarming that 431 students in these misclassified districts are at risk of not receiving the adequate resources needed to ensure equitable educational opportunities for their most needy students.

Overall, the extent to which differential privacy impacts Title 1 Grant eligibility is dependent on the qualification criteria for each type of grant. While there was almost complete agreement between the differentially private and standard data for Concentration Grant eligibility, the two datasets disagreed on many school districts' eligibility for Basic and Targeted Grants. As mentioned previously, this is likely the result of the Concentration Grant

Program favoring larger school districts whose school-aged populations counts will not differ as much under a differentially private algorithm as they small school districts with small populations would. Which leads to the next census-driven U.S. Department of Education program this analysis will investigate: the Rural Education Achievement Program (REAP).

Rural Education Achievement Program (REAP):

Equitable outcomes for school-aged children living in poverty have historically been a priority for the U.S. Department of Education, but increasingly a new subset of children are at risk of being left behind by the public school system: rural children. In 2017, the United States Department of Agriculture (USDA) released a study which highlighted that 467 U.S. counties had working age populations in which at least 20% lacked a high school diploma, double the national average of 10%. The USDA found that 80% of these counties were classified as rural by the Census Bureau (Joseph 2017). Similarly, the New York Times reported in 2017 that only 29% of 18-24 year old adults in rural areas were enrolled in a post-secondary institution, compared with 47% of their urban peers (Congressional Research Service 2017).

Similar to Title 1 Grants, policies that facilitate improvements to the rural education system require an accurate understanding of the number and proportion of school aged children in each school district. One such policy, the Rural Education Achievement Program (REAP), allocates two types of formula grants to schools in rural areas: Small, Rural School Achievement (SRSA) programs that serve LEAs with small numbers of students and Rural and Low-income School (RLIS) programs that serve LEAs with high concentrations of low-income students, regardless of its size. Eligibility for both of these programs strictly depends on Census Bureau data. The program this analysis focuses on, the Small, Rural School Achievement Program, allocates a base grant of $20,000 to school districts which have an average daily attendance of less than 600[6]. For each student over the initial threshold of 50, the district receives an additional $100 up until the maximum grant amount of $60,000 is reached (Congressional Research Service 2017).

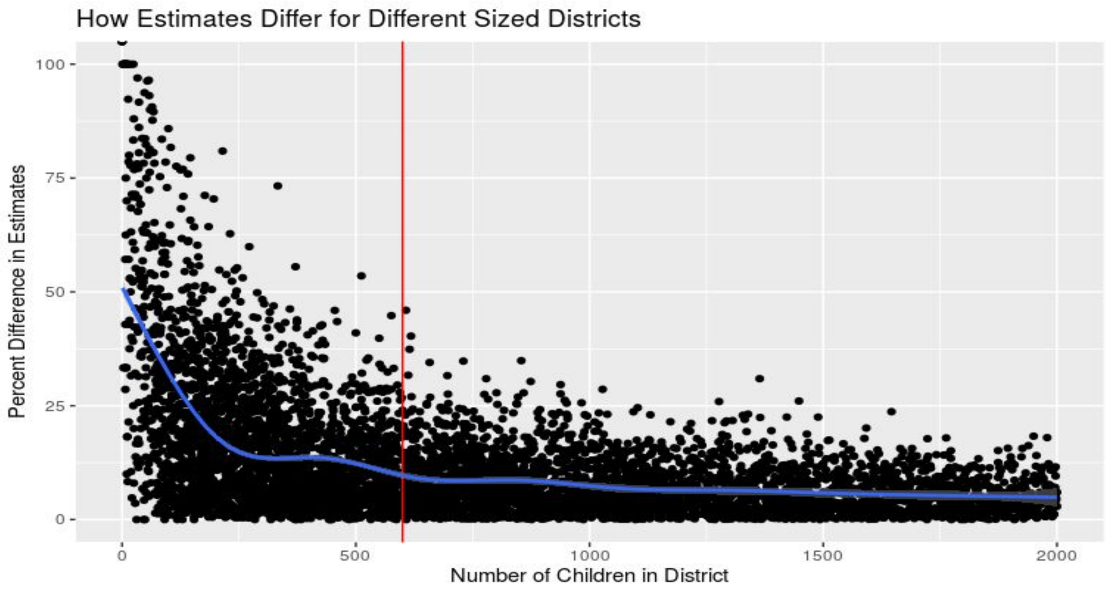
Before I investigated the agreement/disagreement between the two datasets for SRSA Grant qualification, I was interested in visualizing the extent to which rural school districts are disadvantaged in the accuracy of statistics published about them by the Census Bureau under differentially private algorithms in comparison to larger school districts. To conduct this analysis first calculated the percent difference between the total count of school aged children in both datasets using the equation below:

$$\left|\frac{(Standard\ Data\ Children\ Ages\ 5{-}17) - (Differentially\ Private\ Children\ Ages\ 5{-}17)}{(Standard\ Data\ Children\ Ages\ 5{-}17)}\right| * 100 = \text{Percent Difference}$$

I then plotted this calculated percent difference for each school district against their true population counts in the standard 2010 census in Figure 4.

[6] Additionally, a school district can qualify on the sole criteria that it is located in a region with a population density of fewer than 10 people per square mile. For simplicity, this analysis assumed all LEA's have an average daily attendance that is directly congruent to their school-aged population.

Figure 4:



As Figure 4 demonstrates above, as the number of children in the standard dataset increases, the differentially private estimates approach the true population parameter. I chose to include a vertical line where the LEA school-aged population equals 600 since this is the arbitrary threshold that determines whether or not a school district can qualify for a SRSA grant. Around this threshold, the average percent difference between the two datasets still hovers around 15-20%, which causes one to predict that differential privacy could drastically alter and falsey determine which school districts are considered rural and which are not.

The scatter plot also highlights how noisy rural school districts' population counts are in comparison to larger school districts under a differentially private algorithm. When quantified, the average percent difference between the standard 2010 school-aged estimates and differentially private estimates for schools districts with a total school-aged population under 600 is 24.64%. In contrast, the average percent difference for school districts with a total school-aged population at or above 600 is 4.79%. After running a difference in means t-test in R, I found that these two means are significantly different at the p>.0001 level. These results are summarized in Table 6 below:

Table 6:

LEA's 2010 School-Aged Population (5-17)	Average Percent Difference from 2010 Differentially Private Estimates
< 600	24.4%
≥ 600	4.79%

While a LEA's school-aged population is not the sole criteria for qualifying for a REAP Grant, the results above give this analysis a sense of how school districts with small student populations are disproportionately at risk for receiving inaccurate funding for their children as a result of the use of differentially private algorithms in the census. Students who already experience a disadvantage in their education due to their rural residence may find their situations exacerbated if their schools will no longer qualify for the aid promised to them by the federal government. The disagreements in eligibility between the standard procedure 2010 census data and the 2010 differentially private data are shown in Table 7.

Table 7:

	# of Eligible LEAs in Standard Data	# of Non Eligible LEAs in Standard Data
# of Eligible LEAs in Differentially Private Data	2736	109
# of Non Eligible LEAs in Differentially Private Data	144	7869

Table 7 yields a false positive rate of 3.83% and a false negative rate of 1.80%. While both rates seem relatively low, it is concerning that the false negative ratio is higher for SRSA Grants than any of the Title 1 Grant programs. This analysis found that under the differentially private data, 71,262 children residing in rural school districts would not qualify for the SRSA Grants needed to help them close the achievement gap with their urban peers. This is not completely unexpected. Since a differentially private algorithm cannot change a true population statistic into a negative value, small parameters, instead of decreasing in value, actually tend to get larger as a result of the mechanism. Therefore, truly small school districts are falsely granted extra students due to the additional noise differential privacy inserts. While receiving higher population counts is ideal for LEAs who are trying to qualify for Title 1 Grants, where any increase in the number of students living in poverty helps the district qualify for additional government funding, rural school districts whose student population counts are bumped up by the algorithm are actually disadvantaged as they can no longer qualify for REAP Grant programs.

Quantifying Fiscal Error:

Determining LEA eligibility for federal grant programs is only one part of the census's role in the authorization and allocation of fiscal programs to localities. Census data is also used to inform the exact amount of money that a school district receives for a given federal grant program. Oftentimes, federal grant amounts are authorized per individual eligible for the given program residing within some specific geography. For example, under the Title 1 program, eligible school districts are authorized approximately two-fifths of their most recent adjusted State Per Pupil Expenditure (SPPE) for each student that lives at or below the poverty line in their district's boundaries. Similarly, LEAs that qualify for REAP Small, Rural School Achievement (SRSA) Grants receive $100 per student in their district until they reach the maximum authorization threshold of $60,000. After the authorization amount for each district is determined and totaled, the U.S. Department of Education submits the proposal for the federal grant program to be approved and appropriated by Congress. Congress then uses this information to determine a budget for the program which in turn decides the actual allocations that each eligible school district receives.

Since any statistical noise introduced by differentially private algorithms is likely to alter the population counts in low level geographic areas, it is likely that the authorization and allocation amounts to LEAs for these U.S. Department of Education programs will differ significantly between the differentially private and the standard 2010 census data. Through quantifying the difference between the two dataset's allocation and authorization amounts to school districts for each of the chosen grant programs, this analysis provides readers with a rough sense of the dollar cost of this new privacy policy before it is implemented in 2020.

Formula for Basic and Concentration Title 1 Grants:

While Basic and Concentration Title 1 Grants have different eligibility criteria for their programs, their funding formulas are the exact same. For both grants, each school district is authorized, on average, two fifths of their state's annual per pupil expenditure for every formula-eligible child (i.e. every child living at or below the federal poverty line). The formula is written below:

Authorization Amount = .4(State Per Pupil Expenditure)(Formula Eligible Children)

Under the standard 2010 census data, a total of $40,549,666,071 was authorized to qualifying LEAs for the Title 1 Basic Grant program. When the standard 2010 census counts of formula-eligible children were replaced with the differentially private 2010 census counts, a new total of $40,633,992,589 was authorized for the program. The use of the differentially private data resulted in an additional $84,326,518 in total authorized funding for Basic Grants. This analysis discovered that a total of $1,015,431,840 would have been misallocated in 2010 if differentially private algorithms had been used in the census.

A similar level of error is detectable between the two datasets in the authorization amounts for Title 1 Concentration Grants. While there were few disagreements between the differentially private and the standard data in the eligibility of LEAs for Concentration Grants, the total amount of grant money misauthorized for the program under the treated dataset amounts to $832,765,724. While the standard dataset authorized a total of $33,571,213,982, the differentially private dataset authorized $33,708,354,007, a difference of $137,140,025.

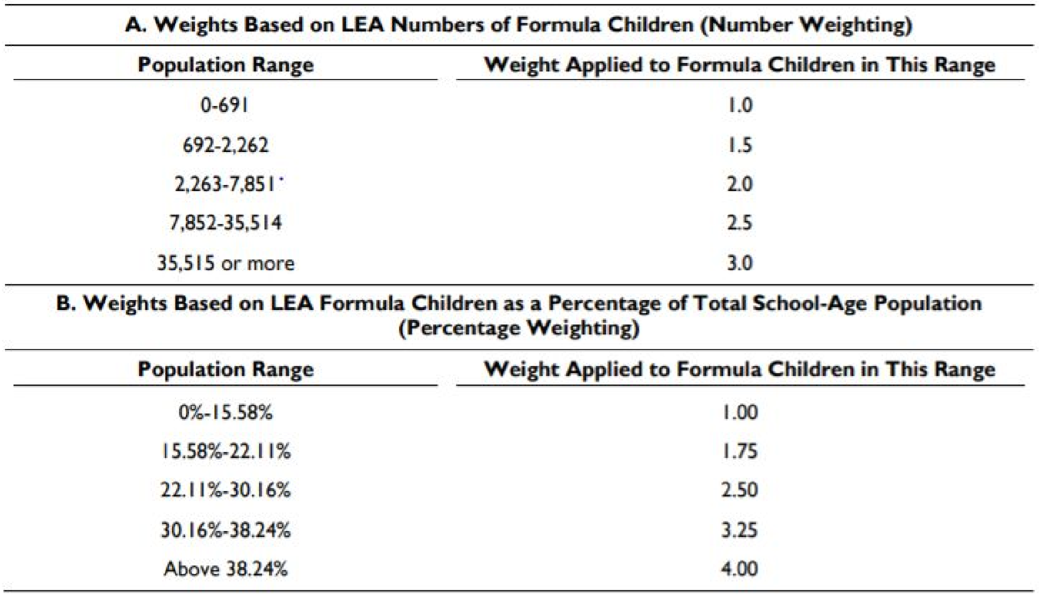
While these discrepancies in the total authorization amounts between the two census datasets is in the magnitude of billions of dollars, it is critical to remember that authorization amounts are not synonymous with allocation amounts. While the U.S. Department of Education submits an authorization amount as the total amount a federal program *should* receive, the amount the program *actually* receives is constrained by the budget set by Congress. In the 2010 fiscal year, the total budget for Title 1 Basic Grants was $6.4 billion. For Title 1 Concentration Grants, the budget was significantly smaller at $1.3 billion (United States Department of Education 2018).

This analysis found that out of the total $6.4 billion allocated to Basic Grants, $160,735,876 would be misallocated under the differentially private data. For Concentration Grants, out of the $1.3 billion budgeted for the program, $34,029,155 would be misallocated if the differentially private data was used. In both cases, this misallocation amount represents about 2% of the total budget for the program.

Funding Formula for Targeted Title 1 Grants:

While similar to other Title 1 Grants in their qualifying criteria, Targeted Title 1 Grants use a more complex authorization formula to determine the amount of federal aid eligible LEAs should receive. Unlike the Basic and Concentration Grant formulas, which authorize an equal amount of funding to each eligible child across all districts in a state, Targeted Title 1 Grants assign weights to each child on the basis of each LEA's formula child rate (i.e. the percent of the age 5-17 population in that area that lives at or below the poverty line) and the number of formula eligible children. The weights for the Targeted Grant formula are shown in Table 8 below:

Table 8:



A. Weights Based on LEA Numbers of Formula Children (Number Weighting)

Population Range	Weight Applied to Formula Children in This Range
0-691	1.0
692-2,262	1.5
2,263-7,851	2.0
7,852-35,514	2.5
35,515 or more	3.0

B. Weights Based on LEA Formula Children as a Percentage of Total School-Age Population (Percentage Weighting)

Population Range	Weight Applied to Formula Children in This Range
0%-15.58%	1.00
15.58%-22.11%	1.75
22.11%-30.16%	2.50
30.16%-38.24%	3.25
Above 38.24%	4.00

As the weights in Figure 8 demonstrate, the higher a school district's number of formula-eligible children and/or formula child rate are, the more funding they will receive per child under the Targeted Grant program. In determining the exact authorization amount, the formula considers two weights per school district: 1.) each LEA's number of formula-eligible children (i.e. number weighting) , and 2.) each LEA's formula child rate (i.e. percentage weighting). Whichever type of weighting results in a higher grant per child is ultimately used in determining the authorization amount for that given school district. In the

2017 fiscal year 88% of LEAs that qualified for Targeted Grants were authorized money under percentage weights as opposed to number weights (Skinner and Rosenstiel 2018).

Once either the percentage or number weights are multiplied by the number of formula-eligible children in the qualifying school district, the authorization amount is calculated using the following formula:

Authorization Amount = .4(State Per Pupil Expenditure)(Weighted Formula Eligible Children)

Using only the 2010 standard data products, a total of $85,005,026,649 would be authorized to a total of 8,684 public school districts across the nation. The 2010 differentially private data products yielded an authorization amount of $85,354,876,607 to 8,693 school districts, for a total difference of $349,849,958. Ultimately, a total of $1,825,586,480 in federal authorizations to LEAs would be misallocated as a result of the differentially private algorithm.

In 2010, the total budget for Title 1 Targeted Grants was $3.3 billion. Under this fiscal constraint, I calculated that a total of $71,934,211 was misallocated when the differentially private data was used as opposed to the standard data. Similar to Basic and Concentration Title 1 Grants, using differentially private algorithms to determine the number of formula-eligible students per school district would result in about 2% of the total Targeted Grant program being misallocated.

Funding Formula for REAP Small, Rural School Achievement Grants:

The Rural Education Achievement Program employs a more rigid formula that requires accurate census child population counts in order to ensure that small school districts receive accurate federal funding. LEAs that qualify for the Small, Rural School Achievement (SRSA) Grant are awarded a base grant of $20,000 plus an additional $100 for each student over the population threshold of 50 in their district until the 550th student is reached, at which point the grant maximizes at $60,000. The formula is summarized in Table 9 below.

Table 9:

Age 5-17 Population in Census Data	**SRSA Authorization Formula**
≤ 50	20,000
> 50 and ≤ 550	20,000 + 100(# of School-Aged Children - 50)
> 550 and ≤ 600	60,000

When the standard 2010 census data was used to input the child population counts in each eligible school district, a total of $100,354,800 was authorized under the SRSA program. When the differentially private 2010 data was used, a new total of $98,647,100 was authorized. Ultimately, the use of differentially private child counts in the funding formula resulted in $1,707,7000 less in the total authorization amount for the SRSA program and $20,719,212 in total misauthorizations.

The total budget for the REAP Small Rural School Achievement Program in 2010 was $84.9 million (United States Department of Education 2010). This appropriation amount is not much smaller than the authorization amounts calculated above. Given this budget constraint, this analysis found that $18,173,864 would be misallocated if the differentially private algorithm were to be used on the child counts. This represents 21.4%, just over a fifth, of the entire SRSA program.

Through quantifying the fiscal error that differential privacy introduces in the authorization and allocation amounts in Title 1 and REAP Grants, this analysis uncovered the magnitude to which this new privacy policy can harm the dissemination of federal money to needy public school districts. Most notably, this analysis discovered the extent to which rural school districts would have suffered had differentially private algorithms been used in 2010. While all three Title 1 Grants saw only 2% of their total funding being misallocated under the treated dataset, the SRSA program had 21% of its total budget misallocated. As the United States increasingly becomes concerned about the urban/rural divide in the context of broadband, healthcare, education, and other sectors, it is critical that the Census Bureau publishes accurate statistics about rural geographic areas so that they can fairly benefit from the federal support they may need. The next section of this paper outlines three policy recommendations I offer as potential alternatives and remedies to the Census Bureau's decision to implement differential privacy in 2020.

Policy Recommendations:

While differentially private algorithms offer impressive privacy guarantees to census respondents, my research concludes that an injection of statistical noise into a dataset with many critical uses in the policy context may be detrimental to specific census-driven federal grant programs at low level geographies. Using findings from this study's analysis and the existing literature on differential privacy, this paper provides three recommendations to the Census Bureau on how to best protect the privacy of its data in the digital age while preserving its statistical accuracy. The three outlined proposals will be described in the order of most to least feasible to implement.

Proposal #1: *Set the Epsilon parameter high enough to promote data accuracy at low level geographies*

As described earlier, differential privacy is constrained by an epsilon parameter which quantifies the amount of privacy loss in a database under the algorithm. As epsilon increases, the statistical accuracy of a database increases proportionally to the detriment of the privacy of individual respondents in the dataset. Therefore, privacy advocates support lower levels of epsilon while Census data stakeholders tend to favor higher values.

The level of epsilon for the 2020 census has yet to be announced by the Census Bureau and is not expected to be released until several months after the decennial census has taken place. The data this paper used for it's analysis used an epsilon value of six and found that this value is still capable of having significant impacts on the eligibility and authorization of federal resources to public school districts.

The first policy recommendation this paper proposes is that the Census Bureau sets its epsilon parameter in its algorithm in 2020 to a value higher than six, in hopes of preserving statistical accuracy, and therefore promoting informed and equitable federal allocations to small geographic areas. A higher value of epsilon would enable the federal government and other data users to explore and appropriately use correct demographic and geographic information provided by the census while still preserving the privacy of individual respondents. Since the Census Bureau has yet to commit to an exact value of epsilon for its 2020 cycle, this policy proposal seems both feasible and sensitive to the short timeline on which the bureau operates.

It is important to consider that the Census Bureau currently faces a difficult situation of trying to compromise with two competing coalitions: social scientists and privacy advocates. These groups are lobbying for two characteristics of the upcoming census data that are increasingly in conflict with one another: statistical accuracy and individual privacy. If the bureau chooses to implement a higher value of epsilon into its 2020 datasets, they will likely be met with hostility and concern from privacy advocates. That being said, privacy advocates have yet to publicly coalesce around a single desired level of epsilon for the 2020 census so it is currently unknown what their response would be to a value higher than six.

Proposal #2: *Explore alternative privacy techniques*

There is compelling evidence as to why the census's old privacy technique of data swapping is no longer effective. After the Census Bureau's internal database reidentification attack alarmed stakeholders that individuals could be identified through microdata tables, the census needed to protect its integrity and legitimacy to encourage participation and foster privacy. Now that the world is well into the digital and big data age, stakeholders should recognize that alternative privacy protections exist which are more effective than data swapping, but less detrimental to statistical accuracy than differential privacy.

The Census Bureau currently employs several methods of disclosure avoidance for different datasets it oversees. The 2010 Census and the American Community Survey Public Use Microdata (PUMS), which contain data at both individual and geographic level, are currently protected through a variety of different privacy mechanisms such as rounding schemes and noise infusion. Rounding schemes collapse exact numerical variable values for an individual into rounded categories. For example, individual and organizational Property Tax payments are rounded according to the following scheme (Gauthier 2019):

$0 remains $0

$1-7 rounded to $4

$8-$999 rounded to nearest $10

$1,000-$49,999 rounded to nearest $100

$50,000+ rounded to nearest $1,000

Another technique, *noise infusion*, is similar to differential privacy in many respects. The difference is that noise infusion only involves the insertion of noise into specific levels of categorical variables. The result is that exact ages of respondents are masked, but the overall grouping frequencies for statistics are not affected. For example, if a census microtable uses the following categories for age: 0-9, 10-19, 20-29, etc., an individual who is actually 17 may be listed as 14 years old after noise is infused to the dataset. Under this technique, it would be impossible to change an individual's age to a number such as 9 or 21, which is outside of his/her age category.

Another way to avoid a database reconstruction attack is to cease the publication of any census microtables that can be retrospectively manufactured to expose individual level characteristics of respondents. In many cases, the microtables that are most susceptible to a reconstruction attack are those that contain the information of relatively few individuals such as a table of the race/ethnicities of those residing on a given census block. To resolve this threat, the Census Bureau currently sets a population threshold of at least 100,000 people for before it publishes a public-use microdata file[7] for a geographic area to avoid individual identification.

While the Census Bureau increasingly fears that computational sophistication undermines these previously used privacy techniques, they have yet to demonstrate that their data is at risk of being reconstructed by an external adversary. As Steven Ruggles pointed

[7] A Public Use Microdata Area (PUMA) is defined by the Census Bureau as a geographic unit used by the US Census for providing statistical and demographic information. PUMAs are typically at the sub-state level. Following the 2000 Census, there were a total of 2,071 PUMAS.

out, these previous disclosure avoidance techniques were only vulnerable to reidentification attacks by high level internal officials at the Census Bureau. Therefore it is not necessarily urgent that the Census Bureau drastically changes its privacy techniques this 2020 cycle.

Synthetic data is another disclosure avoidance method the Census Bureau has explored in its other survey work, but has yet to use in the decennial census. Synthetic data from a survey can be created using posterior predictive models which generate data that has many of the same properties as the original dataset. For each variable in the original dataset, a regression model is created. Then, for each record, the actual value of the variable is replaced by the model's new imputed value. A researcher or data user can then draw random samples from these synthetic survey populations to answer database queries . Since these populations are synthetic, database queries would not result in the reidentification of any individuals.

The privacy techniques described above demonstrate the wide portfolio of options available to the Census Bureau in 2020. This analysis revealed that differential privacy impedes the federal government from making accurate allocations to small geographic areas. It is therefore critical that the bureau considers scaling up alternative privacy techniques that are less prone to statistical error and noise to protect the integrity of decennial census data.

While upscaling other privacy techniques would be a favorable alternative to differential privacy, it is important to consider that the Census Bureau has already exhausted many human and financial resources on developing differentially private algorithms for its datasets. In 2020, the Census will be working under one of its tightest budgets to date, and as a large government agency, it is not quickly adaptable to methodology changes at the last minute. Because the census is currently in the field, the bureau will have little time to publicly release policy changes to Congress for approval in time for the anticipated release of the census's new data products.

Proposal #3: *Allow lower level geographies to challenge and replace census counts*

The third policy proposal this analysis offers allows lower level geographies to have more agency in ensuring that their areas are receiving adequate and accurate funding from the federal government. If the Census Bureau sets the epsilon parameter of their differentially private algorithm at a level that is small enough to distort low level population estimates, I propose that the bureau allow the local governing bodies of these areas to challenge published census data counts with their own (methodologically validated) documentation of individuals residing in their area. For instance, if a rural school district notices that the differentially private algorithm is causing the census to undercount its school-aged population by 25%, its school board or district administrator could submit their own counts of their student enrollment or percent of their students above or below the poverty line.

This policy would be difficult to implement for a variety of reasons. First, it requires school districts to be vigilant and proactive in their attention to census data counts. It is possible that these individuals are too far removed from the processes of the census or too busy with other work to pay attention to inaccuracies in the decennial census. Additionally, this policy would disproportionately favor school districts with extra financial, legal, and human resources that can take the time to conduct their own population count and petition the Census Bureau. Districts with these types of resources may not be the ones who are greatly in need of the funding they are losing from the government as a result of this policy. Despite these limitations, it is important to consider that there may be a more efficient and accurate way to obtain lower level geographic information than through a nationally run survey collection process. The individuals who reside and have knowledge of these communities may be better informatents of the socioeconomic and demographic information for these areas.

Additionally, this policy would require an entire overhaul of the way in which the census currently operates. By allowing lower level geographies to, in a sense, conduct their own censuses, the Census Bureau could be accused of undermining its constitutional obligation to count the entire nation. However, the Bureau could maintain the authority of housing these statistics in one large database and disseminating them to federal programs as needed.

This policy recommendation is not completely unprecedented. In fact, the U.S. Department of Housing and Urban Development (HUD) allows geographies receiving census-based formula grants to challenge Census Bureau population projection data if they believe the projections are not adjusted to reflect recent demographic shifts in their communities (Department of Housing and Urban Development 2002). In the past, American Indian and Alaska Native populations challenged these population projections by submitting their own population estimates which used the same estimation methodology employed by the census for their particular level of geography.

If implemented, this policy has the opportunity to foster greater trust and cooperation between local governments and the Census Bureau. Hard-to-count regions like tribal lands, rural areas, and high minority-population communities would now have agency over conducting a count of their own population. They can ensure that their community receives the federal resources for which they qualify by submitting more accurate counts of their inhabitants. These communities can submit their aggregated population counts to the Census Bureau for the sole purpose of accurate federal allocations. In this scenario, the differentially private data released about the geographical area would have no implication on the accuracy of the funding that the community receives for specific programs.

Conclusion:

Given the quickly evolving and increasingly precarious relationship between individuals and their private data, the Census Bureau must decide how to best ensure that the personal information which they are constitutionally mandated to collect is protected from potential reidentification attacks. In September 2018, the Census Bureau announced its intention to guard its datasets with a differentially private algorithm in order to address ongoing concerns of the adequacy of their existing disclosure avoidance measures. As this analysis has shown, differentially private algorithms ultimately harm the ability of the federal government to accurately determine eligibility for, and distribute fiscal allocations to, low level geographies. Through the investigation of how differential privacy impacts the functionality of two U.S. Department of Education formula grant programs, this analysis found that millions of dollars spent on these programs would be misallocated and thousands of impoverished and/or rural school children would be denied proper funding as a direct result of this policy. Therefore, the Census Bureau should consider other policy options to ensure that federal formula grants can operate as intended after the upcoming release of the 2020 census data.

This analysis recommends several policy options to the Census Bureau: 1.) Set the epsilon parameter of the algorithm to a higher value that is more optimal to the operation of fiscal allocations, 2.) Implement or scale up alternative disclosure avoidance methods, and 3.) Allow local governing bodies to challenge inaccurate census statistics by facilitating their own population counts. Given the limited timeline and the tight budget that restricts the Census Bureau in its 2020 cycle, sudden drastic changes to its data operations and methodologies would be difficult to implement. Regardless, it is important that Americans, as both contributors to and beneficiaries from the Census Bureau's activities, chose to value an accurate and equitable census over a potentially detrimental algorithm.

Works Cited

Auxier, Brooke, et al. “1. How Americans Think about Privacy and the Vulnerability of Their Personal Data.” *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 31 Dec. 2019, www.pewresearch.org/internet/2019/11/15/how-americans-think-about-privacy-and-the-vulnerability-of-their-personal-data/.

Callaghan, Peter, et al. “The Four Factors That Might Prevent Minnesota from Losing a Congressional Seat after the 2020 Census.” *MinnPost*, 6 Jan. 2020, www.minnpost.com/politics-policy/2020/01/the-four-factors-that-might-prevent-minnesota-from-losing-a-congressional-seat-after-the-2020-census/.

“Deploying Differential Privacy for the 2020 Census of Population and Housing.” *Census Bureau*,www.census.gov/content/dam/Census/newsroom/press-kits/2019/jsm/presentation-deploying-differential-privacy-for-the-2020-census-of-pop-and-housing.pdf.

MervisJan, Jeffrey, et al. “Can a Set of Equations Keep U.S. Census Data Private?” *Science*, 17 Jan. 2019, www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private.

“Why Is the U.S. Census So Important?” *Population Reference Bureau*, 17 Sept. 2019, www.prb.org/importance-of-us-census/.

Hansen, Mark. “To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data.” *The New York Times*, The New York Times, 5 Dec. 2018, www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to

-report-less-accurate-data.html.

Mellnik, Ted. “What's Changing for the 2020 Census?” *The Washington Post*, WP Company, 2 Apr. 2019, www.washingtonpost.com/graphics/2019/national/census-2020-technology/.

“Undercounting Hispanics in the 2020 Census Will Result in a Loss in Federal Funding to Many States for Child and Family Assistance Programs.” *Child Trends*, 2019. www.childtrends.org/publications/undercounting-hispanics-in-the-2020-census-will-result-in-a-loss-in-federal-funding-to-many-states-for-child-and-family-assistance-programs.

“Factsheet on the Census, Confidentiality and Japanese American Incarceration.” *The Leadership Conference Education Fund*, civilrightsdocs.info/pdf/census/Census-Confidentiality-Factsheet-AAJC-LeadershipConference.pdf.

Joseph, Marc. “Crisis in Rural American Education.” *HuffPost*, HuffPost, 1 June 2017, www.huffpost.com/entry/crisis-in-rural-american-education_b.

“Title I, Part A Program.” *Title I, Part A Program*, US Department of Education (ED), 7 Nov. 2018, www2.ed.gov/programs/titleiparta/index.html.

“The Rural Education Achievement Program: Title V-B of the Elementary and Secondary Education Act.” *EveryCRSReport.com*, Congressional Research Service, 26 July 2017, www.everycrsreport.com/reports/R44906.html.

U.S. Constitution. Art.1 Section 2

Partners, Georgian. “A Brief Introduction to Differential Privacy.” *Medium*, Georgian Impact Blog, 11 Sept. 2018, medium.com/georgian-impact-blog/a-brief-introduction-to-differential-privacy-eacf8722283b.

Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National

Historical Geographic Information System: Version 14.0 [Database]. Minneapolis, MN: IPUMS. 2019. http://doi.org/10.18128/D050.V14.0

US Census Bureau. “Disclosure Avoidance and the 2020 Census.” *The United States Census Bureau*, 3 Dec, 2019. www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

Ruggles, Steven. Working Paper. “Implications of Differential Privacy for Census Bureau Data and Scientific Research”. December, 2018.

Corporation, Copyright (C) 2019 Caliper. “Maptitude Maps Show How Census Differential Privacy Radically Changes Population Counts.” *Caliper*, www.caliper.com/census-differential-privacy-maps/.

Abowd, John Maron and Schmutte, Ian M., An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices (August 15, 2018). American Economic Review, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3232398

Gauthier, Jason. “Title 13, U.S. Code - History - U.S. Census Bureau.” *Title 13, U.S. Code -*

History - U.S. Census Bureau, www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html.

Moore, Richard A. *Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets*. www.census.gov/srd/CDAR/rr96-04_Controlled_DataSwapping.pdf.

Underhill, Wendy, and Christi Zamarripa. *Differential Privacy for Census Data Explained*, www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx.

Dwork, Cynthia, and Guy N. Rothblum. *Concentrated Differential Privacy*.

U.S. Census Bureau. "How Census Data Help the Nation Respond to Disasters." *The United States Census Bureau*, 14 Apr. 2020, www.census.gov/library/stories/2020/04/how-census-data-help-the-nation-respond-to-disasters.html.

Lapowsky, Issie. "The Challenge of America's First Online Census." *Wired*, Conde Nast, www.wired.com/story/us-census-2020-goes-digital/.

Skinner, Rebecca R., and Leah Rosenstiel. *Allocation of Funds Under Title I-A of the Elementary and Secondary Education Act*. 17 Sept. 2018.

The Department of Housing and Urban Development. "The American Community Survey: CHALLENGES AND OPPORTUNITIES FOR HUD." 2002.

United States Department of Education. *Fiscal Year 2010 Budget Summary and Background Information*.

"Funding Status -- Title I, Part A Program." *Home*, US Department of Education (ED), 7

Nov. 2018, www2.ed.gov/programs/titleiparta/funding.html