EXHIBIT B

Second Expert Report of Michael Barber

Dr. Michael Barber
Brigham Young University
724 Spencer W. Kimball Tower
Provo, UT 84604
barber@byu.edu

25 March 2021

1 Introduction and Qualifications

I am an associate professor of political science at Brigham Young University and faculty fellow at the Center for the Study of Elections and Democracy in Provo, Utah. I received my PhD in political science from Princeton University in 2014 with emphases in American politics and quantitative methods/statistical analyses. My dissertation was awarded the 2014 Carl Albert Award for best dissertation in the area of American Politics by the American Political Science Association.

I teach a number of undergraduate courses in American politics and quantitative research methods.[1] These include classes about political representation, Congressional elections, statistical methods, and research design.

I have worked as an expert witness in a number of cases in which I have been asked to perform and evaluate various statistical methods. Cases in which I have testified at trial or by deposition are listed in my CV, which is attached to the end of my initial report, dated March 9, 2021.

In my position as a professor of political science, I have conducted research on a variety of election- and voting-related topics in American politics and public opinion. Much of my research uses advanced statistical methods for the analysis of quantitative data. I have worked on a number of research projects that use "big data" that include millions of observations, including a number of state voter files, campaign contribution lists, and data from the US Census.

Much of this research has been published in peer-reviewed journals. I have published nearly 20 peer-reviewed articles, including in our discipline's flagship journal, *The American Political Science Review* as well as the inter-disciplinary journal, *Science Advances*. My CV details my complete publication record.

The analysis and explanation I provide in this report are consistent with my training in statistical analysis and are well-suited for this type of analysis in political science and

[1] The political science department at Brigham Young University does not offer any graduate degrees.

quantitative analysis more generally.

I have been asked to evaluate and explain at an approachable level the process of differential privacy (DP), its application to the 2020 Census, and how it fits within the field of probability theory and statistical methods.

2 The Process of Differential Privacy in the US Census

This section provides a very basic explanation of the differential privacy and post-processing procedure that the US Census Bureau plans to implement in the 2020 Census and how the process is a straightforward application of common statistical methods. The process can be divided into three basic steps. While the application of these steps across millions of geographic units and sub-groups of the population requires complicated mathematical and statistical methods as well as immense computational capacity, the concepts are in fact quite simple to describe.[2]

The Census Bureau argues that their method of differential privacy and post-processing does not fall inside the definition of statistical inference because they are not using "the drawing of inferences about a population based on data taken from a sample of that population (pg. 7 of Department of Commerce reply)." However, this definition of statistical inference is overly narrow. Statistical inference also refers to other processes aside from the definition provided by the Department of Commerce. Researchers often use datasets that include the entire population of data and still make inferences, or comparisons that are intended to inform us of differences that exist across the population. For example, suppose I had health information for the entire United States population and was looking at the variation in rates of heart disease. From these data I might learn that there are large differences across the country geographically in the rate of heart disease, as well as differences based on various demographic traits. Furthermore, I might then draw comparisons between the geographic

[2]This description is not intended to be a complete nor technical explanation of the differential privacy and post-processing procedure. Nevertheless, the basic principles outlined here are helpful in understanding how the process works.

differences versus the demographic variation. These inferences are no less "statistical inference" because they came from the population rather than a sample of the population. While our discussion of the variation would differ from a discussion of statistical uncertainty that comes with using samples of data (and the associated sampling error), we would nonetheless still be interested in the variation associated with race, or age, or some other trait compared to the natural variation that occurs across other features of the population. Thus, statistical inference can also include making comparisons across a population, and not just a sample.

This applies to the differential privacy and post-processing method proposed by the Census Bureau because they are engaged in a similar process as described above. Using data on the entire population, they are using a sophisticated statistical algorithm to learn about differences, or variation, in the population. In this case, they are interested in variation in demographic parameters across the country that might lead to leakages of privacy. Once those groups, or subgroups, of people have been identified, they then apply the parameters of their model to inject noise and further adjust that noise via post-processing to produce the confidential dataset. Evens et. al (2020) describe the process in the following way: "privacy researchers typically begin with the choice of a target (confidential) *dataset*, add *privacy-protective procedures*, and then use the resulting *differentially private dataset or analyses* to infer to the confidential dataset (pg. 3, emphasis in original)."[3] Thus the process of differential privacy and post-processing is using information from the population that inform the choice of probability distributions that are then sampled from to generate noise that creates a confidential dataset that infers, or is a "noisy" estimate of, the original population. From top to bottom, the process of choosing the degree of statistical noise to inject into the dataset, the process by which that noise is introduced, and the adjustments made afterward to comply with various constraints, is an exercise in statistical inference.

[3]Evans, Georgina, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. "Statistically valid inferences from privacy protected data." URL: GaryKing. org/dp (2020).

Step 1: Obtain accurate counts of people and geographies

The first step is to gather the actual numbers of people, their race, ethnicity, housing status, and geographic location. The Census Bureau notes that this step is largely accomplished via self-reports from individuals throughout the country with a significant amount of follow-up by Census workers.

Step 2: Inject statistical noise

The second step is to inject a certain degree of statistical noise into the data. This process is referred to as "differential privacy". There are a variety of approaches to differential privacy, and the proposed approach taken by the Census Bureau relies on basic statistical methods. At its core, differential privacy is an exercise in probability theory, and "[p]robability is the foundation and language for statistics."[4] In describing differential privacy as a question of probability and statistical methods, Bambauer et. al (2013) state, "[D]ifferential privacy disclosure occurs when the probability that a query will return a particular result differs from the probability that a query would return that same result if the person were not included in the database. It also ensures that the inclusion of a person who isn't in the dataset wouldn't change the results of a query by too much. The measure of the disclosure for a particular query to a particular individual is the ratio of the probabilities that the query system would return the result with and without the individual's data. Ideally, this ratio would be 1, allowing no disclosure at all. But since this is impossible to achieve if the responses are to be useful, the data curator can select some small level of disclosure that society is willing to tolerate. The closer to 1 the ratio is, the less disclosure has taken place."[5] In other words, differential privacy is a process by which statistical noise is injected into the original data counts so as to obscure the true values in order to lower the probability

[4]Hwang, Jessica., Blitzstein, Joseph K.. Introduction to Probability, Second Edition. United States: CRC Press, 2019.

[5]Bambauer, Jane, Krishnamurty Muralidhar, and Rathindra Sarathy. "Fool's gold: an illustrated critique of differential privacy." Vand. J. Ent. & Tech. L. 16 (2013): 701.

that an individual's identity and accurate information can be inferred from the data. The greater the noise, the lower this probability.

To determine the amount of noise to be injected into a particular quantity of interest (e.g. the number of men residing in a particular census block), the Census Bureau first determines the amount of overall privacy needed (the "privacy budget") and where to allocate that budget (for example how much to apply at the national, state, county, tract, block group, and block level). This budget is referred to by the greek letter epsilon. The size of epsilon determines the amount of statistical noise that is injected into the original, accurate counts.

In an interview with Science Magazine, John Abowd, chief scientist and associate director for research at the Census Bureau and Jerry Reiter, a professor of statistics at Duke University who has worked as a consultant with the Census Bureau discussed how epsilon is chosen. "Abowd says the privacy budget 'can be set at wherever the agency thinks is appropriate.' A low budget increases privacy with a corresponding loss of accuracy, whereas a high budget reveals more information with less protection. The mathematical parameter is called epsilon; Reiter likens setting epsilon to 'turning a knob.' And epsilon can be fine-tuned: Data deemed especially sensitive can receive more protection. The epsilon can be made public, along with the supporting equations on how it was calculated."[6] The Census Bureau has said, regarding the choice of epsilon, "Decisions about the privacy-loss budget (epsilon) for decennial products are made by a committee of senior career Census Bureau data experts, the Data Steward Executive Policy Committee (DSEP). The DSEP will analyze the results of internal and external research on the fitness-of-use of the 2010 Demonstration Data Products to make an informed decision on the level of epsilon for the 2020 Census data."[7] In other words, the Census Bureau is using information from the population and distribution of various demographics in the population to learn about and make statistical inferences regarding the total size of the privacy budget and the degree to which certain

[6] https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private

[7] https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html

places and people's information needs more or less of that privacy budget allocated.

Once the privacy budget has been allocated, the Census Bureau will take draws from a probability distribution (think of a "draw" as akin to rolling a dice, but where the die has more than 6 sides and the probability of each side coming up is not equal) which will then be added to or subtracted from the accurate counts. The chosen value of epsilon has a direct relationship with the particular shape of the statistical distribution.

Probability distributions are a foundational tool upon which much of statistics operates. Using a probability distribution to inject statistical noise allows the researcher to be mathematically rigorous (as opposed to making ad hoc decisions about where and how much noise to inject) in describing the process by which the amount of statistical noise to be introduced is determined while simultaneously making it impossible for a person to reverse engineer the precise values by which counts are added to or subtracted from in any given case because draws from probability distributions are randomly determined. "Randomization is essential; more precisely, any non-trivial privacy guarantee that holds regardless of all present or even future sources of auxiliary information, including other databases, studies, Web sites, on-line communities, gossip, newspapers, government statistics, and so on, requires randomization."[8] Thus, the application of differential privacy can ultimately be considered as a particular application of probability, sampling, and statistics. The greater the statistical noise injected into the data, the lower the probability of a record linkage successfully occurring and privacy being revealed. Similarly, the smaller the statistical noise introduced into the data, the higher the probability of someone successfully identifying individuals included in the data.[9]

In the case of the 2020 Census, the Census Bureau has indicated that the particular probability distribution that they will use is the Laplace distribution, which is displayed in

[8]Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." Foundations and Trends in Theoretical Computer Science 9, no. 3-4 (2014): 211-407.

[9]See Dwork, Cynthia, and Adam Smith. "Differential privacy for statistics: What we know and what we want to learn." Journal of Privacy and Confidentiality 1, no. 2 (2010). for a technical discussion of the ideas presented in this paragraph.

Figure 1: **Example Laplace Distribution** - The Laplace distribution is a symmetric probability distribution. The distribution can be steeper (black line) or flatter (green line) depending on the value set for the shape parameter. The higher the distribution on the vertical axis, the more likely are draws from the distribution (values on the x-axis) to have values in that region. For example, a draw of 1 is more likely than 5 or -3.

Figure 1 below.[10] The Laplace distribution is symmetric and rises to a single peak in the middle of the distribution. The higher the distribution on the vertical axis, the more likely are draws from the distribution to have values in that region. In other words, since the Laplace distribution is centered around zero, small numbers near zero are more likely to be drawn than are larger negative or positive numbers.

The particular spread of the Laplace distribution is determined by setting a shape parameter (epsilon), which can make the distribution more or less "flat." The flatter the distribution (the green line in Figure 1), the more likely are draws to have larger values (either positive or negative) while a steeper distribution (the black line in Figure 1) is more likely to have draws with smaller values. In other words, the choice of the distribution's spread (which is determined by the Census Bureau) injects more or less noise, on average, into the population counts depending on how "flat" the Census Bureau decides to make the Laplace distribution. Once these draws have been taken, the particular values are then

[10] In some cases the Census Bureau has indicated they use a two-sided geometric distribution, which is similar in shape to a Laplace distribution.

added to, or subtracted from, the original, accurate counts. Table 1 below shows a simplified example of this process in three steps. The top table shows the accurate counts of people in an area (such as a census block) based on their gender and race. The middle table then shows how statistical noise based on samples drawn from the Laplace distribution are added to or subtracted from the accurate counts.

The Department of Commerce's reply report states that "Plaintiffs assert that differential privacy is a 'statistical method' — and perhaps it is in a colloquial sense — but the reasons they offer in support of that conclusion are untethered from the express statutory definition of 'statistical method' found in Section 209's text (pg. 7)." While I have not been asked to speak to the relevance of differential privacy to Section 209, it is curious that the Department of Commerce refers to differential privacy as a statistical method "in a colloquial sense." In fact, it is difficult to know what this even means. It is hard to imagine how differential privacy, which at its most basic level is adding or subtracting values sampled from a probability distribution function, could be seen as anything but an exercise in statistical methods. Probability, sampling, and the use of probability distributions, sit at the very foundations of statistics. It would be hard to find a statistics textbook that didn't include a discussion of these ideas or that didn't devote significant page space to the development of probability theory and probability distributions.[11]

[11] See for example:
Diez, David., Barr, Christopher., Çetinkaya-Rundel, Mine. OpenIntro Statistics. United States: OpenIntro, Incorporated, 2019.
Imai, Kosuke., Bougher, Lori D.. Quantitative Social Science: An Introduction in Stata. United States: Princeton University Press, 2021.
Hwang, Jessica., Blitzstein, Joseph K.. Introduction to Probability, Second Edition. United States: CRC Press, 2019.
all of which are used in introductory statistics courses at Harvard and Princeton Universities.

Step 3: Address fractions, negative numbers, and internal consistency

In some cases the process of differential privacy would be complete after the researcher adds or subtracts from the original data the particular random draws that arose from sampling from the chosen statistical distribution. However, census data require several additional steps to address constraints that arise from the need for the differential privacy process to align with other objectives related to the use of census data. These steps are collectively referred to by the Census Bureau as "post-processing."

The first issue centers on the need for counts of people, housing units, and other statistics to be reported as integers (as opposed to fractions). The middle panel of Table 1 illustrates this point. The particular draws from the Laplace distribution have been added to or subtracted from the original data. One problem is that there are now fractions of people living in this particular census block. To resolve this issue, fractional values are rounded to become integers. The second issue arises from cases in which the draw from the Laplace distribution subtracts more than the original number of people who occupy a particular cell in the table. This is especially likely to happen in cases with small counts of people, such as in census blocks. This results negative values, which are of course, not possible. To resolve this issue, these values are truncated so that they are no longer less than zero. A simplified example of this step is displayed in the bottom panel of Table 1.

The steps of integer rounding and resolving negative counts would be trivial except that the Census Bureau has committed to providing invariant (i.e. accurate) counts of people for redistricting purposes at the state level. However, when a block (or subgroup within a block) is rounded or adjusted to no longer be negative, this results in an overall change in the total population, as illustrated in Table 1 below. Thus, an equal number of people must be subtracted from another block (or subgroup within a block) to maintain the correct population numbers across the various states. Furthermore, a similar problem must be resolved with geographies that are nested within other geographies (i.e. the number of

people in all blocks in tract X should add up to the total population reported in tract X, even after the statistical noise has been added). Because noise is injected independently into each histogram, the totals are inconsistent with each other, both within and across geographic levels. This is an incredibly complex problem to solve since the number of ways in which blocks (or subgroups of blocks) could be adjusted to maintain the correct population at the state level while also making the data internally consistent across geographies is enormous.

Table 1: **A simplified example of differential privacy and post-processing:**

	Race			
	White	Black	Other	
Male	5	2	0	
Female	3	4	3	
			Total Population:	17

After adding statsitical noise via sampling from probability distribution:

	Race			
	White	Black	Other	
Male	5+3=8	2+0=2	0-5=-5	
Female	3+2.25=5.25	1+.5=1.5	3-1= 2	
			Total Population:	13.75

After post-processing to remove fractions and negative counts:

	Race			
	White	Black	Other	
Male	8	2	0	
Female	5	2	2	
			Total Population:	19

To accomplish this objective, the Census Bureau uses what is known as a "least squares optimization," which is another common statistical method. In this case, the problem to optimize over is incredibly large and unusually difficult given the size of the dataset as well as the numerous constraints imposed as a part of the optimization problem. Abowd et.

al (2020) provide a technical description of this least squares optimization problem.[12] The use of optimization via the method of least squares is an extremely common application of statistical inference and is widely used across the social sciences, natural sciences, and many other disciplines.[13]

In a presentation describing the differential privacy process and the 2020 Census, Michael Hawes, a Senior advisor for data access and privacy, described this post-processing optimization procedure as "statistical inference creating non-negative integer counts from the noisy measurements."[14] In other words, the procedure is a particular use of statistical inference to locate the optimal solution (i.e. closest to the DP injected counts) to the problem of cell counts that need to be non-negative integers whose sum totals up to the accurate count of people at the state level, among other constraints. By their own admission, the Census Bureau is using statistical inferential methods to implement the post-processing procedure.

3 Enumeration and DP

In their report, the Department of Commerce appears to draw a hard distinction between the "enumeration" period of the census and the "disclosure avoidance" methods that are applied to the census data after they are enumerated. This distinction is, however, a matter of semantics and not one of substance. This is because, for all intents and purposes, users of the census data, including state legislatures and other redistricting bodies, will only have access to the noise-injected data and not the original, accurate, enumerated data. The Census Bureau argues that because they provide the accurate state-level data for the purposes of redistricting, that the differential privacy and post-processing procedure are not

[12] https://www2.census.gov/adrm/CED/Papers/CY20/202008AbowdBenedettoGarfinkelDahletal-The%20modernization%20of.pdf

[13] These are only a small sample of statistics textbooks that discuss statistical inference and least squares methods in detail: Silvey, S.D.. Statistical Inference. Japan: Taylor & Francis, 1975. Berger, Roger L.., Casella, George. Statistical Inference. United States: Cengage Learning, 2021. Stock, James H.., Watson, Mark W.. Introduction to Econometrics. United States: Pearson Education, 2015. Freedman, David A.. Statistical Models: Theory and Practice. United States: Cambridge University Press, 2009. Greene, William H.. Econometric analysis. United Kingdom: Pearson/Prentice Hall, 2008.

[14] https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf, slide 24

a part of the enumeration procedure. However, this is only partially true. In addition to the state level totals, redistricting bodies also need accurate counts at the sub-state level. While the state-level data are used to allocate seats for the US House of Representatives, the districts themselves require more fine-grained data to ensure equal population across districts within states as well as other racial and geographic-based measures to ensure the creation of certain majority-minority districts or the protection of other communities of interest, as required by law.

The adding and subtracting of counts that occurs during the differential privacy and post-processing stages of the enumeration process will impact the overall counts of people that are used to partition states into their various legislative districts. Importantly, in 2010 the Census Bureau provided accurate enumerations at both the state level and census block level, which allowed for not only an accurate allocation of legislative seats across the states, but also the accurate creation of legislative districts from the combination of census blocks within states. This will not be the case if the Census Bureau goes forward with their plan for differential privacy and post-processing of the data.

I, Michael Barber, am being compensated for my time in preparing this report at an hourly rate of $400/hour. My compensation is in no way contingent on the conclusions reached as a result of my analysis.

Michael Barber

March 25, 2021