Survey Operations and Differentially Private Noise Injection in the Decennial Census[*]

Quentin Brummet

NORC at the University of Chicago

December 14, 2020

Preliminary Draft – Please do not cite without permission

Abstract

We consider the effect of the U.S. Census Bureau's announcement that it will use differentially private (DP) noise injection in the 2020 Census on survey operations. The decennial census forms the backbone of many statistical operations that provide data for downstream use cases, making it important to understand whether the use of DP will lead to lower data utility. We first discuss at a high level the various pieces of survey operations that rely on census data, and then perform comparisons between released 2010 Census data and a series of DP demonstration files. We focus especially on the effect of DP data on sample design, showing that for large-scale survey operations there are likely to be very few if any effects on sample efficiency or coverage. However, surveys targeting smaller populations may see decreases in survey coverage, meaning survey collectors may be faced with a decision of whether to accept the worse coverage or increase costs by pursuing alternative methodologies or fielding strategies. These results are consistent across versions of DP demonstration data products, with modest improvements in newer versions of the DP demonstration data.

[*] I am grateful for support from the Alfred P. Sloan Foundation (award G-2019-12589). Thanks also go Claire Brockway for research assistance and to Rupa Datta, Lee Fiorio, Dave Van Riper, and Kirk Wolter for helpful discussions. Any errors and omissions are my own.

1 Introduction

Sample surveys provide essential information for a variety of stakeholders. Across domains such as health, education, criminal justice, or employment, the information provided by sample surveys enables informed decision making and allows for the administration of a variety of federal, state, and local programs. These sample surveys are part of a larger statistical ecosystem, and almost all are reliant on other data sources in order to function properly. Arguably, the most important of these data sources in the United States comes from the decennial census, which provides the backbone of many survey operations such as frame development, sample design, and weight development. Because of the importance of census data, maintaining the quality of decennial census data is imperative for these surveys to continue providing high-quality information to their users.

In this paper, we consider the effect of the U.S. Census Bureau's announcement that it will use differentially private (DP) noise injection techniques to protect respondent confidentiality in the 2020 census. This announcement comes in response to concerns over increasing amount of publicly available data and growing computational power that make it easier for "attackers" to make improper use of statistical data, such as identifying respondent information from aggregate anonymized statistics. As currently proposed, the use of DP will likely provide significant improvements in the ability to protect the privacy of Census respondents, but may come at the cost of lower data utility given that additional noise will be added by the DP mechanism. Because of this lessened data utility, many observers have suggested that the switch to DP is too large a departure from traditional disclosure avoidance methods and that any potential gains in privacy protection are not worth the cost of lower utility in Census data products**.**

As a specific example of how DP in the census may impact data utility, we provide evidence of how survey operations may be affected using a series of demonstration products released by the US Census Bureau. We first document descriptive differences in data between DP and the originally released 2010 Census data, showing results that vary slightly across data release but in general point to noisier data for smaller geographies and geographies off the traditional geographic “spine” (i.e., geographies other than block, block group, tract, county, state, and nation).

We then discuss the various pieces of survey operations that rely on census data. Survey frames may use Census data, and we show that additional noise in census data could lead to additional operational expenditures in order to construct high-quality survey frames. In addition, a variety of other pieces of the survey life cycle use census data, including weighting, benchmarking, and sample prioritization. For many of these, it is difficult to construct estimates of how DP data will affect the specific use case, but they nonetheless may represent important effects of the change in data quality.

We then examine in depth the effect of DP data on sample design. Many important government and private sector surveys, including the General Social Survey, Current Population Survey, and National Health Interview Survey, use census data for sampling and weighting (Smith et al. 2019, Wolter et al. 2015, Parsons et al. 2014). Therefore, understanding the effects of DP census data on sample design is important for all downstream users of these other surveys. Our results show that for large-scale survey operations there are likely to be very few if any effects on sample efficiency or coverage. However, for smaller operations our results show potential decreases in survey coverage. Therefore, survey collectors may be faced with a decision of whether to suffer worse coverage or increase costs by pursuing additional methodologies or fielding strategies. These results are consistent across data releases, the July 2020 version of the DP demonstration data appearing modestly better than other versions.

There are a few caveats to note regarding this analysis. Most importantly, the extent to which data utility is hurt by DP will vary substantially by the implementation of the DP system, so it is important to evaluate the effects of DP separately for each iteration. In addition, given that the DP noise injection process is transparent and well documented, researchers may be able to adjust their analyses to account for the additional error in the data. However, doing this measurement error correction is made more difficult by the post-processing steps that are included in the Census’s DP implementation, and the release of DP data without post-processing would lead to improved inference (Seeman et al. 2020). Finally, in the context of survey operations specifically there may be additional steps that a survey collector may be able to take in order to mitigate potential issues caused by additional DP error (e.g., moving to a new frame or drawing sample using a procedure that accounted for measurement error). Nonetheless, doing so would involve a transition period and would naturally come with an inherent cost.

The rest of the paper is structured as follows. Section 2 provides a high-level overview of DP and the Census Bureau's DP implementation, and Section 3 discusses the data used in the analysis as well as a high-level discussion of how the DP demonstration data relates to previously released 2010 Census data. Section 4 then discusses how DP data might affect survey operations, and Section 5 presents an analysis of how DP census data might be expected to affect sample design. Section 6 then concludes and discusses implications of this analysis.

2 Differential Privacy

The adoption of DP represents a significant departure from traditional Census disclosure limitation techniques, and provides privacy guarantees that are mathematically provable without making assumptions related to the computing power, knowledge, or sophistication of potential attackers (Garfinkel et al. 2018, McKenna 2018). We note that DP itself is a definition of privacy, and there are a myriad of techniques that adhere to the definition.[1] In a typical example, DP adds noise to the data, where the amount of noise is calibrated in an attempt to optimally balance the negative effects of the noise on data utility with privacy as defined by DP.

2.1 Differential Privacy as a Definition of Privacy

Consider the case of an attacker who is attempting to use census data in an improper way. This improper use could take multiple forms, including attempts to reidentify a respondent and obtain their specific data

[1] See Dwork and Roth (2014) or Vadhan (2018) for overviews of DP.

or more general purposes such as using census data to predict a respondent's race. As described in Dwork and Roth (2014), the basic intuition behind DP is that an attacker's beliefs about an individual should not improve much based on whether or not a specific individual's information is included in the DP data set. Taking the example of predicting a respondent's race, there may be a number of pieces of information available to an attacker that allow them to accurately predict the respondent's race, regardless of whether census data is released. DP attempts to ensure is that the attacker's prediction accuracy is not improved *much* based on whether or not the respondent responds to the census.

More formally, the output of the DP mechanism is almost equally likely to be generated from a neighboring data set as the data set used to create the statistic. To define this formally following Dwork and Roth (2014), let M be a mechanism for masking a statistic that takes as an input a data base $x \in \mathbb{N}^{[X]}$ and produces a masked statistic $M(x)$. The mechanism is ε-DP if for all $S \subset Range(M)$ and for all $x, y \in \mathbb{N}^{[X]}$ where $\|x - y\|_1 \leq 1$:

$$P(M(x) \in S) \leq e^{\varepsilon} P(M(y) \in S)$$

Here, ϵ is the DP measure of privacy, and the value of ϵ controls the amount of noise added to the final data. As mentioned previously, this value is a policy decision and must be set in an attempt to balance privacy and data utility (Abowd and Shmutte 2018). Depending on the level of ϵ, the same DP algorithm can produce results that are almost entirely noise or results that are essentially identical to those that would be obtained on the original data.[2]

The intuition is that ϵ represents a bound for how much influence any individual's data can have on the data. When ϵ is small, this provides a strong privacy guarantee that does not rely on having to specify how attackers may attempt to improperly use the data. Nonetheless we note some caveats regarding interpretation of ϵ. First, McClure and Reiter (2012) show that the value of ϵ may or may not correspond

[2] The definition above takes data as fixed, with the only randomness being introduced by the privacy protection mechanism. This is not necessary for all definitions of DP, and Vadhan (2017) or Kifer and Machanavahhjala (2014) provide further discussion and examples.

to actual reidentification risk. In addition, interpretation of privacy protection may be difficult for relatively large values of ϵ. For example, if $\epsilon = 0.10$, this intuitively represents a roughly 10% increase in the probability of a bad event happening if a research subject chooses to share their data (Wood et al. 2018). However, if $\epsilon = 4$, this relative increase would be e^4, or a 50-fold increase in the probability of a bad event. If $\epsilon = 8$, the same increase is $e^8 \approx 3{,}000$.

2.2 Differentially Private Noise Injection in the 2020 Census

Van Riper et al. (2020) provide a detailed and accessible introduction to the workings of the DAS, but we summarize important details here.[3] At a high level, the algorithm works by dividing the data into a "histogram" of various bins. As an example, a bin might include all female residents of a specific age group and racial/ethnic group in a given block. After this, a random draw is added to the count of observations in each cell of data and statistics of interest can then be calculated off of these noise-injected counts. The DAS makes use of the insights from the matrix mechanism described in Li et al. (2015) in order to make this as efficient as possible by taking into account the overlapping structure of the tables released as part of the decennial census.

The DAS applies noise hierarchically, ensuring that the DP data satisfy standard hierarchical relationships in decennial census data. In other words, noise is added at different geographic levels so that statistics for more aggregated geographic areas contain less noise than statistics from less aggregated geographic areas. This will be important for the results shown below: any use of decennial census data for larger geographies will be much less affected by DP than uses that rely on precise information for smaller geographic levels.

[3] Reiter (2019) and boyd (2019) also provide descriptions of the potential uses of differential privacy in federal statistical agencies. Abowd et al. (2019), Leclerc (2019), and Ashmead et al. (2019) provide technical details of the DAS.

In addition to this noise injection process, the DAS includes a post-processing step in order to deal with other complications that arise in constructing DP data for the decennial census.   This includes constraints on the data set so that there can be no negative population counts for any geographies, as well as constraints to respect logical relationships (for example, the number of individuals within a county must equal the sum of individuals in each individual tract within the county). In addition, post-processing ensure that DP data follow "invariants", a group of statistics that are released accurately without any perturbation. The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) has announced that for the 2020 Census, invariants will include state-level population counts as well as block-level housing unit totals and group quarters totals by group quarters type. Invariants complicate the construction of the DP algorithm, because it limits the options that the DAS has for reallocating population and housing in order to achieve data that meets all constraints. Raw population counts at low levels of geography will have noise injected and not necessarily reflect true population totals.

Ruggles et al. (2019) argue that using DP as a definition of privacy is too strong for the case of the decennial census, and not consistent with the statutory obligations of the Census Bureau. Other observers have noted that the lower data utility may also make decennial census data difficult to use for topics such as measuring health disparities or COVID-19 mortality rates (Santos-Lozada et al. 2020, Hauer and Santos-Lozada 2020).

3 Data

Our primary data used in the analysis are assembled by the Minnesota Population Center (Manson et al. 2020).[4] These data files cover varying geographic levels and contain original data from the 2010 Census (designated as "SF1") and data from the Census DP demonstration products (designated "DP). We consider three sets of demonstration data released by the Census Bureau:

[4] See https://www.nhgis.org/differentially-private-2010-census-data for the full data.

1. “Demonstration Data”: these data were released by Census in fall of 2019 and discussed at length in the Workshop on 2020 Census Data Products at the National Academies of Sciences Engineering and Medicine.[5]
2. “May 2020”: This demonstration product was released by Census incorporating feedback from the workshop, with improvements especially focused on the post-processing part of the algorithm.
3. “September 2020”: This data set was more restricted in content in that it only included the information needed for redistricting (specifically, tables in the PL94-171 redistricting file).

In each successive iteration, the Census Bureau has included additional enhancements to the DP algorithm. Importantly, following feedback to the initial demonstration data, the May 2020 products utilized a multi-step post-processing routine that was no longer solely focused on the traditional Census geographic hierarchy of block, block group, tract, tract group (a new construct for the 2020 Census), county, state, and nation. This has the advantage of making the data less error-prone for “off-spine” geographies such as school districts, cities, or census designated places. The September 2020 products included additional improvements for post-processing, especially post-processing related to the American Indian and Alaskan Native (AIAN) population.

Note that the SF1 data contains its own privacy protections, most importantly an unknown amount of swapping. Therefore, the comparisons below are inherently comparing two data sources with some perturbation added with neither representing the “ground truth”. While it is likely the effects of swapping and other statistical disclosure control methods are small relative to the effect of DP, the Census Bureau is unable to release the exact extent of swapping for fear of compromising the integrity of the privacy protection. In what follows, we will generally compare results on the two data sets without noting the specifics of the SF1 disclosure methodology, but all results should be interpreted with the understanding that SF1 includes at least some small amount of error due to disclosure avoidance processes in 2010.

[5] See https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518 for full presentations from this workshop.

To give a sense of how DP and SF1 data relate, we first consider the distribution of percent differences across these files. Table 1 shows the distribution of percent difference in county-level population counts for major racial and ethnic groups across the DP demonstration products. On the whole, the results show that for county-level counts there is relatively little difference across DP and SF1 data in terms of total population counts. For example, at the 5th percentile the DP data show 2.21 percent less population than the SF1 data. This improves over the releases of the DP files and is 0.08 percent in the most recent September 2020 file. For smaller racial and ethnic groups the percent differences can be much larger, though this largely reflects counties with very small populations. Nonetheless, we see qualitatively similar patterns across the demonstration files for these subgroups.

Note that the distribution is shifted to the left in the sense that the median percent difference for many counts is less than zero. This is a result of the post-processing, where no counties are allowed to experience negative counts. The intuition is that after the initial addition of noise there may be negative counts. However, because no geographic area can have a negative population, the DAS will force the count to be zero. Because total state-level population must be invariant this leads the DAS algorithm to subtract population from other geographic areas. This leads to a pattern where there tend to be more geographic areas with lower populations.

As an example of a smaller geographic area, Table 2 displays percentage differences at the school district level. Here, we can see that total counts are more variable. For example, at the 5th percentile total population in the DP Demonstration Data is 3.7% lower than SF1 data and at the 95th percentile is 6.77% larger than SF1 data. This is a much wider range than for counties (or tracts that are shown below in Table 3), and reflects the fact that especially for the Demonstration Data the DAS algorithm is focused on the on-spine geographies. The more recent versions of demonstration data incorporate other geographic entities into the post-processing step and show improvements in terms of school-district total population. In the most recent September 2020 data the comparable numbers are 1.94% and 2.37% respectively.

Racial and ethnic subgroups show a similar pattern, with percent differences that are larger than county-level differences in Table 1.

Table 3 then considers percent differences in Census tracts, which are even smaller in population (averaging about 4,000 residents) and form the basis for many survey operations considered below. The percent differences in total population are actually less variable than those for school districts, likely reflecting the additional accuracy due to being on the spine of census geographic hierarchies. Note that for smaller racial and ethnic groups such as AIAN individuals, there are instances where the percentage difference is -100% (i.e., the tract loses all AIAN population). These tend to be in areas with very small AIAN populations so are not necessarily of concern in and of themselves, but if they occur across enough Census tracts could lead to issues with coverage in survey operations. We discuss this issue in more detail in the coming sections.

4 Survey Operations and Census Data

Census data is used across all phases of the survey life cycle. These include uses in areas such as frame construction, sample design, fielding, and weighting/benchmarking. Below, we discuss at a high level the uses of Census data in surveys and then provide a more in-depth discussion of how census data may be used in sample design related to oversampling or targeting of rare populations.

4.1 DP Census Data in Survey Operations

At the very start of many surveys, census data are often invaluable as a tool in constructing sampling frames. Many survey collectors perform their own listing operations in order to create frames, and rely on decennial Census data in order to determine when these operations are needed. For example, the 2010 National Frame at NORC uses data from the United States Postal Service (USPS) as an initial list of addresses, and then conducted an in-person listing operation some geographic areas where the census count of occupied housing units differed substantially from the count of DSF address. These types of in-person operations are expensive, and rely on census data in order to target resources where they are most needed. While USPS data has improved over time, more generally there will always be potential new sources of data that could be incorporated into a survey operation and will need to be validated. To the extent that DP adds additional uncertainty into the released census data, this will further degrade the ability to do these validation exercise.

As an example of how the change to DP data might impact frame construction, Figure 1 plots the tract-level ratio of units in the DSF to units in the 2010 Census using either the DP Demonstration Data (y axis) or SF1 Census Data (x axis). We see that the dots in blue primarily follow the 45-degree line, but that there is a cluster of dots in the middle for which there is some significant variation around the line. On the whole, these results indicate that the use of census counts for this validity exercise will be less reliable and that survey collectors may wish to investigate other approaches**.**

In addition to frame construction, census data end up being used in less structured manners throughout the fielding of a survey. For example, being able to understand which census blocks or block groups within a tract have the highest populations may help field workers prioritize where to survey first. Given the informal nature of these uses, it is difficult to understand exactly how the move to DP might affect these uses. Nonetheless, given that estimates at lower levels of geography such as block and block groups are likely to be measured with significantly more error in DP data these uses may be more difficult in the future.

At the end of the survey, census data are also used for benchmarking and weighting. These uses tend to primarily use data from very high geographies (e.g., racial/ethnic compositions for an entire state). Given the fact that state-level estimates are likely to be very accurate in the DP data, it is unlikely that the move to DP as currently implemented will have much effect on these uses.

4.2 DP Census Data and Sample Design

In many sample designs, Census data is used to target survey operations in order to contain survey costs. This can happen either by oversampling a population as part of a larger survey or if a subgroup is rare enough, screening potential sample members in order to create a sample of only individuals from the rare subgroup. In either case, the intuition is that a sample is designed with that has a larger group of individuals from areas with high concentrations of the target population. Because this relies on having accurate data to differentiate between geographic areas, DP noise injection may lead to a loss of efficiency in this process.

Much of the following discussion mirrors Brummet et al. (2020), who study a similar problem using 1940 Census data.[6] Denote the entire adult population size as N, which is divided into two sampling strata of sizes N_L and N_H, such that $N_L + N_H = N$. Stratum L is the low-density stratum consisting of tracts in which less than 30% of the individuals are from a target population, and stratum H consists of tracts with at least 30% of the target population. An equal number of screener interviews are then conducted for all sampled enumeration districts regardless of strata.

Let $f_H = n_H/N_H$ be the sampling fraction for adults in stratum H, and let $f_L = f_H/b$ be the sampling fraction stratum L, where b controls the degree to which the sample is tilted towards the high-density stratum. As b increases, the sample becomes more heavily represented by individuals in the high-density stratum. This leads to increased efficiency for the screening operation (i.e., reduced cost per completed interview for the target population), but increased sampling variance for estimators of survey parameters of interest. Given proper weighting of the survey data, no bias is introduced by the oversampling procedure.

[6] For further discussion of oversampling and stratified sampling more generally, refer to Cochran (1977), Kalton and Anderson (1986), Biemer and Lyberg (2003), or Kalton (2009).

If we wish to obtain a sample of 1000 members, then we will need to draw a sample of the following size:

$$1000\frac{N_l + N_h b}{d_L N_L + d_H N_H b}$$

where d_L is the fraction of adults in the low-density stratum in the target population and d_H is the fraction in the high-density stratum from the target population. As the density of the target subgroup decreases, more screeners will be needed in order to achieve a given sample size. For the purposes of this exercise, we assume that any area with a known 0% density of the target population is excluded from consideration. These areas could be included in the sample, but this would come with a significant increase in costs that in the context of our current examples would be prohibitive.

With DP data, the strata may be misallocated, which could lead to decreases in efficiency. In addition, because the sample will only include areas with non-zero population of the target group, DP data may lead to a lack of coverage for rare subgroups.

5 Sample Efficiency and Coverage Changes Due to Differential Privacy

The efficiency and coverage of a sample survey operation will depend on how precisely subgroup concentration as measured in DP data reflects the underlying census data. Given that in practice it is common to use census tracts as a geographic unit for sampling purposes, we first consider the relationship between concentration of racial and ethnic subgroups in DP and SF1 data. Figure 2 plots this relationship for the Demonstration Data. The x-axis shows the fraction in a particular subgroup in that tract in the SF1 data, and the y-axis shows the fraction in the same subgroup in the Demonstration Data. The results show that for Whites, who are the majority racial/ethnic group in many tracts, the DP demonstration data is more likely to show a lower concentration than the SF1 data. This is a property of the DP algorithm, where the noise tends to lead to subgroups being more evenly spread than in the SF1 data. Looking at rarer subgroups, we see that the majority of dots do track with the 45-degree line, but that especially towards the left of the graph there is more dispersion. Of particular note are tracts that lie on the x-axis itself. For these tracts, the SF1 data show that there are individuals living in the census tract, but the DP data show no individuals from that subgroup present. As will be discussed below, these tracts could pose an issue for survey operations as survey collectors would need to decide between covering these tracts in their survey (which will significantly increase costs), or accepting potential coverage losses.

Figures 3 and 4 show this same relationship for the May 2020 and September 2020 demonstration products, respectively. The results are similar across the files, with some subgroups perhaps tightening more towards the 45-degree line than others in the later demonstration products.

5.1 Sample Efficiency and Coverage in a Nationwide Survey

These graphical results suggest that clearly there is significant variation between DP and SF1 data, but it is an open question as to whether this would impact a downstream use case of using DP data for sample design. To examine this question, Table 4 shows the relative efficiency of a survey operation using DP data as opposed to SF1 data. Here, the rows indicate the target subgroup for the survey, and the columns indicate the vintage of the DP data that would be used. Each cell contains the relative increase in costs for the survey, where for example 1.01 would be interpreted as a 1% increase in expected costs. Each cell in the table is extremely close to 1, suggesting that in the context of a nationwide survey of major racial/ethnic groups that the expected effects of the move to DP on survey costs would be minimal.

Even though efficiency is unlikely to be effected, there could effects of DP on the coverage of the survey due to geographic areas showing 0 sample size for the target population. To investigate this, Table 5 shows the expected coverage for surveys using DP data relative to SF1 data. The table is laid out similarly to Table 4, with cells representing the coverage of the survey. Coverage here is defined as the fraction of the target population that would be included in the sample frame for a survey using DP data for sample design.[7] The results show that for most racial/ethnic subgroups, the coverage is very close to 1, suggesting that coverage is unlikely to suffer in nationwide surveys for these populations (e.g., a nationally-representative survey that wished to oversample African-American households). However, for AIAN individuals, the coverage in the Demonstration Data is 0.963, indicating that roughly 4% of the AIAN population would be dropped from a survey frame using this data. This number increases in the May 2020 data, but then decreases in the September 2020 data. This highlights that the effects of DP are dependent on the specific DP method that is applied to the data, and it is still a little uncertain how the eventual DP data utilized in the 2020 Census may relate to these DP data products.

5.2 Sample Efficiency and Coverage in a Statewide Survey

The larger the geographic scale of this analysis, the less likely it is that DP data will have an impact. Therefore, in order to obtain an understanding of how smaller-scale surveys may be impacted we examine hypothetical statewide surveys using a similar methodology. Given the very small results on survey efficiency shown above, we focus specifically on coverage. Table 6 shows the expected coverage from a statewide survey operation of specific demographic subgroups, where these results are directly comparable the Demonstration Data Results in Table 5. Given the results in Table 5 showed that the three data sets provide qualitatively similar results, we limit this analysis specifically to the Demonstration Data.

The results show that there is a wide variation across states and racial/ethnic groups, but that many have coverage very close to 100%. Examining the table, it appears that smaller states with rarer populations may be more prone to lower coverage. In order to examine this in greater detail, Figure 5 plots how the

[7] We again note that because the SF1 data include their own privacy protections, this estimate is not exact. Nonetheless, it is likely that the extent of swapping the SF1 data is sufficiently small to not materially affect the results. In addition, these results are similar (or if anything more optimistic) than results in Brummet et al (2020), who use 1940 DP Census data to investigate a similar question with access to the underlying data with no privacy protections added.

expected coverage of hypothetical surveys relates to the concentration of the target subgroup in that state. The results show that coverage is high for the majority of states and subgroups, but that for settings with rare populations DP data lead to significant drops in coverage. This indicates that in settings with very rare populations that DP census data may not be particularly useful for sample design. We note that in these settings, it is quite possible that the subgroup of interest was sufficiently rare that even with true data it would be difficult to do a sufficient job of oversampling or targeting. Nonetheless, additional noise from DP will make the data less usable.

6 Conclusion

In this paper, we consider the effects of the change to using DP to protect decennial census data on survey operations. We show that in the context of nationwide surveys, the effects on both efficiency of the survey as well as coverage are projected to be minimal. Nonetheless, for rarer populations in specific geographic areas the DP data could lead to drops in coverage. Moreover, because DP data are less accurate for "off-spine" geographies such as school districts, survey operations or other analyses examining these geographies may be more impacted.

We conclude by noting that the implications of this change in disclosure protection for analysts and survey organizations. The implementation of DP leads to additional error that could be thought of as measurement error for analytic purposes. But unlike many forms of measurement error, DP-induced measurement error can be traced back to an exact form. This means that data users can adjust their analyses to be aware of the additional noise. While the adjustments will never be able to correct specific estimates at a small area level, they can lead to correct statistical inference about important questions of interest (Seeman et al. 2020). We emphasize that these procedures will work best if the DP data are released without post-processing, as the DAS post-processing is done in a manner that is not amenable to being represented in a closed-form manner. Therefore, access to data without post-processing would enable analysis to significantly improve their inference and analyses that use decennial census data.

References

Abowd, John, Robert Ashmead, Simson Garfinkel, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, and William Sexton. 2019. "Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge." Working Paper. Retrieved from https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf.

Abowd, John M., and Ian M. Schmutte.2019. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review* 109 (1): 171–202.

Ashmead, Robert, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. 2019. "Effective Privacy after Adjusting for Invariants with Applications to the 2020 Census." Working Paper. Retrieved from https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints__With_applications_to_the_2020_Census.pdf.

Biemer, Paul P. and Lyberg, Lars E. 2003. *Introduction to Survey Quality* (Vol. 335). John Wiley & Sons.

boyd, danah. 2019. "Differential Privacy in the 2020 Census and the Implications for Available Data Products." Working Paper. Retrieved from https://arxiv.org/abs/1907.03639.

Brummet, Quentin, Edward Mulrow, and Kirk Wolter. 2020. "The Effect of Differentially Private Noise Injection on Sampling Efficiency and Funding Allocations: Evidence from the 1940 Census." Working Paper.

Cochran, William G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Dwork, Cynthia and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science*, 9 (3-4): 211-407.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." WPES'18 Proceedings of the 2018 Workshop on Privacy in the Electronic Society, pp. 133-137. "

Hauer, Mathew, and Alexis R. Santos-Lozada. 2020. "Differential Privacy in the 2020 Census Will Distort COVID-19 Rates." SocArXiv. October 18. doi:10.31235/osf.io/mvh5b.

Kalton, Graham and Dallas W. Anderson. 1986. Sampling Rare Populations. *Journal of the Royal Statistical Society: Series A (General*), 149(1), pp.65-82.

Kalton, Graham. 2009. Methods for Oversampling Rare Populations in Sample Surveys. *Survey Methodology*, 35(2), pp.125-141.

Kifer, Daniel and Ashwin Machanavajjhala. 2014. Pufferfish: A Framework for Mathematical Privacy Definitions. *ACM Transactions on Database Systems (TODS)*, *39*(1), pp.1-36.

Leclerc, Philip. 2019. "Guide to the Census 2018 End-to-End Test Disclosure Avoidance Algorithm and Implementation." Working Paper. Retrieved from https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0938_2018_E2E_Test_Algorithm_Description.pdf.

Li, Chao, Gerome Miklau, Micahel Hay, Andrew McGregor, and Vibhor Rastogi. 2015. "The Matrix Mechanism: Optimizing Linear Counting Queries under Differential Privacy." *The VLDB Journal,* 24 (6): 757—781.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 15.0 [dataset]. Minneapolis, MN: IPUMS. 2020. http://doi.org/10.18128/D050.V15.0.

McClure, David and Jerome P. Reiter, 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy,* 5(3), pp.535-552.

McKenna, Laura. 2018. "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing." Center for Economic Studies Working Paper 18-47, U.S. Census Bureau.

Parsons, Van L, Chris Moriarity, Kimball Jonas, et al. 2014. Design and Estimation for the National Health Interview Survey, 2006–2015. National Center for Health Statistics. Vital Health Stat 2(165).

Reiter, Jerome P. 2019. "Differential Privacy and Federal Data Releases." *Annual Review of Statistics and its Application* 6 (85): 85—101.

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research." *American Economic Review, Papers and Proceedings*, 109: 1-7.

Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Verdery. 2020. "How Differential Privacy Will Affect Our Understanding of Health Disparities in the United States." *Proceedings of the National Academies of Sciences,* 117 (24): 13405-13412.

Seeman, Jeremy, Aleksandra Slavkovic, and Matthew Reimherr. 2020. "September. Private Posterior Inference Consistent with Public Information: A Case Study in Small Area Estimation from Synthetic Census Data." In *International Conference on Privacy in Statistical Databases* (pp. 323-336). Springer, Cham.

Smith, Tom W., Michael Davern, Jeremy Freese, and Stephen L. Morgan. 2019. "General Social Surveys, 1972-2018: Cumulative Codebook, Appendix A." Chicago: NORC, 2019. National Data Program for the Social Sciences Series, no. 25.

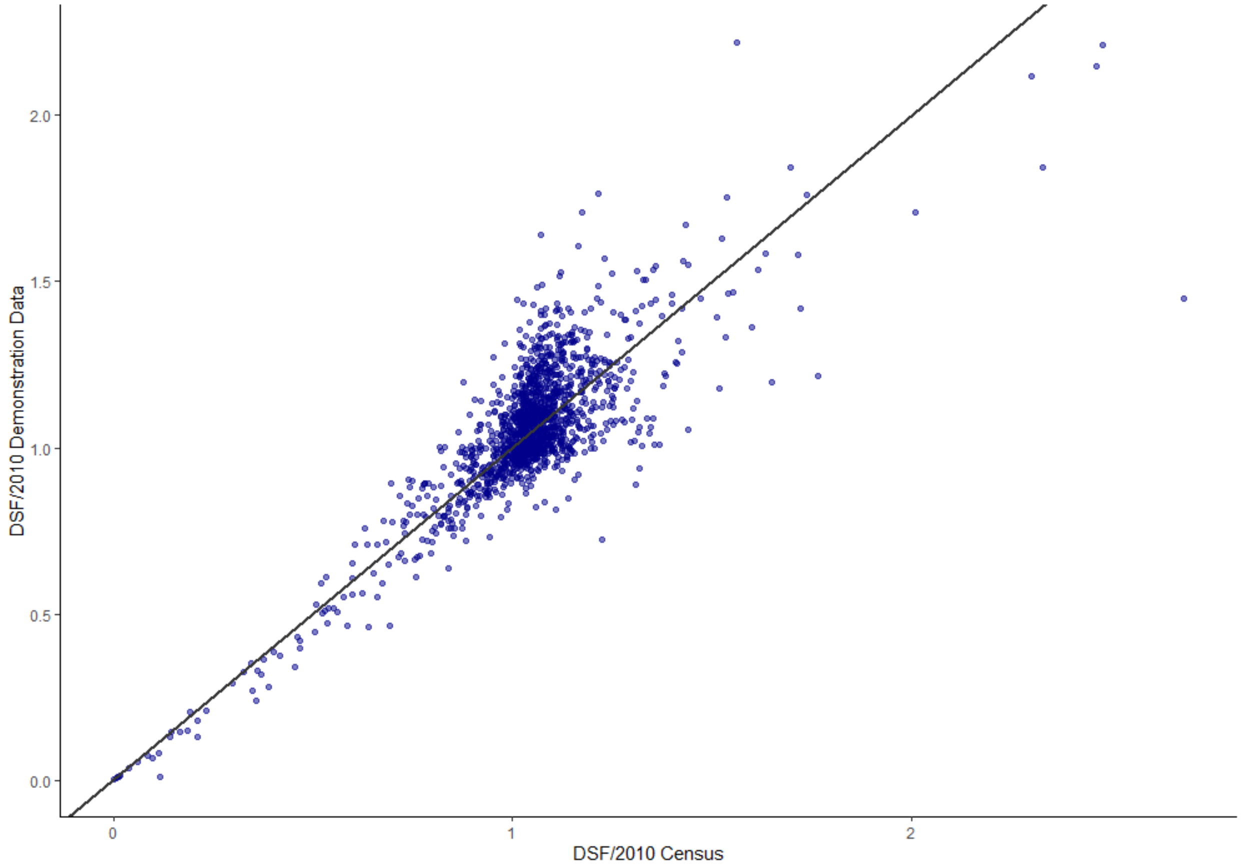
Vadhan, Salil, 2017. The Complexity of Differential Privacy. In *Tutorials on the Foundations of Cryptography* (pp. 347-450). Springer.

Van Riper, David, Tracy Kugler, Jonathan Schroeder, and Steven Ruggles. 2020. "Differential Privacy and Racial Residential Segregation." Paper presented at the Association for Public Policy Analysis and Management Annual Meetings.

Wolter, Kirk M., Anne E. Polivka, and Antoinette Lubich. 2015. "Evolution of the Current Population Survey," Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Inc., New York.

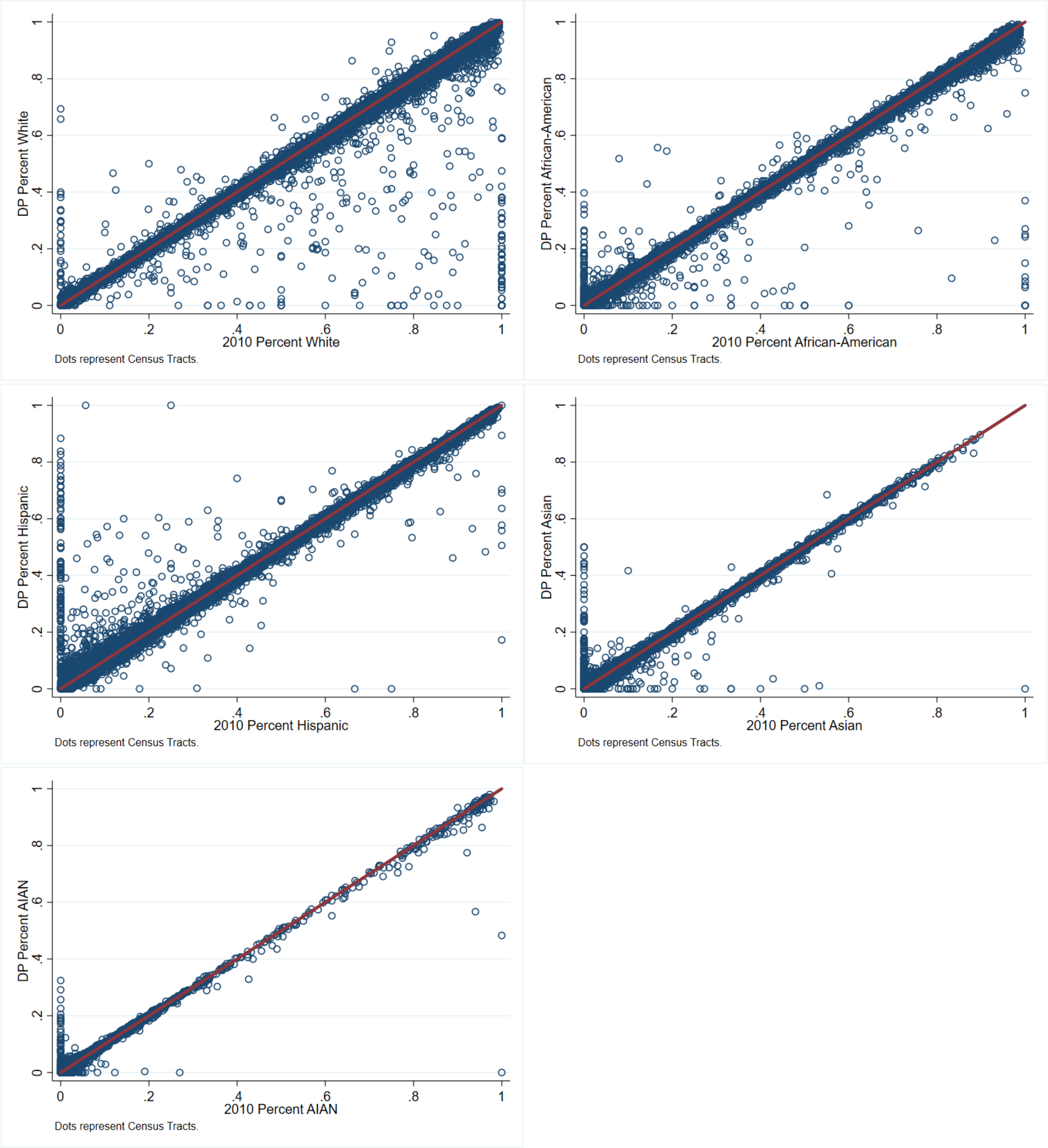
Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. 2018. Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law* 21 (1): 209.

Figure 1: Ratio of Delivery Sequence File Units to Census Occupied Units, Demonstration Data



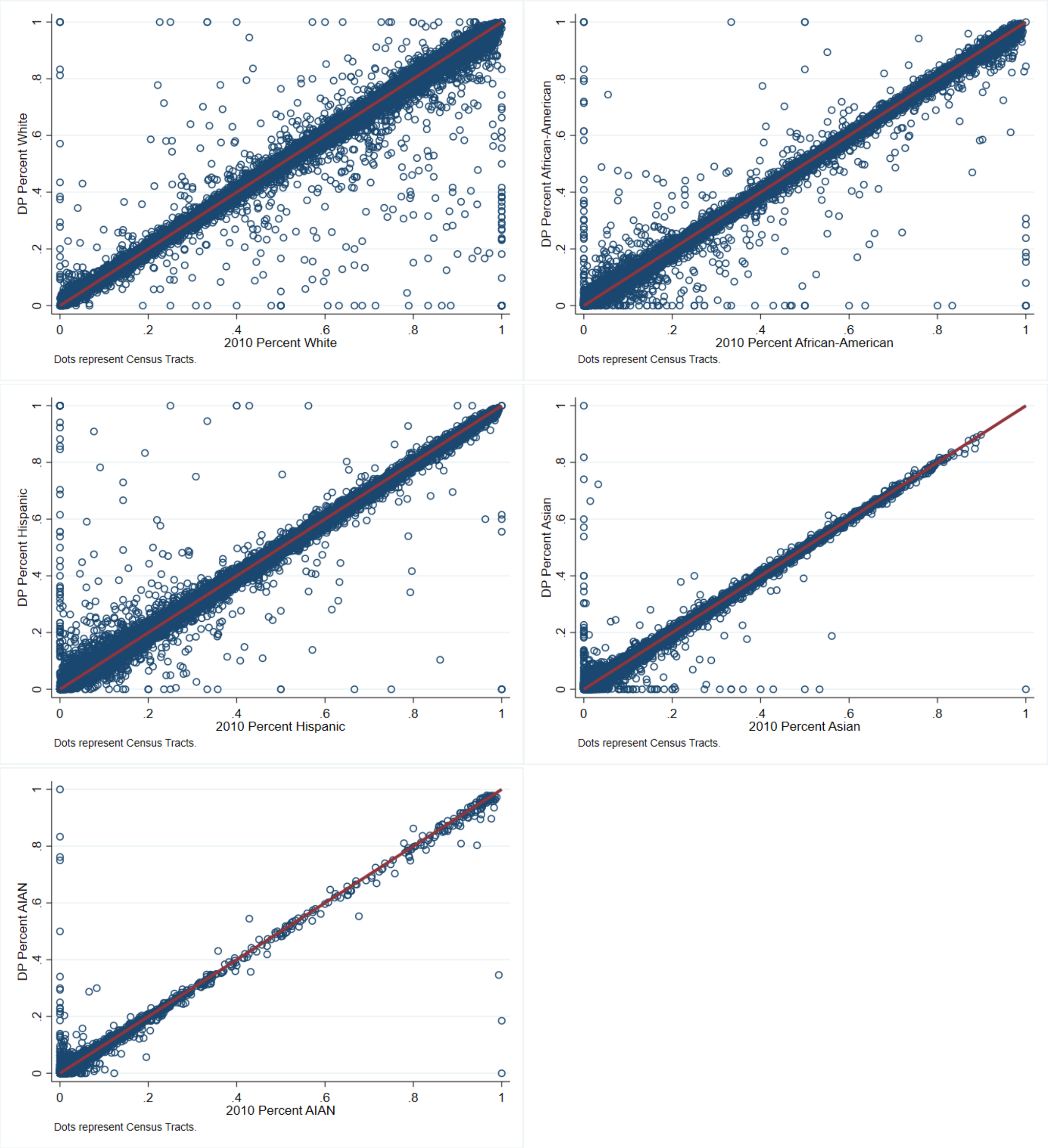
Source: USPS Delivery Sequence File and 2010 SF1 and Demonstration Data. Each dot represents a Census tract, and axes refer to the ratio of units in the DSF to the number of occupied units in Census data.

Figure 2: Comparison of Census Tract-Level Racial and Ethnic Composition between DP and SF1 Data, Demonstration Data



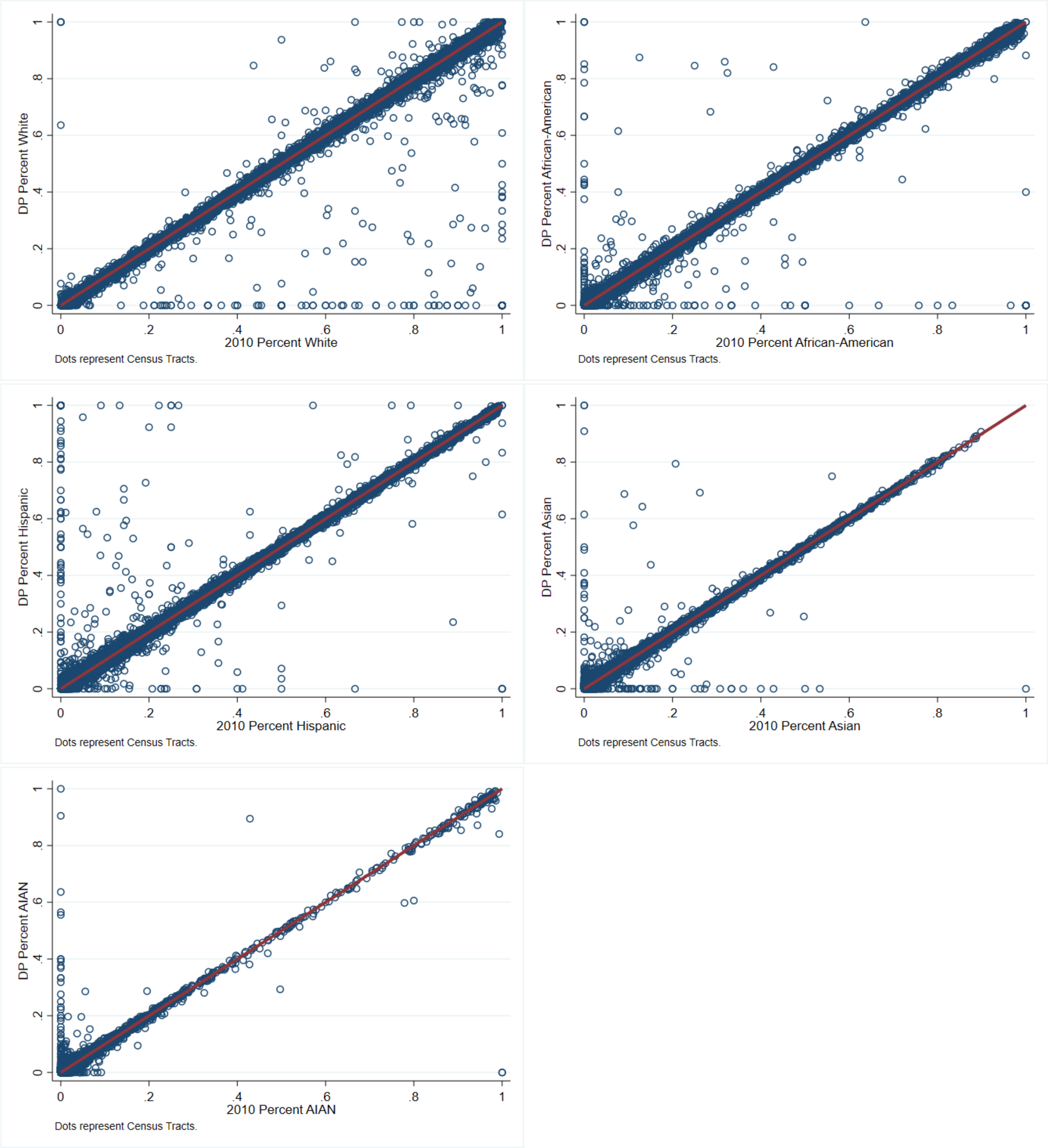
Source: 2010 SF1 and Demonstration Data. Each dot represents a Census tract, x-axes refer to the fraction of the given subgroup in the tract in the released 2010 Census SF1, while the y axes refer to the fraction of a given subgroup in the tract in the DP data.

Figure 3: Comparison of Census Tract-Level Racial and Ethnic Composition between DP and SF1 Data, May 2020 Data



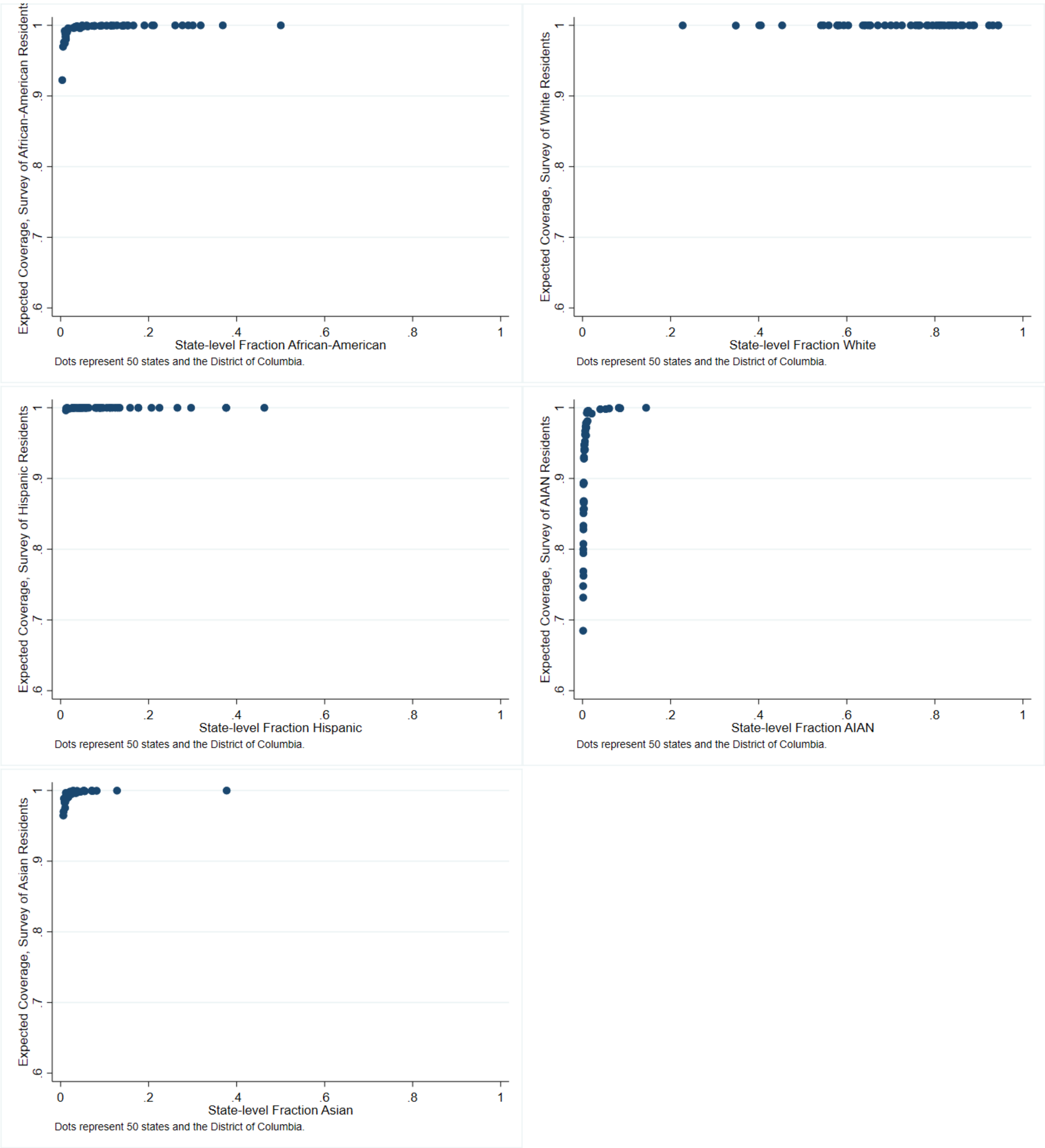
Source: 2010 SF1 and May 2020 DP data. Each dot represents a Census tract, x-axes refer to the fraction of the given subgroup in the tract in the released 2010 Census SF1, while the y axes refer to the fraction of a given subgroup in the tract in the DP data.

Figure 4: Comparison of Census Tract-Level Racial and Ethnic Composition between DP and SF1 Data, September 2020 Data



Source: 2010 SF1 and September 2020 DP data. Each dot represents a Census tract, x-axes refer to the fraction of the given subgroup in the tract in the released 2010 Census SF1, while the y axes refer to the fraction of a given subgroup in the tract in the DP data.

Figure 5: Coverage by Concentration of a Racial Group within a State



Source: 2010 SF1 and Demonstration Data. Each dot represents a state or the District of Columbia, x-axes refer to the fraction of the given subgroup in the state in the released 2010 Census SF1, while the y axes refer to the coverage in a hypothetical survey operation. Coverage is defined as the fraction of a given subgroup in the state in 2010 that would be included in the sampling frame for the survey.

Table 1: Percent Difference in Population Counts at the County Level

	Percentile of Percent Difference between DP and 2010 Census Data						
Total Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-0.1	-0.1	0.0	0.0	0.0	0.1	0.2
May 2020	-0.4	-0.2	-0.1	0.0	0.0	0.1	0.2
Demonstration Data	-2.2	-1.2	-0.5	-0.1	0.1	0.2	0.2
White Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-0.8	-0.4	-0.1	0.0	0.1	0.2	0.4
May 2020	-1.6	-0.8	-0.2	0.0	0.1	0.2	0.3
Demonstration Data	-0.8	-0.4	-0.1	0.0	0.1	0.2	0.4
Black Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-84.0	-45.5	-4.9	-0.1	0.5	11.9	43.7
May 2020	-48.3	-26.6	-3.9	-0.1	1.0	25.0	94.6
Demonstration Data	-44.2	-19.7	-2.7	-0.1	1.0	25.8	92.6
Asian Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-93.8	-73.3	-17.7	-0.2	5.0	34.3	77.8
May 2020	-66.7	-44.7	-14.1	-0.2	8.3	60.0	141.4
Demonstration Data	-64.5	-40.0	-10.6	-0.2	9.3	56.7	129.7
Hispanic Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-15.8	-8.9	-2.1	0.0	1.5	10.4	25.7
May 2020	-16.2	-7.8	-1.7	0.0	5.7	23.6	45.6
Demonstration Data	-7.0	-3.7	-1.0	0.4	9.8	38.1	78.9
AIAN Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-87.5	-57.1	-12.1	-0.1	7.1	29.3	60.0
May 2020	-58.3	-40.9	-15.3	-0.7	7.1	39.4	0.4
Demonstration Data	-47.1	-27.8	-9.1	-0.2	8.6	43.8	88.9

Source: 2010 SF1 and Demonstration Data. Numbers reflect percent differences: (DP Count – 2010 Census Count)/(2010 Census Count) * 100.

Table 2: Percent Difference in Population Counts at the School District Level

	Percentile of Percent Difference between DP and 2010 Census Data						
Total Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-1.9	-1.0	-0.2	0.0	0.2	1.2	2.4
May 2020	-3.2	-1.8	-0.4	0.0	0.6	2.6	5.0
Demonstration Data	-3.7	-2.0	-0.5	0.1	1.1	3.8	6.8
White Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-3.1	-1.6	-0.5	0.0	0.4	1.5	3.0
May 2020	-4.9	-2.7	-0.8	0.0	0.6	2.6	5.0
Demonstration Data	-4.4	-2.4	-0.7	0.0	0.6	3.0	6.1
Black Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-100.0	-23.2	-0.2	7.7	83.5	213.9
May 2020	-80.8	-50.0	-10.8	0.0	25.0	150.0	400.0
Demonstration Data	-88.9	-50.0	-8.2	0.0	22.9	127.0	300.0
Asian Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-100.0	-41.1	-0.9	16.1	108.3	260.0
May 2020	-100.0	-63.3	-21.1	0.0	35.3	172.7	400.0
Demonstration Data	-100.0	-66.7	-20.0	0.0	30.8	150.0	384.4
Hispanic Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-48.1	-27.6	-6.9	0.0	0.4	1.5	3.0
May 2020	-37.0	-21.6	-5.6	0.6	15.9	56.3	106.7
Demonstration Data	-31.3	-18.5	-4.4	1.3	18.8	64.4	116.7
AIAN Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-100.0	-23.2	-0.2	7.7	83.5	213.9
May 2020	-96.2	-66.7	-31.3	-1.1	27.3	106.3	232.0
Demonstration Data	-100.0	-66.7	-25.8	0.0	28.1	100.0	214.3

Source: 2010 SF1 and May 2020 DP Data. Numbers reflect percent differences: (DP Count – 2010 Census Count)/(2010 Census Count) * 100.

Table 3: Percent Difference in Population Counts at the Tract Level

	Percentile of Percent Difference between DP and 2010 Census Data						
Total Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-0.4	-0.3	-0.1	0.0	0.1	0.3	0.4
May 2020	-0.9	-0.6	-0.3	0.0	0.3	0.7	1.0
Demonstration Data	-1.3	-0.9	-0.5	0.0	0.5	1.3	2.0
White Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-2.5	-1.3	-0.5	0.0	0.5	1.1	1.8
May 2020	-3.6	-2.1	-0.9	0.0	0.8	1.9	3.6
Demonstration Data	-2.3	-1.4	-0.6	0.0	0.6	1.5	3.0
Black Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-51.7	-8.8	-0.1	5.2	31.8	75.0
May 2020	-54.5	-32.5	-8.1	0.0	9.3	51.9	114.3
Demonstration Data	-50.0	-25.5	-5.4	0.0	7.6	41.1	100.0
Asian Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-100.0	-27.3	-0.6	13.3	73.3	175.0
May 2020	-76.0	-52.6	-17.5	0.0	21.7	106.9	250.0
Demonstration Data	-83.3	-51.6	-15.6	0.0	17.8	100.0	240.0
Hispanic Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-36.5	-19.6	-5.2	0.0	4.8	20.7	41.7
May 2020	-33.1	-2.1	-0.9	0.0	0.8	1.9	3.6
Demonstration Data	-27.8	-16.7	-4.9	0.0	9.7	36.7	69.7
AIAN Population	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**
September 2020	-100.0	-100.0	-100.0	-13.6	41.2	164.3	318.2
May 2020	-100.0	-91.9	-50.0	-5.6	50.0	175.0	360.0
Demonstration Data	-100.0	-100.0	-50.0	-4.0	41.7	150.0	311.7

Source: 2010 SF1 and September 2020 DP Data. Numbers reflect percent differences: (DP Count – 2010 Census Count)/(2010 Census Count) * 100.

Table 4: Sampling Efficiency for Hypothetical Survey Operations of Racial/Ethnic Subgroups

	September 2020	May 2020	Demonstration Data
Non-Hispanic Black	1.000	1.001	1.001
Non-Hispanic White	1.000	1.000	1.000
Non-Hispanic AIAN	1.002	1.005	1.000
Non-Hispanic Asian	1.000	1.001	1.001
Hispanic	1.000	1.001	1.001

Source: 2010 SF1, Demonstration Data, May 2020 DP data, and September 2020 DP data. Efficiency is defined as the expected ratio of costs between a survey operation using DP data and a survey operation using SF1 data (for example, 1.001 would reflect a 0.1% expected increase in costs).

Table 5: Coverage for Hypothetical Survey Operations of Racial/Ethnic Subgroups

	September 2020	May 2020	Demonstration Data
Non-Hispanic Black	0.999	1.000	1.000
Non-Hispanic White	1.000	1.000	1.000
Non-Hispanic AIAN	0.935	0.983	0.963
Non-Hispanic Asian	0.994	0.999	0.999
Hispanic	1.000	1.000	1.000

Source: 2010 SF1, Demonstration Data, May 2020 DP data, and September 2020 DP data. Coverage is defined as the fraction of a given subgroup in the state in 2010 that would be included in the sampling frame for the survey.

Table 6: Coverage for Hypothetical Survey Operations of Racial/Ethnic Subgroups within a given state, September 2020 DP Data Release

State	Non-Hispanic Black	Non-Hispanic White	Non-Hispanic AIAN	Non-Hispanic Asian	Hispanic
Alabama	1.00	1.00	0.95	0.99	1.00
Alaska	1.00	1.00	1.00	1.00	1.00
Arizona	1.00	1.00	1.00	1.00	1.00
Arkansas	1.00	1.00	0.97	0.99	1.00
California	1.00	1.00	0.94	1.00	1.00
Colorado	1.00	1.00	0.97	1.00	1.00
Connecticut	1.00	1.00	0.81	1.00	1.00
Delaware	1.00	1.00	0.93	1.00	1.00
District of Columbia	1.00	1.00	0.76	1.00	1.00
Florida	1.00	1.00	0.89	1.00	1.00
Georgia	1.00	1.00	0.89	1.00	1.00
Hawaii	0.99	1.00	0.86	1.00	1.00
Idaho	0.97	1.00	0.99	1.00	1.00
Illinois	1.00	1.00	0.73	1.00	1.00
Indiana	1.00	1.00	0.85	0.99	1.00
Iowa	1.00	1.00	0.87	0.99	1.00
Kansas	1.00	1.00	0.98	1.00	1.00
Kentucky	1.00	1.00	0.83	0.98	1.00
Louisiana	1.00	1.00	0.96	0.99	1.00
Maine	0.98	1.00	0.96	0.98	1.00
Maryland	1.00	1.00	0.87	1.00	1.00
Massachusetts	1.00	1.00	0.80	1.00	1.00
Michigan	1.00	1.00	0.94	0.99	1.00
Minnesota	1.00	1.00	0.98	1.00	1.00
Mississippi	1.00	1.00	0.95	0.98	1.00
Missouri	1.00	1.00	0.94	0.99	1.00
Montana	0.92	1.00	1.00	0.96	1.00
Nebraska	1.00	1.00	0.96	0.99	1.00
Nevada	1.00	1.00	0.98	1.00	1.00
New Hampshire	0.99	1.00	0.83	1.00	1.00
New Jersey	1.00	1.00	0.75	1.00	1.00
New Mexico	1.00	1.00	1.00	0.99	1.00
New York	1.00	1.00	0.87	1.00	1.00
North Carolina	1.00	1.00	0.98	1.00	1.00
North Dakota	0.98	1.00	1.00	0.98	1.00
Ohio	1.00	1.00	0.77	0.99	1.00
Oklahoma	1.00	1.00	1.00	0.99	1.00
Oregon	0.99	1.00	1.00	1.00	1.00

State	Non-Hispanic Black	Non-Hispanic White	Non-Hispanic AIAN	Non-Hispanic Asian	Hispanic
Pennsylvania	1.00	1.00	0.68	1.00	1.00
Rhode Island	1.00	1.00	0.95	1.00	1.00
South Carolina	1.00	1.00	0.93	0.99	1.00
South Dakota	0.98	1.00	1.00	0.98	1.00
Tennessee	1.00	1.00	0.89	0.99	1.00
Texas	1.00	1.00	0.93	1.00	1.00
Utah	0.99	1.00	0.99	1.00	1.00
Vermont	0.97	1.00	0.86	0.99	1.00
Virginia	1.00	1.00	0.89	1.00	1.00
Washington	1.00	1.00	1.00	1.00	1.00
West Virginia	1.00	1.00	0.79	0.97	1.00
Wisconsin	1.00	1.00	0.97	1.00	1.00
Wyoming	0.98	1.00	0.99	0.99	1.00

Source: 2010 SF1 and Demonstration Data. Coverage is defined as the fraction of a given subgroup in the state in 2010 that would be included in the sampling frame for the survey.