The New Sample Design for the Annual Survey of Local Government Finances

Noah Bassel, Bac Tran
U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20233

Abstract

The Annual Survey of Local Government Finances (ALFIN) is conducted by the U.S. Census Bureau to collect data on state and local government financial activity. Every five years a full census of all local governments is taken, and approximately two years subsequent to that the Economic Statistical Methods Division (ESMD) designs and selects a new sample using the last census to create a frame. In the 2019 sample selection cycle ESMD introduced major changes to the sample design of ALFIN; new certainty criteria were added, optimal allocation was used for the first time, and a two-phase probability proportional-to-size design was replaced by a single phase stratified simple random sample. We show that while both the new and old sample design meet the initial requirements the new design gives superior performance along multiple dimensions. In particular the new design improves precision for key variables over the long term life of the sample, allows for unbiased estimation of the sampling variance, and allows for the easy incorporation of alternative estimators that are robust to influential units. Both sample designs are evaluated through a Monte Carlo simulation experiment using data from the 2012 and 2017 Census of Governments-Finance.

Key Words: Government Statistics, Sample Design, Variance Estimation, Robust Estimation

1 Introduction

The Annual Survey of Local Government Finances (ALFIN) is conducted by the U.S. Census Bureau to collect data on state and local government financial activity. Published estimates for the ALFIN are aggregated from the five local government types: counties, cities, townships, special districts, and independent school districts, in conjunction with data collected from the Annual Survey of School Finances. The Census Bureau publishes local level aggregates from the ALFIN along with corresponding state level aggregates from the Annual Survey of State Government Finances for all 50 states and the District of Columbia at three levels of aggregation: local only, state only, and state & local combined. Statistics from these two surveys are used to estimate the government component of the Gross Domestic Product, allocate some federal grant funds, and provide information to assist in public policy research. For more information on published statistics see: https://www.census.gov/programs-surveys/gov-finances.html

Every five years, in years ending in “2” and “7,” the Census Bureau conducts the Census of Governments (CoG). The finance component of the CoG, known as CoG-F, collects public financial data (expenditures, revenues, debts, and assets) for the approximately 90,000 governments in the United States. About two years after every CoG-F, Census Bureau staff redesign and select a new sample of local governments. In survey years ALFIN consists of three parts: a census of the 50 state governments, a census of the approximately 14,000 independent school districts provided by the Annual Survey of

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. Approval ID: CBDRB-FY20-ESMD001-009.

School Finances, and a probability sample of size 10,500 covering the approximately 76,000 local general purpose and special district governments. Some data on expenditures, revenues, and financial assets related to public employee retirement systems and public insurance systems are also provided via the Annual Survey of Public Pensions (ASPP). The sample is designed to provide a CV of 3% or less for total expenditures, total revenues, and long-term debts for all states and at all levels of aggregation, and a CV of 5% or less for sales tax and property tax for all states at the combined state & local level of aggregation. Additionally due to its larger sample size ALFIN is used as a benchmark for non-property taxes (sales, personal income, and corporate income) in the local component of the Quarterly Summary of State and Local Tax Revenue (QTax) survey. It is therefore desirable that ALFIN deliver estimates of high precision for these taxes at the national level even though this is not a formal requirement.

Prior samples had utilized a two-phase stratified probability proportional-to-size (πps) design with stratum sample sizes based on historical allocations. In order to improve estimate precision and to simplify estimation of variance the 2019 sample design introduced optimal allocation and the use of stratified simple random sampling (STSI). For this research, we conducted an evaluation utilizing data from the 2012 and 2017 CoG-F to compare the performance of the two sample designs. In the following sections we provide more details on the prior sample design and the new sample design, and the results of our simulation study.

2 Competing Sample Designs

In documenting the two sample designs and in subsequent sections we will make use of the following notation. Let U be the universe of size N which is studied via a sample $S \subset U$ of size $n \leq N$. Under a generalized single-phase sample design let I_k denote the inclusion indicator for unit k, such that $I_k = 1$ if $k \in S$ and $I_k = 0$ otherwise. The first order inclusion probabilities are then denoted by $\pi_k = P(I_k = 1)$, and the second order joint inclusion probability for units k and l are correspondingly denoted by $\pi_{kl} = P(I_k = 1, I_l = 1)$. Under a generalized two-phase sampling design we first draw $S_1 \subset U$ and then subsequently draw $S_2 \subset S_1$. The first and second phase inclusion indicators are defined as $I_{1k} = 1$ if $k \in S_1$ and $I_{1k} = 0$ otherwise, and $I_{2k} = 1$ if $k \in S_2$and $I_{2k} = 0$ otherwise. The first and second phase inclusion probabilities for unit k are $\pi_{1k} = P(I_{1k} = 1)$ and $\pi_{2k} = P(I_{2k} = 1|I_{1k} = 1)$, and the first and second phase joint inclusion probabilities for units k and l are denoted by $\pi_{1kl} = P(I_{1k} = 1, I_{1l} = 1)$ and $\pi_{2kl} = P(I_{2k} = 1, I_{2l} = 1|I_{1k} = 1, I_{1l} = 1)$. Finally we define $\pi_k^* = \pi_{2k}\pi_{1k} = P(I_{2k} = 1|I_{1k} = 1)P(I_{1k} = 1)$ and $\pi_{kl}^* = \pi_{2kl}\pi_{1kl} = P(I_{2k} = 1, I_{2l} = 1|I_{1k} = 1, I_{1l} = 1)\, P(I_{1k} = 1, I_{1l} = 1)$.

2.1 Prior Sample Design

In 2014 ALFIN was sampled according to the following procedure:

1) A frame was created using data from the most recent CoG-F (i.e. 2012).
2) All units on the frame were assigned a measure of size as the maximum of total expenditures and a ratio adjusted second variable depending on sampling type (total taxes for counties, total revenue for cities and towns, and long-term debt for special districts).
3) Any unit meeting one or more of the following criteria was included in the sample as an initial certainty: counties with a population of 500,000 or more, cities or towns with a population of 200,000 or more, and all local governments in Hawaii and the District of Columbia. Additional certainties were also taken on ad hoc basis in order to meet CV requirements for key variables. For example in 2014 all special

district governments in California that collected sales taxes in the 2012 Census of Governments were included as initial certainties. In 2014 258 sample units were included as initial certainty units under these criteria.

4) All the remaining units were stratified by state and type, with cities and towns consolidated into a single sampling type.
5) Initial sample sizes for primary strata were based on historic sampling rates, with adjustments made as needed in order to meet CV requirements.
6) Any unit with a measure of size large enough such that its probability of selection would be greater than or equal to 1 under a proportional-to-size sampling scheme is included with certainty. That is if x_{kh} is a unit's measure of size and n_h is the stratum sample size as determined in step 5 any unit over the cutoff $x_k \geq \sum_k x_k / n_h$ is taken with certainty. These units are referred to as second certainties in order to distinguish them from initial certainties.
7) The remaining units are sampled using systematic πps sampling, with either population (for counties and municipalities) or a numeric code corresponding to government function (for special districts) as the control variable for ordering the frame.
8) A cut-off point was calculated for the second phase of the design using the cumulative square root of the frequency method (Dalenius & Hodges, 1959), to distinguish between small and large government units in the municipal and special district strata.
9) All units in the large cutoff stratum were retained in the second phase of the sample, while units in the small cutoff strata were subsampled at a rate of 60%. Simple random sampling was used to subsample general purpose municipal governments and systematic simple random sampling was used to subsample special district governments.
10) A small number of governments (generally small special district governments that have neither expenditures nor long-term debts) have a measure of size of 0. Rather than being excluded via cutoff sampling these units are assigned to separate "non-activity" strata and sampled using simple random sampling.

The modified cutoff sampling procedure described in steps 7-9 is done in order to reduce the number of non-contributory sub-counties, and to reduce respondent burden on the smallest units while retaining estimate precision (Cheng 2012).

2.2 Proposed Sample Design

The 2019 ALFIN sample design was conducted according to the following modified sequence:

1) A frame was created using data from the most recent CoG-F (i.e. 2017).
2) Every unit on the frame was assigned a measure of size in the same manner as under the 2014 design.
3) Any unit that met one or more of the following criteria was taken as an initial certainty: counties with a population of 500,000 or more, cities or towns with a population of 200,000 or more, all local governments in Hawaii and the District of Columbia, any unit that accounted for 10% or more of the state total for a key variable, and any unit which made the largest contribution in its state to a variable subject to macro editing. The last two criteria were added on the recommendation of subject matter experts.

4) All the remaining units were stratified by state and type, with cities and towns consolidated into a single sampling type. We refer to these state and type groups as primary strata.
5) Initial sample sizes for primary strata were determined by power allocation (Bankier, 1988).
6) Units over a size cutoff were taken with certainty as under the old design. This time we used the conservative cutoff point $x_k \geq 0.8 * (\sum_k x_k / n_h)$.
7) The remaining units within each primary stratum were substratified using the equal aggregate method (Wright 1983, Särndal et al 1992), subject to the constraints that a substratum must have at least 6 sample units, and no primary stratum could be divided into more than 7 size based substrata (see Chambers and Clark 2012, Cochran 1977).
8) A simple random sample was selected in each substratum.

The choice of power allocation in step (5) for determining sample sizes of primary strata is due to the need for estimates of adequate precision across all states. Under power allocation the sample size for the *h*th primary stratum is $n_h = n \frac{(t_{xh})^\alpha cv_{yh}}{\sum_{h=1}^{H}(t_{xh})^\alpha cv_{yh}}$, where $\alpha \in [0,1]$ is a tuning constant, *n* is the total sample size, t_{xh} is the total sum of the measure of size of all units in stratum *h*, $cv_{yh} = S_{yh}/\bar{y}_{U_h}$, and y_{hk} is a variable of interest (in the case of ALFIN a unit's total revenues as of the last CoG-F were used). Setting $\alpha = 1$ and $y_{hk} = x_{hk}$ gives the Neyman allocation whereas $\alpha = 0$ gives the uniform allocation; in creating the 2019 ALFIN sample we followed the common practice of setting $\alpha = 1/3$. Power allocation therefore represents a compromise between Neyman allocation (which is optimal for national level aggregates but at the cost of high sampling variances in small strata) and uniform allocation (which oversamples the smallest strata, increasing respondent burden for smaller units and decreasing precision in larger strata). For more details see Bankier, 1988. The choice of the equal aggregate method for determining substratum boundaries and allocations in (7) is due to the findings of Wright and others that a stratified design constructed in this manner will experience only a trivial loss of efficiency compared to a πps design (see Särndal et al 1992).

In practice the two designs share many similarities. The size variable is defined in the same way under both designs. Additionally as we would expect in most economic surveys there are a large number of certainty cases due to the fact that a majority of most published totals are accounted for by a small number of very large units. In our simulation study the old design had 5,581 certainties while the new design had 5,835 certainties in total with a great degree of overlap between the two samples' lists of certainty cases. While the old design had only 258 initial certainties and the new design had over 2,600 initial certainties due to the newly added certainty criteria, 70% of the new certainties would have been taken as either initial or second certainties under the old sample design. However the subtle differences between systematic πps sampling and stratified simple random sampling turn out to be highly consequential as we shall see.

3 Evaluation

3.1 Simulation Study

Our initial evaluation uses data from the Finance components of the 2012 and 2017 Census of Governments. The universe is the intersection of 2012 data with 2017 data, including only the units surveyed during both census years, and hence ignores the effect of births and deaths over the life of the sample. The universe for this evaluation is comprised of 90,144

units. As mentioned in the introduction some data on expenditures, revenues, and assets of public insurance and retirement systems are provided via the Annual Survey of Public Pensions. However the contribution of pension systems to total expenditures and total revenues of all local governments in the United States were 2.82% and 4.58% respectively as of the 2017 Census of Governments, and in all individual states insurance and pension systems accounted for less than 10% of total revenue and expenditures of local governments. Between the minimal contributions made by these pension systems to the national and state totals and the high sampling rates of the Annual Survey of Public Pensions our analysis can safely ignore the effect of these items and the sample design of the ASPP. By way of comparison insurance trust funds accounted for 35.2% of total cash and securities held by all local governments in the United States, and in some individual states insurance and retirement systems account for more than 60% of total cash and security holdings.

The 2012 CoG-F provides the auxiliary data, and serves as the sampling frame. The 2014 and 2019 sampling designs are both applied to select 1,000 replicated samples from the frame created from the 2012 CoG-F data. For each sample replicate we estimate the 2012 totals for all key variables and ensure that they meet initial precision requirements as we would when selecting our production sample, and in turn estimate the 2017 state totals for all key variables at all required levels of aggregation in order to see how well both sample designs perform several years after initial selection. In general we are interested in estimating a total of the *p*th key variable in state *j*, $Y_{jp} = \sum_{k \in U_j} y_{kp}$. For the proposed single-phase STSI design we will generally utilize the Horvitz-Thompson estimator $\hat{Y}_{jp}^{HT} = \sum_{k \in s_j} \breve{y}_{kp} = \sum_{k \in s_j} y_{kp}/\pi_k$, while for the two-phase πps design we will generally use the two phase analogue to the Horvitz-Thompson, the double expansion estimator $\hat{Y}_{jp}^{DE} = \sum_{k \in s_j} \breve{y}_{kp} = \sum_{k \in s_j} y_{kp}/\pi_{kp}^{*} = \sum_{k \in s_j} y_{kp}/(\pi_{1k}\pi_{2k})$. During the analysis we computed the relative root mean squared error (RRMSE) and relative bias for the survey estimator under both designs. We also evaluated the relative efficiency of alternative survey estimators under both designs.

3.1.1 Relative Root Mean Squared Error (RRMSE)

We used the mean squared error (MSE) as a primary measure for evaluating estimator quality under the competing sample designs. In this study, we calculate MSE for the survey estimator of key totals under both designs over all sample replicates. The Monte Carlo MSE for an estimator in a tabulation cell is calculated as:

$$\widehat{MSE}_{MC}(\hat{Y}_{jp}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{jp}^{(r)} - Y_{jp}\right)^2$$

where $\hat{Y}_{jp}^{(r)}$is the estimated total of a variable of interest for one sample replicate (r), and Y_{jp} is the true total. As mentioned previously we take R= 1,000. In order to compare sample designs we normalize the MSE by its corresponding cell value, giving us the Relative Root Mean Squared Error (RRMSE):

$$\widehat{RRMSE}_{MC}(\hat{Y}_{jp}) = 100 \times \left(\frac{\sqrt{\widehat{MSE}_{MC}(\hat{Y}_{jp})}}{Y_{jp}}\right)\%$$

We want to see not only which design gives a lower mean squared error for the key variables, but also in particular how well each design performs at meeting CV requirements where possible. It is not expected that any sample design will meet the initial CV requirements in all tabulation cells in subsequent survey years. The sample selected in 2019, for example, must cover the survey years 2019-2021 and 2023 (2022 being a census year), and over time the correlation between a unit's measure of size and its reported values for key variables can decrease. Long-term debts in particular can be highly volatile from year to year.

3.1.2 Relative Bias

The bias of an estimator is measured as the difference between its expected value and the true value of the parameter being estimated. In our evaluation, relative bias is calculated for a key variable as:

$$\widehat{RB}_{MC}\left(\hat{Y}_{jp}\right) = 100 \times \left(\frac{1}{R}\sum_{r=1}^{R}\frac{\hat{Y}_{jp}^{(r)} - Y_{jp}}{Y_{jp}}\right)\%$$

In accordance with theory we expect the Horvitz-Thompson estimator and the double-expansion estimator to be unbiased under both sample designs, that is $E\left[\hat{Y}_{jp}^{HT}\right] = Y_{jp}$. However a proposed alternative estimator that will be evaluated in this study is not necessarily unbiased. Additionally we are interested in the relative bias of standard estimators of variance under our two designs, defined as:

$$\widehat{RB}_{MC}\left[\hat{v}\left(\hat{Y}_{jp}\right)\right] = 100 \times \left[\frac{1}{R}\sum_{r=1}^{R}\frac{\hat{v}^{(r)}\left(\hat{Y}_{jp}^{(r)}\right) - V\left(\hat{Y}_{jp}\right)}{V\left(\hat{Y}_{jp}\right)}\right]\%$$

3.1.3 Relative Efficiency

Suppose that two estimators $\hat{Y}_{jp}^{1}$ and $\hat{Y}_{jp}^{2}$ are used to estimate the same population parameter, Y_{jp}. These two estimators could be the survey estimator from two different sample designs, or the survey estimator and a robust estimator under the same sample design. The Monte Carlo relative efficiency of the 2nd estimator using the 1st as a reference is defined as the ratio of their Monte Carlo mean square errors:

$$\widehat{RE}_{MC}\left(\hat{Y}_{jp}^{1}, \hat{Y}_{jp}^{2}\right) = 100 \times \left[\frac{\widehat{MSE}_{MC}\left(\hat{Y}_{jp}^{2}\right)}{\widehat{MSE}_{MC}\left(\hat{Y}_{jp}^{1}\right)}\right]\%$$

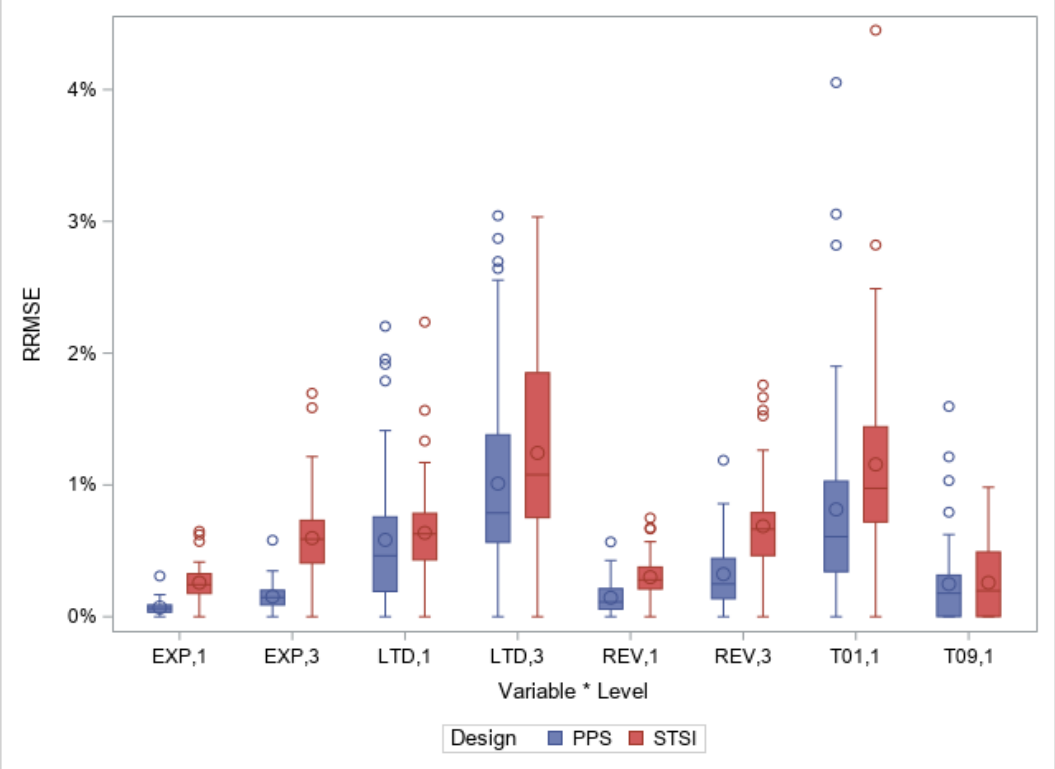
Note that a relative efficiency of less than 100% means that the second estimator is more efficient than the first.

3.1.4 Sample Performance

Figure 1 shows the distribution of RRMSEs under both designs at the time of selection from the frame created from the 2012 CoG-F. As previously mentioned design requirements stipulate that at the time of selection the sample must give a CV of 3% or less for the key variables of total expenditures (EXP), total revenues (REV), and long-term debts (LTD) for all states at both the local only (level 3) and state & local (level 1) levels of aggregation, and a CV of 5% or less for property tax (T01) and sales tax (T09) for all states at the state & local level (level 1) of aggregation. Note that totals for state governments only (level 2) are based on a full census, and so by definition have no

sampling error; sampling variances for level 1 totals of a given variable will therefore always be less than or equal to those of level 3 totals for the same variable. As Figure 1 shows both designs meet all CV requirements at the time of design, with the πps design giving on average a slightly lower RRMSE. In fact the STSI design gave a lower RRMSE for only 22% of cells that are subject to CV requirements at the time of sample design. However the loss of efficiency compared to probability proportional-to-size sampling is trivial—as can be seen by comparing the median RRMSE of both the πps and STSI designs.

Figure 1. Distribution of RRMSEs for Sample Designs at Initial Selection



Data Source: U.S. Census Bureau, 2012 Census of Governments: Finance

As shown in Table 1 both designs also give excellent precision at the national level for three QTax variables at the time of design and selection:

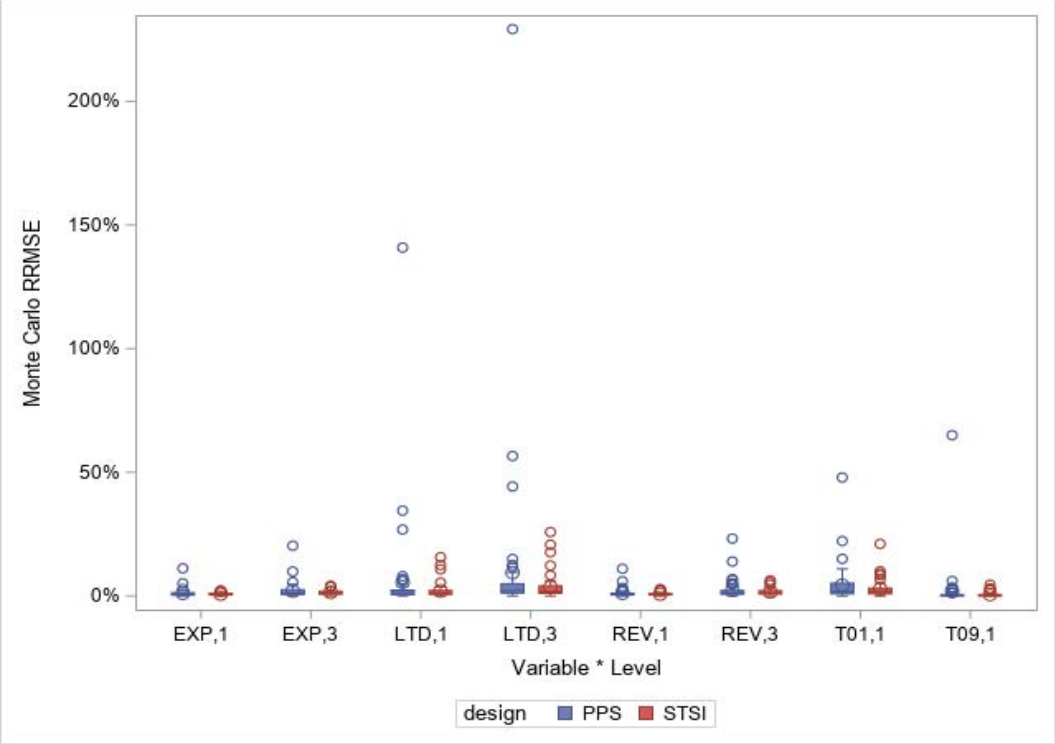
Table 1: Design RRMSEs of Key Tax Variables at Initial Selection

Variable	Level	$\widehat{RRMSE}_{MC}(\hat{Y}^{DE}_{\pi ps})$	$\widehat{RRMSE}_{MC}(\hat{Y}^{HT}_{STSI})$
T09 (sales tax)	3	0.45%	0.36%
T40 (income tax)	3	0.43%	0.62%
T41 (corporate income tax)	3	0.30%	0.59%

Data Source: U.S. Census Bureau, 2012 Census of Governments: Finance

However, checking the performance of the two sample designs 5 years after the initial design gives a very different picture. As shown in Figure 2 the STSI design is much more efficient for all key variables when used to give estimates for 2017.

Figure 2. Distribution of RRMSEs for Sample Designs 5 Years After Selection



Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

The πps design encounters particular problems with variables that are not always well correlated with a unit's measure of size. As can be seen from the maximum RRMSE of over 200% for the variable of long-term debts at the local government only level this variable can have extremely high sampling variances under the old sample design in part because it is possible for small units with large survey weights to also report very large values for this variable in subsequent survey years. Consistent with theory stratified simple random sample designs are more resilient to changes in unit size over time. In fact, we find that for this variable the old sample design runs into problems even at the national level of aggregation as shown in Table 2.

Table 2. RRMSEs of Key Variables at National Level of Aggregation Under Both Sample Designs 5 Years After Selection

Variable	Level	$\widehat{RRMSE}_{MC}(\hat{Y}^{DE}_{\pi ps})$	$\widehat{RRMSE}_{MC}(\hat{Y}^{HT}_{STSI})$
EXP	1	0.29%	0.11%
EXP	3	0.51%	0.20%
LTD	1	5.73%	0.39%
LTD	3	9.24%	0.63%
REV	1	0.27%	0.09%
REV	3	0.56%	0.20%
T01	1	0.85%	0.33%
T09	1	1.12%	0.15%

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

Similarly the πps design offers much less precision for one of the variables necessary for benchmarking the QTax survey. Because most of the local governments that collect personal and corporate income taxes are relatively large and therefore much of the national total for these variables comes from certainty cases, it is expected that both designs will offer reasonable precision. By contrast governments of all sizes collect sales taxes and this variable is potentially vulnerable to high sampling variances as shown in Table 3. Even five years after the initial sample is drawn the STSI design maintains an RRMSE of less than 1% for the three benchmark totals, while the πps design has an RRMSE of nearly 5% for sales tax.

Table 3: RRMSEs of Key Tax Variables 5 Years After Initial Design

Variable	Level	$\widehat{RRMSE}_{MC}(\hat{Y}_{\pi ps}^{DE})$	$\widehat{RRMSE}_{MC}(\hat{Y}_{STSI}^{HT})$
T09 (sales tax)	3	4.90%	0.63%
T40 (income tax)	3	0.50%	0.69%
T41 (corporate income tax)	3	0.32%	0.61%

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

As noted previously we do not expect any sample design to meet all CV constraints in years subsequent to initial selection. However it is desirable that the survey violate as few constraints as possible. Here again we find that the STSI design outperforms the πps design. Table 4 shows the number of constraint violations for each key variable at each relevant level of aggregation for both designs. Note that here there are 52 cells to consider at each level: the 50 states, the District of Columbia, and the national aggregate.

Table 4. Number of Cells Where Each Design Exceeds CV Requirements in 5th Sample Year

Variable	Level	Number of Cells	Violations: $\hat{Y}_{\pi ps}^{DE}$	Violations: $\hat{Y}_{STSI}^{HT}$
EXP	1	52	2	0
EXP	3	52	9	4
LTD	1	52	10	6
LTD	3	52	22	23
REV	1	52	3	0
REV	3	52	11	6
T01	1	52	14	6
T09	1	52	2	0
Total	Total	416	73	45

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

In addition to the design constraints Census Bureau standards only allow the publication of totals with a CV of 30% or less. Under our old design based on the RRMSEs observed in our simulation study we would potentially be required to suppress 7 tabulation cells for the key variables as shown, a problem that the new sample design avoids entirely.

Table 5. Required Suppressions of Key Variables

Variable	Level	Suppressions: $\hat{Y}_{\pi ps}^{DE}$	Suppressions: $\hat{Y}_{STSI}^{HT}$
LTD	1	2	0
LTD	3	3	0
T01	1	1	0
T09	1	1	0

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

3.1.5 Variance Estimation

The 2014 sample design presents two challenges from the perspective of estimation of variance: the challenge of variance estimation for systematic sampling, and the challenge of estimating the additional variance due to the two phase design. Systematic πps sampling is widely used by survey practitioners because it allows the selection of a sample with optimal probabilities while avoiding the programming and computational difficulties of alternative methods of selecting a fixed-size without replacement sample where units have unequal probabilities of selection. Implemented properly the method affords the opportunity to combine the efficiency of proportional-to-size sampling (assuming the variables of interest are highly correlated with the size variable) and the ability of systematic sampling to exploit both hidden and explicit stratification in the population. (Wolter 2007) In practice however systematic sampling generally suffers from the drawback that unbiased estimators for the sampling variance are not easily derived, with this problem becoming particularly acute when unequal selection probabilities are used. One possible solution to this problem is to use the Yates-Grundy-Sen estimator of the variance for a (non-systematic) πps sample of fixed size:

$$\hat{v}_{YGS}\left(\hat{Y}^{HT}\right) = -\frac{1}{2}\sum_{k\in S}\sum_{l\in S}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\right)\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2$$

Unfortunately this estimator comes with several important caveats. In general standard variance estimators are not unbiased for systematic sampling and must be used with caution if at all. Additionally the double sum is cumbersome, and the joint selection probabilities (π_{kl}) may not be readily available and may therefore need to be approximated. Commonly used approximations of the second order selection probabilities (such as that of Hajek) often lead to variance estimators that are highly sensitive to the strength of the correlation between the measure of size and the study variable of interest (Haziza et al 2004). Kirk Wolter proposes several alternative estimators but concedes that no single estimator is optimal in all cases (Wolter, 2007). Bengt Rosén proposed an approach based on the heuristic of "successive" πps sampling as an alternative to the more commonly used heuristic of with replacement sampling (Rosén 1991). All the estimators mentioned so far would also need to be combined with additional methods in order to account for the two phase design (Beaumont et al 2015). In a typical survey year ALFIN must publish totals and CVs for thousands of tabulation cells and a simple variance estimator that is available via standard statistical software packages is therefore desirable.

Many survey practitioners use the variance estimator for probability proportional-to-size with replacement (pps wr) sampling:

$$\hat{v}_{WR}\left(\hat{Y}^{HT}\right) = \frac{1}{n(n-1)}\sum_{k\in S}\left(\frac{ny_k}{\pi_k}-\hat{Y}^{HT}\right)^2$$

This estimator of the sampling variance is conservative, that is $E\left[\hat{v}_{WR}\left(\hat{Y}^{HT}\right)\right] > V\left(\hat{Y}^{HT}\right)$ and easy to compute due to the avoidance of double sums and second order inclusion probabilities, and therefore widely used in practice (Särndal et al 1992). It is easily implemented in SAS™ software using the SURVEYMEANS procedure. Of particular relevance it was assumed to be sufficiently conservative to account for the additional variance introduced by the two phase design utilized in 2014 (Cheng, 2012) and was used in survey years 2014-2018 to estimate the variance of published totals from ALFIN.

In contrast stratified simple random sampling encounters no such difficulties. The stratum variance is estimated unbiasedly by the standard formula for simple random sampling without replacement (where $\bar{y}_h = \frac{1}{n_h}\sum_{k\in s_h} y_k$ is the stratum sample mean):

$$\hat{v}_h(\hat{Y}_h) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h - 1}\sum_{k\in s_h}(y_k - \bar{y}_h)^2$$

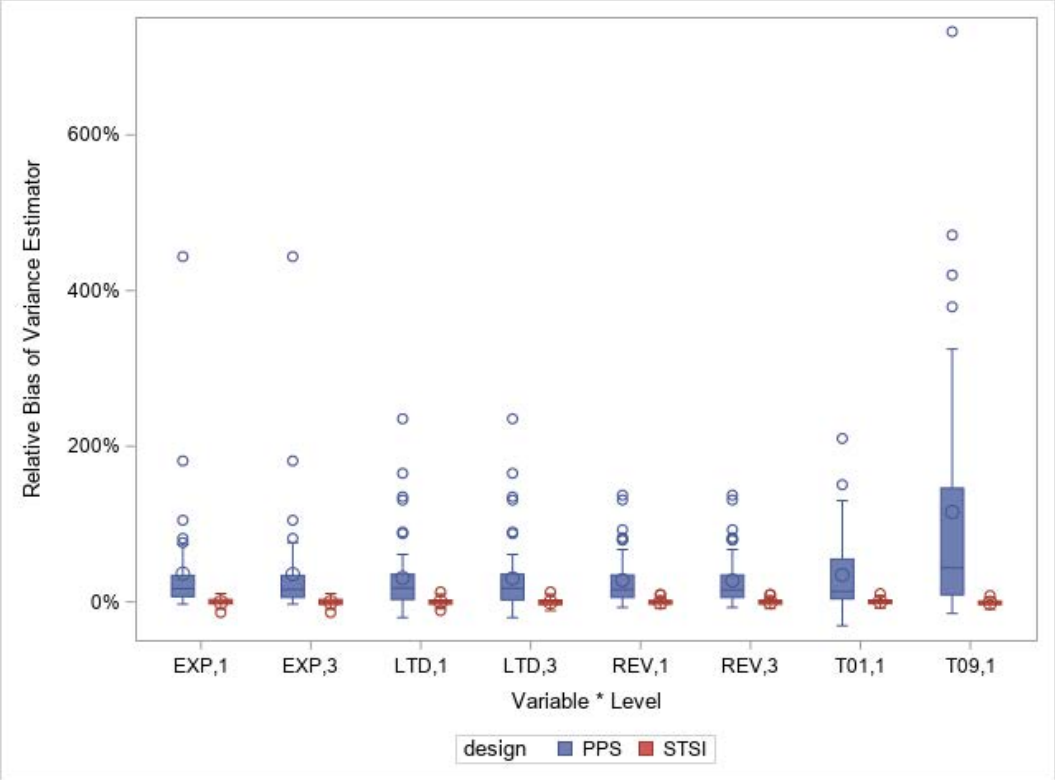
And in turn the variance for a given total is estimated by the sum of stratum variance estimators:

$$\hat{v}_{ST}(\hat{Y}^{HT}) = \sum_{h=1}^{H}\hat{v}_h(\hat{Y}_h)$$

This estimator is also easily implemented via the SURVEYMEANS procedure in the SAS™ programming language. While many other options are available for estimating the sampling variance of a stratified simple random sample (again see Wolter 2007) for this analysis we will use the standard formula.

Figure 3 shows the distribution of the relative bias of variance estimators across tabulation cells that are checked against CV requirements. In some cases $\hat{v}_{WR}$ is punishingly conservative, with a maximum relative bias of over 700% for the key variable of sales tax (T09). Even the median value of the relative bias for the estimated variance of sales tax is over 40% while as expected the mean and median relative bias for $\hat{v}_{ST}$ are approximately 0%. But in many cases $\hat{v}_{WR}$ is also more likely to underestimate the sampling variance as well.

Figure 3. Relative Bias of Standard Variance Estimators Under Both Designs



Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

Retaining the old sample design would therefore likely require further research on a suitable method for estimation of variances.

3.1.6 Robust Survey Estimators

One further motivation for the use of stratified simple random sampling is that it facilitates the introduction of estimators that are robust to influential units. If Y is a total of interest and $\hat{Y}^{HT}$ the Horvitz-Thompson estimator of the total under a generalized design, we measure the influence of the kth unit on the estimated total by its conditional bias, where I_k is the inclusion indicator:

$$B_k^{HT}(I_k = 1) = E_p\left(\hat{Y}^{HT} - Y | I_k = 1\right) = \left(\frac{1}{\pi_k} - 1\right) y_k + \sum_{l \neq k} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) y_l$$
$$= \sum_{l \in U} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) y_l$$

Recall that the sampling variance of the Horvitz-Thompson estimator under a generalized design can be written as:

$$V_p\left(\hat{Y}^{HT}\right) = \sum_{k \in U} \sum_{l \in U} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) y_k y_l = \sum_{k \in U} B_k^{HT}(I_k = 1) y_k$$

That is the sampling variance of the Horvitz-Thompson estimator of a population total can be thought of as a weighted sum of each population unit's influence function, or conversely a unit's influence can be thought of as its contribution to the sampling variance with more influential units contributing more to the total sampling variance. Certainty units (that is units with $\pi_k = 1$) can be shown to have a conditional influence of 0.

In practice a unit's influence function depends on unknown quantities (namely variables of interest for non-sampled units) and must be estimated for each sample unit. One common estimator of a sample unit's conditional bias is:

$$\hat{B}_k^{HT}(I_k = 1) = \sum_{l \in s} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}\right) y_l$$

Note that the determination of a unit's influence requires the joint inclusion probabilities, π_{kl}. As mentioned in the previous section, these are not always readily available under many sample designs, including the two-phase πps design utilized in 2014, and would have to be estimated or approximated. However the joint selection probabilities are easily available under a stratified simple random sampling design and a unit's estimated influence function can be shown to be:

$$\hat{B}_k^{HT}(I_k = 1) = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_k - \bar{y}_h)$$

That is a unit will be influential if its reported value is far from the stratum sample mean.

Once we have estimated each sample unit's influence function we are able to estimate the robust HT of Beaumont, Haziza, and Ruiz-Gazen:

$$\hat{Y}^{RHT} = \hat{Y}^{HT} - \frac{1}{2}\left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}\right)$$

$$\hat{B}_{min}^{HT} = min_{k\in s}\left(\hat{B}_k^{HT}(I_k = 1)\right), \hat{B}_{max}^{HT} = max_{k\in s}\left(\hat{B}_k^{HT}(I_k = 1)\right)$$

Under regularity conditions this estimator is design consistent, with $\hat{Y}^{RHT} - Y = O_p\left(Nn^{-1/2}\right)$. For more information on influence in finite population inference and robust survey estimation see Beaumont, Haziza, and Ruiz-Gazen, 2013.

Implementation of a similar robust estimator under the two-phase πps design presents a more formidable challenge. Using the results of Favre-Martinoz, Haziza, and Beaumont (2016), we note that under a two phase design the optimal robust survey estimator is similarly defined as (where DE denotes the double-expansion estimator):

$$\hat{Y}^{RDE} = \hat{Y}^{DE} - \frac{1}{2}\left(\hat{B}_{min}^{DE} + \hat{B}_{max}^{DE}\right)$$

$$\hat{B}_{min}^{DE} = min_{k\in S_2}\left(\hat{B}_k^{DE}(I_{1k} = 1, I_{2k} = 1)\right), \hat{B}_{max}^{DE} = max_{k\in S_2}\left(\hat{B}_k^{DE}(I_{1k} = 1, I_{2k} = 1)\right)$$

However we do not have a simple expression for the unit level influence functions as under the STSI design. Instead we have:

$$\hat{B}_k^{DE}(I_{1k} = 1, I_{2k} = 1) = \sum_{l\in S_2} \frac{\pi_{1k}}{\pi_{1kl}} \frac{\pi_{2k}}{\pi_{2kl}} \left(\frac{\pi_{kl}^*}{\pi_k^*\pi_l^*} - 1\right) y_l$$

Second order selection probabilities at the first phase would need to be obtained via an approximation such as that of Hartley and Rao, 1962:

$$\begin{aligned}\pi_{1kl} \approx{}& \frac{(n-1)}{n}\pi_{1k}\pi_{1l} + \frac{(n-1)}{n^2}\left(\pi_{1k}^2\pi_{1l} + \pi_{1k}\pi_{1l}^2\right) - \frac{(n-1)}{n^3}\pi_{1k}\pi_{1l}\sum_{l=1}^{N}\pi_{1l}^2 \\ &+ \frac{2(n-1)}{n^3}\left(\pi_{1k}^3\pi_{1l} + \pi_{1k}\pi_{1l}^3 + \pi_{1k}^2\pi_{1l}^2\right) \\ &- \frac{3(n-1)}{n^4}\left(\pi_{1k}^2\pi_{1l} + \pi_{1k}\pi_{1l}^2\right)\sum_{l=1}^{N}\pi_{1l}^2 \\ &+ \frac{3(n-1)}{n^5}\pi_{1k}\pi_{1l}\left(\sum_{l=1}^{N}\pi_{1l}^2\right)^2 - \frac{2(n-1)}{n^4}\pi_{1k}\pi_{1l}\sum_{l=1}^{N}\pi_{1l}^3\end{aligned}$$

For this particular sample design the robust estimator would in practice be difficult to automate and implement in a production setting. It is worth noting that Favre-Martinoz et al generally apply their results to an arbitrary design in the first phase followed by an implicit Poisson sample at the second phase as a way of conducting robust estimation in the presence of unit non-response, and in this setting the estimator is much more tractable. In particular their simulation study examines the case of a simple random sample in the first phase and a Poisson sample (with probabilities obtained via a propensity score method) in the second. For more details about robust estimation in two phase sampling see Favre-Martinoz et al 2016.

In our simulation study we applied the relevant robust estimator to cells where either sample design encountered variances over initial CV requirements for the variable of long-term debts at the local level of aggregation in the 2017 survey year. This set of cells were chosen because as seen previously this variable is the most likely to give high sampling variances at both the design and estimation phases. In all cases where the STSI design encounters a high sampling variance the robust estimator produces modest to large reductions in the mean squared error with only modest downward bias.

Table 6. Robust Estimator for Long-Term Debts in Problem States, STSI Design

State	$\widehat{RRMSE}_{MC}(\hat{Y}^{HT}_{STSI})$	$\widehat{RRMSE}_{MC}(\hat{Y}^{RHT}_{STSI})$	$\widehat{RB}_{MC}(\hat{Y}^{HT}_{STSI})$	$\widehat{RE}_{MC}(\hat{Y}^{HT}_{STSI}, \hat{Y}^{RHT}_{STSI})$
AL	4.19%	3.06%	-0.63%	53.48%
AR	4.37%	3.53%	-0.98%	65.20%
CO	5.19%	4.60%	-1.49%	78.62%
GA	3.04%	2.42%	-0.49%	62.97%
ID	4.66%	3.88%	-0.93%	69.26%
IA	3.50%	3.19%	-0.66%	83.06%
KS	3.02%	2.34%	-0.52%	60.06%
KY	4.13%	3.76%	-0.96%	83.03%
LA	3.20%	2.77%	-0.70%	75.37%
ME	3.09%	2.97%	-0.51%	91.94%
MA	3.09%	2.41%	-0.30%	61.21%
MS	3.75%	3.03%	-0.54%	65.35%
MO	25.80%	15.81%	-2.29%	37.57%
MT	5.97%	4.72%	-0.63%	62.46%
NH	8.36%	6.72%	-1.32%	64.54%
NJ	3.74%	3.02%	-0.49%	65.05%
ND	4.62%	3.82%	-1.08%	68.52%
PA	3.28%	2.51%	-0.47%	58.42%
SC	20.71%	13.75%	-1.06%	44.09%
SD	12.28%	9.16%	-1.51%	55.59%
VT	3.72%	3.54%	-0.71%	90.78%
WV	5.82%	5.67%	-0.91%	95.00%
WY	17.65%	13.80%	-2.93%	61.15%

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

By contrast while the robust double-expansion estimator does produce major gains in efficiency in some states (such as Pennsylvania and Wyoming) it does not always produce a more efficient estimate, and in extreme cases (such as Colorado and Oklahoma) produces a much higher mean squared error than the simple non-robust survey estimator. Additionally while the robust survey estimator under the STSI design does not produce any tabulation cell with an absolute relative bias of more than 3%, under the πps design we encounter 4 cells where the absolute relative bias of the robust estimator is larger than 3% (Colorado, Oklahoma, Pennsylvania, and Wyoming). Extreme caution would need to be applied if using this estimator under the old sample design.

Table 7. Robust Estimator for Long-Term Debts in Problem States, πps Design

State	$\widehat{RRMSE}_{MC}(\hat{Y}^{DE}_{\pi ps})$	$\widehat{RRMSE}_{MC}(\hat{Y}^{RDE}_{\pi ps})$	$\widehat{RB}_{MC}(\hat{Y}^{RDE}_{\pi ps})$	$\widehat{RE}_{MC}(\hat{Y}^{DE}_{\pi ps}, \hat{Y}^{RDE}_{\pi ps})$
AL	9.50%	11.24%	1.24%	139.90%
CO	8.61%	83.96%	-6.05%	9516.04%
FL	4.74%	8.60%	-1.09%	328.31%
ID	4.01%	3.93%	-0.42%	95.76%
KY	3.65%	3.86%	0.28%	111.67%
ME	3.18%	2.35%	-0.69%	54.49%
MA	11.54%	5.80%	-0.33%	25.28%
MS	4.31%	2.62%	-0.72%	37.01%

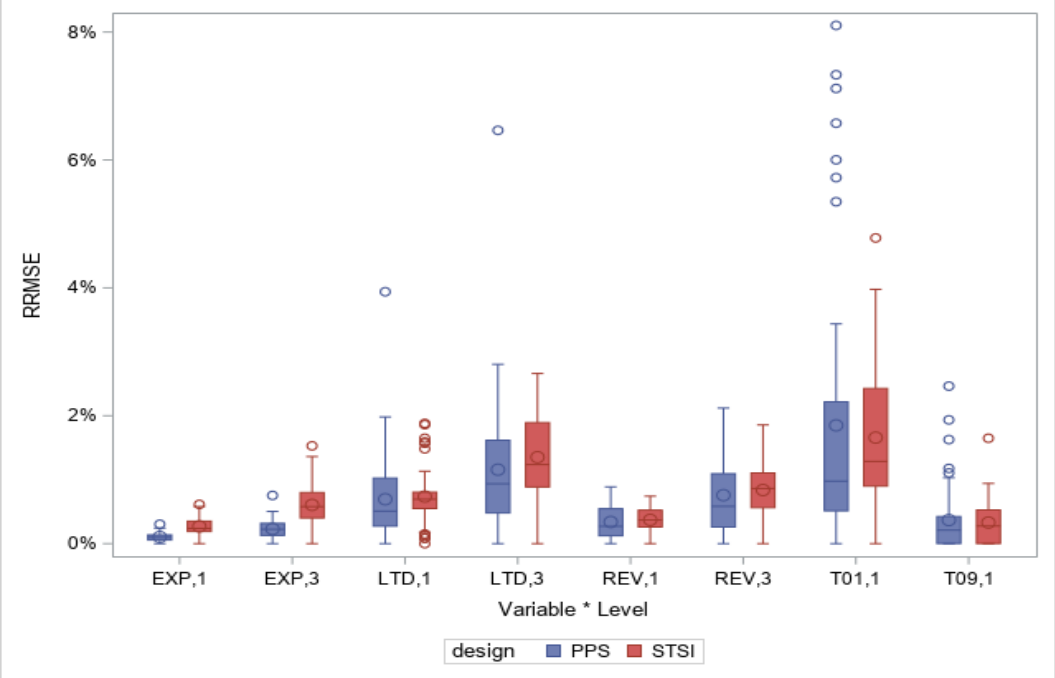
State	$\widehat{RRMSE}_{MC}(\hat{Y}_{\pi ps}^{DE})$	$\widehat{RRMSE}_{MC}(\hat{Y}_{\pi ps}^{RDE})$	$\widehat{RB}_{MC}(\hat{Y}_{\pi ps}^{RDE})$	$\widehat{RE}_{MC}(\hat{Y}_{\pi ps}^{DE}, \hat{Y}_{\pi ps}^{RDE})$
MO	44.24%	43.37%	0.72%	96.12%
MT	5.70%	1.88%	-1.88%	10.82%
NJ	6.04%	5.31%	-1.10%	77.28%
ND	12.51%	13.80%	-1.77%	121.60%
OH	4.23%	7.30%	-2.48%	297.31%
OK	3.71%	28.47%	4.43%	5888.93%
PA	229.24%	114.79%	3.11%	25.07%
SC	5.07%	5.13%	-0.22%	102.49%
SD	14.91%	14.96%	0.18%	100.64%
VT	3.51%	2.62%	-0.84%	55.57%
VA	3.18%	1.82%	-0.57%	32.81%
WV	5.66%	7.34%	0.65%	167.97%
WY	56.56%	29.56%	-3.10%	27.32%

Data Source: U.S. Census Bureau, 2012 and 2017 Census of Governments: Finance

4 Overview of the 2019 Production Sample

In evaluating the new sampling design a simulation study was also performed using the frame created from the 2017 CoG-F, with the distribution of RRMSEs for key variables from the simulation study shown in Figure 4.

Figure 4. Distribution of RRMSEs for Key Variables at Time of 2019 Sample Design



Data Source: U.S. Census Bureau, 2017 Census of Governments: Finance

As can be seen from the median RRMSE across all states and consistent with our previous simulation study the STSI design's loss of efficiency at the time of initial sampling compared to the πps design is trivial. Additionally under the 2014 stratum allocations the

πps design does not meet all CV constraints for long-term debts and property taxes, and would have required subsequent adjustments in sampling rates for underperforming states in order to meet these initial requirements.

Table 8. CV Requirement Violations at Time of 2019 Sample Design Research

Variable	Level	Violations: πps Design	Violations: STSI Design
EXP	1	0	0
EXP	3	0	0
LTD	1	1	0
LTD	3	1	0
REV	1	0	0
REV	3	0	0
T01	1	7	0
T09	1	0	0

Data Source: U.S. Census Bureau, 2017 Census of Governments: Finance

At the time of sample selection there were also 1,058 birth units, defined as governments that were discovered or created since the 2017 CoG freeze. Because these units were not active at the time of the 2017 CoG-F there was little to no auxiliary information available for them, and other selection methods were needed. A birth sample of 350 units was selected. All counties, cities, towns, and independent school districts were taken with certainty. Additionally, any special district with long-term debts in the 2018-2019 period based on administrative records was also taken with certainty, along with special districts in states with 3 or fewer special district births. The remaining units were stratified by state, with allocations determined by a hybrid method that took the maximum of (a) the proportion of the national total measure of size contributed by that state in 2017 and (b) simple proportional allocation based on the number of special districts in that state. In prior years most birth units have been extremely small, and contributed very little to key aggregates.

5. Conclusions and Future Research

As shown in our simulation study the new 2019 sample design outperforms the old 2014 sample design according to multiple criteria: precision of survey estimators for key variables over time, ease of obtaining an unbiased estimate of the sampling variance, and the ability to troubleshoot tabulation cells with high variances with estimators that are reliably robust to influential units and simple to implement in a production environment. The difficulties inherent in variance estimation and robust estimation for the two-phase πps design might be justified if the design resulted in increased estimate precision compared to a less complicated design, but in fact we found the opposite to be the case. In all areas where further research is called for the simplicity of stratified simple random sampling and the existence of a large literature on this design make further improvements possible as well. While $\hat{v}_{STSI}$ is not unbiased for all cells in our analysis we believe the availability of alternative estimators of sampling variance in addition to the standard formula for simple random sampling without replacement will allow us to continue improving variance estimation in future research.

Many areas for future study remain. As mentioned in previous sections the two-phase design was implemented in 2014 in order to increase unit response rates and reduce

respondent burden by cutting the number of small units in the sample. We believe that direct control over the number of small units via stratification by size will achieve the same goals, but response rates in future survey years will need to be monitored carefully in order to ensure that this objective is achieved. The new sample design gives us the ability to estimate variances via the standard estimator for simple random sampling without replacement, but it may be worth investigating alternative estimators such as BRR and the stratified jackknife to see if they can offer further improvements. Of the five key variables that are subject to CV requirements three (total revenues, total expenditures, and long term debts) are derived items. Long-term debts for example are a sum of the items 44T (Long-term Debt Outstanding, End Of Fiscal Year, Public Debt For Private Purposes) and 49U (Long-term Debt Outstanding, End of Fiscal Year, Unspecified Public Purposes) for which estimated totals are also published. Any attempt at introducing a robust survey estimator for long-term debts must therefore ensure that estimates for individual states are consistent with estimates of the national total, and that estimates of individual items are consistent with estimates for derived items. An estimator of the mean squared error will also have to be investigated for robust survey estimators, with the generalized bootstrap as the most promising method (Beaumont et al 2013).

References

Bankier, M.D. (1988). Power Allocations: determining sample sizes for subnational areas. *The American Statistician* 42, 174-177.

Beaumont, J.-F., Haziza, D. & Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* **100**, 555-569.

Beaumont, J.-F., Beliveau, A. & Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology* 3, 524-542

Chambers, R. and Clark, R. (2012). An Introduction to Model-Based Survey Sampling with Applications. New York, NY: Oxford University Press.

Cheng, Yang. 2012. New Technique for Modifying the Cutoff Sample and Its Application. Governments Division Report Series, Research Report #2012-2

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). New York, NY: Wiley.

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, *54*, 88-101.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scandinavian Journal of Statistics* **43**, 1019-1034

Hartley, H.O. and Rao, J.N.K. (1962), Sampling with Unequal Probabilities and Without Replacement, The Annals of Mathematical Statistics 33, 350-374.

Haziza, D., Mecatti, F. & Rao, J. N. K. (2004), Comparison of variance estimators under Rao-Sampford method: a simulation study. *Proceedings of the Survey Methods Section*, American Statistical Association.

Rosén, Bengt (1991). Variance Estimation for Systematic PPS-Sampling. Stockholm, Sweden: Statistics Sweden.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.

Wolter, K (2007). *Introduction to Variance Estimation* (2nd ed.). New York, NY: Spring Science+ Business Media LLC.

Wright, R.L. (1983). Finite Population Sampling With Multivariate Auxiliary Information. *Journal of the American Statistician* 78, 879-884.