

R Institute for Policy Research Northwestern University *Working Paper Series*

Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds

Zachary H. Seeskin

Graduate Research Assistant, Institute for Policy Research Doctoral Candidate, Department of Statistics Northwestern University

Bruce D. Spencer

Faculty Fellow, Institute for Policy Research Professor of Statistics Northwestern University

Version: May 18, 2015

Address correspondence to Bruce D. Spencer, Department of Statistics, Northwestern University, 2006 Sheridan Rd., Evanston IL 60208-4070 Email: <u>bspencer@northwestern.edu</u>.

This research was supported by NSF grant SES-1129475, "NCRN-SN: Census Bureau Data Programs as Statistical Decision Problems".

DRAFT Please do not quote or distribute without permission.

Abstract

How much accuracy is needed in the 2020 census depends on the cost of attaining accuracy and on the consequences of imperfect accuracy. The cost target for the 2020 census of the United States has been specified, and the Census Bureau is developing projections of the accuracy attainable for that cost. It is desirable to have information about the consequences of the accuracy that might be attainable for that cost or for alternative cost levels. To assess the consequences of imperfect census accuracy, Seeskin and Spencer consider alternative profiles of accuracy for states and assess their implications for apportionment of the U.S. House of Representatives and for allocation of federal funds. An error in allocation is defined as the difference between the allocation computed under imperfect data and the allocation computed with perfect data. Estimates of expected sums of absolute values of errors are presented for House apportionment and for federal funds allocations.

1. Introduction

The U.S. Constitution requires that the population be enumerated decennially, for purposes of allocating Representatives among the states.

Representatives shall be apportioned among the several States according to their respective numbers, counting the whole number of persons in each State, excluding Indians not taxed. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

 The Constitution of the United States, Article I, Section 2, as amended by the 14th amendment

Although the Constitution requires a census, it does not say how accurate the census should be. Accuracy and cost are closely related. The Census Bureau can increase accuracy by spending more money, at least up to a point. As the great demographer Nathan Keyfitz (1979, 46) noted, "Asking why the census cannot [accurately] count 100 percent of the population in a free society is like asking why books contain typographical errors, why manufactured products often have defects, or why the police cannot catch all criminals." Accuracy can be increased through investment of more resources in the census, but the accuracy will never be perfect.

The current strategy for choosing census accuracy is to specify a cost target and optimize the accuracy that can be attained for that cost. The cost target, as we understand it, is consistent with recommendation of the National Research Council (2011, Recommendation 3) that the cost per housing unit for 2020 be kept at the same (inflation-adjusted) level as for 2010. The 2010 census was estimated to be quite accurate for the total U.S. population, so that the national net undercount was estimated to be nearly zero (Census Bureau 2012). However, most uses of the census depend on population sizes for geographic areas or demographic subgroups, and the census was estimated to have a 0.8 percent national overcount of non-Hispanic whites, a 2.1 percent national undercount of blacks, a 1.5 percent undercount of Hispanics (Census Bureau 2012). There were net overcounts for some states and net undercounts for others.

The question of what accuracy is attainable for the specified is cost is complex and is being studied by the Census Bureau. In this paper, we address the related question of the consequences of a given

1

profile of census accuracy. We refer to profiles of accuracy rather than levels, because census accuracy, like census statistics, is multi-dimensional. For example, census population numbers are produced for state and local governments and much smaller areas, and for demographic subgroups both nationally and by geography.

To understand the consequences of imperfect accuracy, one needs to know how census data get used. The most visible uses of the census results include intergovernmental allocation of funds by formulas using population statistics, apportionment of the U.S. House of Representatives, and redrawing of Congressional district boundaries. When the census population numbers contain errors, the fund allocations, Congressional apportionment, and district sizes are different than what they would be if the census numbers had no error. We will refer to the differences as, respectively, errors in allocation (misallocations) of funds, errors in apportionment (malapportionment), and errors in district sizes. Not all uses of the census are or can be known, and it is important to acknowledge that some of the most important uses of the census may be the least visible, including research in social, economic, behavioral, medical, and policy areas and applications of that research. The role of census data in policy development and decision-making by the Congress and the White House, by state and local governments, and by businesses and other organizations has not received sufficient study, but we conjecture that it is important. For example, surveys are widely used sources of information, and almost all national population surveys – whether government or private sector, whether by internet, mail, phone, or in-person – directly or indirectly use decennial census numbers for adjusting their results.

In this paper we measure the distortions in allocations of representation and funding among states that are projected to occur at alternative profiles of accuracy. The funding formulas and the apportionment algorithm are treated as fixed, and the allocations that would occur with error-free statistics are treated as *true values* for the allocations. When actual statistics are used to compute the allocations, the resulting allocations are said to be empirical or *estimated*. The difference between the estimated allocation and the true allocation is called the *error in allocation* or, more simply, the *misallocation*. The term "error" is standard usage in statistics and does not imply that someone made a mistake. The value of improving accuracy to reduce misallocations is a political question that relates to the question of how much it is worth spending on the census, but that we do not address here.

We represent accuracy in terms of the multivariate distribution of errors in census estimates for states and the District of Columbia (DC). The focus on the states and DC level is consistent with the uses of census data for apportionment of House seats among states and for allocation of federal funds to states

2

and DC. The error in an estimate or statistic is the difference between the statistic and its true value. The mean squared error (MSE) is the expected value of the squared error, and it is equal to the square of the bias plus the square of the standard deviation. The root mean squared error (RMSE) is the square root of the MSE, and the relative root mean squared error (relative RMSE) is the RMSE expressed as a percentage of the true value being estimated. If the estimate is unbiased, the RMSE is equal to the standard deviation or standard error of the estimate, and the relative RMSE is equal to the coefficient of variation (c.v.). Initially, we will consider several alternative accuracy profiles for the census, as shown in Table 1. Alternative profiles are discussed in Section 4.

Accuracy Profile	Description
base case	errors for all states (and DC) are independently normally distributed, with zero bias, and equal relative RMSE
correlated case	same as base case, except estimates for all areas have constant correlation of 0.5
accurate small states case	errors for all states (and DC) are independently normally distributed, with zero bias; the smallest 25 states and DC have zero relative RMSE; the largest 25 states have constant relative RMSE, such that the weighted average of RMSEs for all 50 states and DC, weighted by the 2010 state census populations, equals the corresponding relative RMSE for the base case
differential bias case	errors for all states (and DC) are independently normally distributed, with equal relative RMSE and with absolute value of the relative bias equal to the c.v.; the bias for the smallest 25 states and DC has one sign, and the bias for the largest 25 states has the opposite sign

Table 1. Alternative Accuracy Profiles for the Census

The Constitution requires that seats in the U.S. House of Representatives be allocated to the states "according to their respective Numbers", and the law since 1941 has required that the so-called Method of Equal Proportions be used to calculate the apportionment (Balinski and Young 1982); Section 3 provides further explanation. Define the true allocation of seats a_j^{true} as number of seats given to state j when the Method of Equal Proportions is applied to the true population numbers (i.e., if the census were perfectly accurate). The empirical or estimated allocations a_j^{est} are the numbers of seats given when the apportionment method is applied to census population numbers. The difference $a_j^{est} - a_j^{true}$ is the error in allocation, or misallocation, to state j. The number of misallocated seats is defined as the sum of absolute errors (i.e., sum of absolute values of errors in apportionments),

$$\sum_{j} |a_{j}^{est} - a_{j}^{true}|.$$
⁽¹⁾

We used computer simulations to analyze the distribution of the sum of absolute errors. Figure 1 shows the strong linear relationship between the relative RMSE of state population estimates and the expected number of seats going to the wrong states in House apportionment.



Figure 1. Expected sum of malapportioned seats in the U.S. House of Representatives under alternative profiles of census accuracy.

It should be noted that Figure 1 shows the expected number of seats misallocated, and that for any particular census with a given accuracy profile the number of seats misallocated could be much greater than the expected number. For example, in the base case with relative RMSE equal to 1%, the expected number of malapportioned seats is estimated to be 3.4, but there is a 1 in 6 chance that 6 or more seats are malapportioned; with relative RMSE equal to 4%, there is a 3 in 4 chance that 12 or more seats are malapportioned.

Blumerman and Vidal (2009) identified 140 federal grant and direct assistance programs that apparently distributed funds at least partly on the basis of population and income data. These programs distributed

approximately \$446.4 billion in FY 2007. For analysis, we selected the 8 largest programs (in terms of FY 2007 obligated amount) with certainty, which accounted for approximately 4/5 of the total FY 2007 obligations of the 140 programs, and we selected a disproportionate stratified sample of 10 of the remaining 132 programs, so that larger programs had a higher chance of selection. We used sampleweighting methods to get unbiased estimates of totals for all 140 programs. For each sampled funds allocation program we estimated the expected sum of absolute errors in state allocations that would arise from alternative profiles of census accuracy. We found that for 4 of the 18 selected programs, the allocations would not be affected by error in the most recent census, indicating that, although census numbers have been used to determine past allocations, 2020 census error will not affect subsequent allocations. For a few programs, the role of census numbers was limited to affecting sampling weights for estimates of inflation and unemployment. We multiplied the estimated sum of absolute errors by 10 to approximate the effect for the ten-year period until the 2030 census, using the simple but perhaps unrealistic assumption that the legislated allocation programs now in existence will continue in their current form and current funding levels for 2021 – 2030, and no new allocation programs dependent on census results will be introduced. The weighted estimates of the expected absolute errors in federal fund allocations to states for ten years are shown in Figure 2.



Figure 2. Expected sum of absolute errors in federal fund allocations to states over 10 years, as related to census accuracy.

2. Methods

2.1. Overview

Similar but not identical methods are used to estimate the effects of census accuracy on apportionment and fund allocation. For the former, the true apportionment occurs when the 2020 census numbers for states have zero error. Let **P** and $\hat{\mathbf{P}}$ denote the vectors of true and estimated population sizes of states for the census year, 2020. The future population sizes **P** of states are unknown, and our beliefs about them can be summarized by a prior distribution, developed on the basis of population forecasts. This is described in Appendix 4. From specifications of the accuracy profiles (Table 1), and conditional on the true population sizes **P**, we can derive the probability distribution census estimates, $\hat{\mathbf{P}}$. House apportionment is determined purely by the population sizes, and the distribution of the errors in apportionment, i.e., apportionment based on $\hat{\mathbf{P}}$ minus apportionment based on **P**, can be derived.

6

Unlike apportionment, which depends only on state population sizes in 2020, formula-based allocations of funds depend on a wide variety of population statistics and other statistics. It would be very complex to jointly forecast the values of all such statistics ahead to 2020, and the results would likely be uncertain. Therefore, we took the simpler approach of obtaining the latest values we could of the statistics used to calculate allocations for the 18 programs we studied, and treating those as if they were error-free. To allow for error in the census, we used the accuracy profiles (Table 1) to first develop a distribution of census error, and we developed a distribution of error in the population statistics used to compute the allocations.

2.2. Apportionment

Since 1941, House apportionment has been determined from census populations using the Method of Equal Proportions (also known as Hill's Method and Huntington's Method). If fractions of seats could be allocated, then state *j* could simply receive its quota, q_j , defined as the number of seats *h* in the House of Representatives (currently h = 435) times the fraction of all 50 states' census population held by state *j*. Letting p_j denote the population of state *j* and *p* denote the population of all 50 states, we have $q_j = (p_j / p) \times h$. The allocations a_j of seats to state *j* must be whole numbers, however – no fractional allocations are allowed. The Method of Equal Proportions chooses positive integers a_j that minimize $\sum_i a_j (p_j / a_j - p / h)^2$ when the quotas q_j are given (Balinski and Young 1982, 1980, 1975; Spencer 1985). The apportionments are computed by the Census Bureau and provided to the President, who transmits them to Congress. Computation of apportionment is described by Balinski and Young (1977) and by the Census Bureau at <u>www.census.gov/population/apportionment/about/computing.html</u>. The Method of Equal Proportions is computed in practice by first awarding the first fifty seats one to each state. Seats 51 to 435 are awarded iteratively, each one to the state with the largest value of $p_i / \sqrt{n_i(n_i - 1)}$, where p_i is the census population of state *i* and n_i is the number of seats already awarded.

The sensitivity of the apportionment to census accuracy depends in part on the values of the underlying true populations of the states. The requirement that the numbers of seats held by states must be integers implies that, for some configurations of states' populations, a change of just a single person can cause the numbers to shift (Keyfitz 1979). For such configurations, even the smallest errors in census numbers will shift the allocations of seats. To analyze how sensitive apportionments are to changes in

census quality, we considered a joint distribution for the true state populations and the census numbers for states. To formulate the joint distribution, it is sufficient to consider the distribution for the true population sizes and the conditional distribution for census error given the true population. The error distribution was specified using an accuracy profile from Table 1, with alternative levels of average relative RMSE. The mean of the distribution of true population sizes was set equal to population projections prepared by the Weldon Cooper Center (2013a, 2013b). The true population sizes were taken to be independent, and the coefficient of variation for each state's population was chosen to be consistent with the past level of error in state population forecasts with similar time horizon. See Appendix 4 for details.

We estimated the distribution of sums of absolute errors in states' allocations of seats by drawing population numbers (true and census numbers) from the joint distribution, as described above. Then the true apportionments a_j^{true} and estimated apportionments a_j^{est} were calculated for each state j, and the sum of absolute misallocations (1) was calculated. This process was repeated 5,000 times independently, and the average sum of absolute misallocations was used to estimate the expected number of malapportioned seats.

2.3. Allocation of federal funds

As we have mentioned, Blumerman and Vidal (2009) identified 140 federal grant and direct assistance programs that distributed approximately \$446.4 billion in funds in FY 2007 at least partly on the basis of population and income data from the U.S. Census Bureau. The largest of these is the Medical Assistance Program, also known as Medicaid. Grants to states are equal to state medical expenditures times the Federal Medical Assistance Percentage (FMAP). The FMAP depends on per capita income, which is calculated as the ratio of census population to Bureau of Economic Analysis (BEA) personal income. The formula can be written as

FMAP = min
$$\left\{ max \left[1 - 0.45 \left(\frac{I_i / I_+}{P_i / P_+} \right)^2, 0.50 \right], 0.83 \right\},$$
 (2)

where I_i is the BEA personal income, P_i is the census population of state i, $I_+ = \sum_j I_j$, and $P_+ = \sum_j P_j$.

For analysis of total misallocated funds across all 140 programs, we selected the 8 largest programs (in terms of FY 2007 obligated amount) with certainty, which accounted for about 80.1% of the total FY

2007 obligations. From the remaining 132 programs, we selected a disproportionate stratified sample of 10 so that larger programs had a higher chance of selection. The programs we selected are shown in Table 2. We used sample-weighting methods to get unbiased estimates of totals for all 140 programs. For each program, results are weighted by the ratio of N_h , the number of total programs in stratum h, to n_h , the number of programs sampled in the stratum. We formed the weighted sum of estimates for sampled programs to estimate the total for all 140 programs. Thus, in total, the 18 programs represent \$445.6 billion in funds distributed in 2007 on the basis of population and income data from the U.S. Census Bureau. Note that this number is slightly different from the \$446.4 in funds actually distributed in FY 2007 due to the random sampling of programs. For each program studied, we estimated the expected annual amount of absolute errors in state allocations that would arise from alternative levels of census data quality. Sampling errors and approximate 95% confidence intervals were also estimated using theory for stratified samples. Table 2 shows the sampled programs.

					EV 2007	Weighted
					FY 2007	FY 2007
Strat			CFDA		Obligation	Obligation
h	N _h	n _h	No.	Program	(\$Billions)	(\$Billions)
			93.778	Medical Assistance Program (Medicaid)	\$203.5	\$203.5
			17.225	Unemployment Insurance	\$35.9	\$35.9
			20.205	Highway Planning and Construction	\$34.2	\$34.2
			10.551	Suppl. Nutrition Assistance Program (SNAP)	\$30.3	\$30.3
			93.558	Temporary Assist. for Needy Families (TANF)	\$16.5	\$16.5
			84.063	Federal Pell Grant Program	\$13.7	\$13.7
			84.010	Title I Grants to Local Educ. Agencies (LEAs)	\$12.8	\$12.8
1	8	8	84.027	Special Education Grants to States	\$10.8	\$10.8
			93.600	Head Start	\$6.9	\$10.3
2	3	2	93.767	State Children's Insurance Program (CHIP)	\$5.9	\$8.9
			10.557	Special Supplemental Nutrition Program for	\$5.5	\$16.6
				Women, Infants, and Children (WIC)		
3	6	2	93.596	Child Care Mandatory and Matching Funds	\$2.9	\$8.7
			93.575	Child Care and Development Block Grant	\$2.1	\$12.3
4	12	2	93.667	Social Services Block Grant	\$1.7	\$10.2
			84.365	English Language Acquisition Grants	\$0.6	\$4.9
5	16	2	84.181	Special Ed. – Grants for Infants and Families	\$0.4	\$3.5
			66.460	Nonpoint Source Implementation Grants	\$0.2	\$9.5
6	95	2	16.458	Title V Delinquency Prevention Program	\$0.1	\$3.0
Total	140	18				\$445.6

Table 2. Sampled programs allocating federal funds.

Table 3 presents the variety of the kinds of statistics used to allocate funds across the 18 sampled programs. Annual mid-year population estimates from the Population Estimate Program are used in 9 of the 18 programs. Two programs use model-based estimates for small-area populations that include Census Bureau population data in the models. Title I Grants to Local Education Agencies uses Small Area Income and Population Estimates for school district school-age children in poverty. The Supplemental Nutrition and Assistance Program for Women, Infants and Children uses a model-based estimate of the number of children age 1 to 4 below 185% of the poverty line. Two programs use American Community Survey (ACS) estimates. Special Education Grants to State uses information on state Free Appropriate Public Education age children in poverty from ACS Public Use Microdata. English Language Acquisition Grants uses ACS data on Limited English Proficiency children and foreign-born children. Current Population Survey (CPS) unemployment rates help determine whether states are eligible for additional Unemployment Insurance assistance. The CPS uses decennial census information for its sampling frame. Three programs, Supplemental Nutrition and Assistance Program, Pell Grants and Head Start, all make awards based on poverty thresholds. The poverty thresholds are estimated using the Consumer Price Index for all Urban Workers (CPI-U), which is estimated in part with a sampling frame that uses the decennial census.

						Non-	Latest
		Model-		CPS		Census	Census
	Midyear	based	ACS	Unempl.	CPI-	Stats	Not
Program (Dept.)	Pop. Est.	Pop. Est.	Pop. Est.	Rate	Urban	Used	Used
Medicaid	•					•	
СНІР	•					•	
Child Care Mandatory & Matching	٠						
Child Care and Development	٠					•	
Social Services Block Grant	•						
Special Ed. –Infants & Families	•						
Title V Delinquency Prevention	٠						
Title I Grants to LEAs	٠	٠				•	
Special Ed. – States	•		٠				
WIC		٠					
English Language Acquisition			•				
Unemployment Insurance				٠			
SNAP					٠		
Pell Grants					•	•	
Head Start					•		
Highways							•
TANF							•
Nonpoint Source Implementation							•

Table 3. Statistics used in formulas for allocating federal funds.

Many of the programs that we study use multiple census-based statistics. Further, five programs also use non-census statistics in formula-based allocation. For example, Medicaid awards use both census population numbers and BEA personal income. Surprisingly, we found that for 3 of the 18 selected programs, the allocations would not be affected by error in the most recent census: Highway Planning and Construction, Temporary Assistance for Needy Families, and Nonpoint Source Implementation Grants. These three programs have used census data for past allocations, but future allocations are fixed to previous shares. Several analytic simplifications were necessary for analyzing the effect of census error on the allocations. Except as noted, the simplifications were chosen to have the effect of overstating the effect of census error on error in allocation.

- (i) Unlike apportionment, which depends only on census population, the fund allocation programs involve other statistics in addition to census population. To fully model the diverse sources of error is too vast an undertaking for this project. Spencer (1980a, 67-150) demonstrates the kind of investigations that would be needed. For example, BEA personal income is used in multiple allocation formulas, but its accuracy is unknown. We use an approximation that conditions on the observed values of the non-census statistics. If we represent the allocation to a state by f(x,y), where y denotes the census estimates and x denotes other statistics, then the expected absolute misallocation may be expressed as $E | f(x,y) f(x^{true}, y^{true}) |$, where x^{true} and y^{true} denote the true values of x and y. We approximate this by $E | f(x,y) f(x, y^{true}) |$, conditioning on the observed values of x. Work in progress suggests that the approximation overstates the effect of census error in some general scenarios and that the potential understatement is smaller than the potential overstatement.
- (ii) For cases where population enters the allocation formula as a mid-year population estimate, which adjusts the census estimate for births, deaths and net migration since the census, we approximated the relative error in the postcensal estimate by the relative error in the underlying base census number. This approximation overstates the effect of census error on the postcensal estimate, since the errors in estimates of change due to births, deaths, and net migration are only somewhat dependent on the census base (Spencer 1980b). Specifically, the relative effect of census error on the census base overstates the relative effect of census error on the sum of the census base and other components only somewhat affected by census error.
- (iii) Model-based and ACS population estimates are used to calculate the proportion of the population in a group or area. The proportion is multiplied by a census or postcensal estimate of total population to estimate the number in the group or area. Here too, we approximated the relative error in the model-based or ACS estimate of population of the subgroup by the relative error in the underlying base census number. Since the errors in model-based and ACS estimates of fractions are largely independent of the census base, the effect of census error on the census base approximates the effect of census error on the product of the census base and the modelbased or ACS estimate of the population proportion.

12

- (iv) To model the effect of census error on CPS unemployment rates, we used differential net undercount estimates by age, sex and race in 2010 and applied these to unemployment estimates for these three groups to study the relationship between census error and unemployment rate error. We made the simplifying assumption that the effect of undercount by age, race and sex on unemployment rate estimates is proportional to the effect of state census errors on unemployment rate estimates. For CPI-U, we proceeded similarly using differential price indices for renters and owners together with information on renter and owner net census undercount. More details about the procedures are described in Appendix 5.
- (v) Title I Grants to LEAs provide grants to sub-state areas, namely school districts. We take the simplifying approach of studying errors in allocation at the state-level alone. Our models apply the state relative errors to each LEA population estimate within the state, which we conjecture slightly to some extent understates the effect of census error on the LEA-level Title I allocations
- (vi) For programs that depend upon multiple census-based statistics, we assume the same relative errors apply to all statistics, which overstates the effect of census error.

We estimated the expected sum of absolute errors in allocations for the year for which the most recent data was available. In order to obtain estimates corresponding to FY 2007, we ratio-adjusted the estimates of sum of absolute errors by the ratio of the FY 2007 program obligations to the allocations for the year for which allocations were analyzed. Typically, this was a downward adjustment. We conducted 5,000 independent simulations of census numbers and found absolute errors for each federal program analyzed. As previously described, we used sampling theory to estimate the total expected misallocated funds for FY 2007 for all 140 programs.

3. Results

3.1. Apportionment

Table 4 presents estimates of the expected number of malapportioned House seats, or House seats going to the wrong state, under the alternative accuracy profiles. Figure 1 is based on Table 4. Table 5 shows that the probability distribution of number of malapportioned seats for the base case accuracy profile with relative RMSE equal to 1%. The number of malapportioned seats can substantially exceed the expected number.

	Estimated Expected Number of Malapportioned Seats					
	Average	Relative	RMSE of S	tate Popu	lation Nu	mbers
Accuracy Profile	0.0%	0.5%	1.0%	2.0%	3.0%	4.0%
Base Case	0.00	1.79	3.38	6.66	10.00	13.32
Correlated Case	0.00	1.32	2.46	4.74	7.11	9.33
Accurate Small States Case	0.00	1.88	3.59	7.03	10.56	14.01
Differential Bias Case	0.00	1.59	2.96	5.70	8.51	11.44

Table 4. Estimated expected number of malapportioned seats in the U.S. House, with various census accuracy profiles. (Estimated standard errors for all numbers do not exceed 0.05.)

Relative RMSE	Prob	Probability that number of misallocated seats equals or exceeds k						
of census numbers	<i>k</i> = 2	<i>k</i> = 4	<i>k</i> = 6	<i>k</i> = 8	<i>k</i> = 10	<i>k</i> = 12		
0.5%	.714 (.006)	.166 (.005)	.014 (.002)	.001 (.000)	.001 (.000)	.001 (.000)		
1.0%	.934 (.004)	.561 (.007)	.168 (.005)	.024 (.002)	.002 (.001)	()		
2.0%	.998 (.001)	.956 (.003)	.760 (.006)	.427 (.007)	.147 (.005)	.034 (.003)		
3.0%	1.000 ()	.998 (.001)	.975 (.002)	.867 (.005)	.640 (.007)	.339 (.007)		
4.0%	1.000 ()	1.000 ()	.998 (.001)	.983 (.002)	.914 (.004)	.751 (.006)		

Table 5. Probability distribution of number of House seats misallocated, base case census accuracyprofile. -- signifies number < 0.05%. Number in parentheses is estimated standard error of probability.</td>

3.2. Formula-based allocations of federal funds

Estimates of the expected sum of absolute errors in total FY 2007 federal funds allocations are shown in Table 6. We used the latest available values of x and y in estimating the expected sum of absolute errors in allocations. Then, we ratio-adjusted the estimates of $E|f(x,y)-f(x,y^{true})|$ by the ratio of the FY 2007 program obligations to the allocations for the year whose allocations based on x and y were analyzed. Typically, this was a downward adjustment.

	Estimated Expected Misallocated Funds in One Year					
		(\$ I	Billions)			
	Average	Relative RMS	E of State Pop	ulation Numb	ers	
Accuracy Profile	0.5%	1.0%	2.0%	3.0%	4.0%	
Base Case	1.02	2.05	4.10	6.09	8.08	
(standard error)	(0.04)	(0.07)	(0.13)	(0.19)	(0.25)	
Correlated Case	0.75	1.50	3.01	4.51	5.99	
(standard error)	(0.03)	(0.05)	(0.09)	(0.14)	(0.18)	
Accurate Small						
States Case	1.02	2.05	4.05	6.11	8.09	
(standard error)	(0.04)	(0.07)	(0.14)	(0.20)	(0.27)	
Differential Bias						
Case	0.92	1.83	3.66	5.45	7.25	
(standard error)	(0.04)	(0.06)	(0.11)	(0.16)	(0.22)	

Table 6. Estimated expected absolute misallocations of federal funds in one year, with various census accuracy profiles.

4. Sensitivity Analysis

The base case accuracy profile specifies that census numbers are independent, normally distributed, unbiased and have equal coefficients of variation across all states. Work in progress will estimate a response surface and will provide estimates of expected amounts of misallocated funds and malapportionment as a function of parameters describing the distributions of census numbers. The

response function will account for means and variances of state census numbers, the correlation between state estimates, and whether or not distributions have heavy tails (*t* distribution as compared to normal distribution). This will allow the assessment of the sensitivity of distortions in allocations to a variety of distributional aspects of census error.

Here, we examine various assumptions in the base case one at a time and study how results change for alternative distributions of census error. For each analysis, we study the impact on one example of formula-based allocation, Medicaid for FY 2012, and on apportionment. Attention is restricted to small relative RMSE levels.

4.1. Correlated errors

To study the sensitivity of findings to the assumption of independently distributed census numbers across states, we considered scenarios where state relative errors were multivariate normally distributed with a constant correlation coefficient. We examined constant correlations of -0.02, 0.00 (corresponding to independence), 0.25, 0.50, 0.75 and 0.90. Note that the smallest possible constant correlation for an $n \times n$ non-negative definite correlation matrix is -1/(n-1). When n is 51 (for 50 states and D.C.), then the minimum possible constant correlation is -0.02. All scenarios are constructed so that the average relative RMSE is 1.0%.

Figure 3 presents results for Medicaid, and Figure 4.2 presents results for apportionment. Both figures demonstrate that sums of absolute errors in allocation and in apportionment decrease sharply with increased correlation of state relative errors. For Medicaid, \$1.6 billion in funds is misallocated on average with independent errors, but when relative errors have a constant correlation of 0.9, only \$0.5 billion is misallocated on average. Similarly, 3.4 seats are expected to be malapportioned with independent errors, but only 1.2 seats are expected to be malapportioned when relative errors have a constant correlation of 0.9.

16



Figure 3. Expected sum of absolute errors in Medicaid allocations FY 2012, as related to the correlation between state relative errors.



Figure 4. Expected sum of malapportioned seats in the U.S. House of Representatives, as related to the correlation between state relative errors.

We believe the reason for this relationship is that when relative errors are highly correlated, there is less variation in the census estimates of state population shares. The population shares are the quantities truly of interest for distributing Medicaid funds or for apportioning the House of Representatives. Thus, we find that misallocation and malapportionment decrease with the correlation in relative errors for state census numbers.

It is also possible that the findings understate the effects of census error if state errors are negatively correlated. However, the understatement would likely be small due to bounds on negative correlations required for a non-negative definite correlation matrix.

4.2. Heavy-tailed distributions

Our main analyses only consider normally distributed state census numbers. For the 1.0% relative RMSE scenario, if a state has population p, then the state errors are normally distributed with mean 0 and standard deviation 0.01 p. We considered how results would change for distributions with heavier tails. In particular, we examined the t distribution with 4 degrees of freedom. In order to compare distributions with the same variances, the state errors are distributed $t_4(.01p/\sqrt{2})$, as the t_4 distribution has a standard deviation of $\sqrt{2}$.

Results are presented in Table 7. For Medicaid, expected misallocated funds change from \$1.56 billion in the scenario with normally distributed census numbers to \$1.40 billion in the scenario using the scaled t_4 distribution. Similarly, expected malapportionment changes from 3.41 seats in the scenario with normally distributed census numbers to 3.09 seats in the scenario using the scaled t_4 distribution. From this, we conclude that using normally distributed census numbers when the distributions have heavy tails leads to overstating the effects of census error on uses.

Distribution	Medicaid (\$ Bill.)	Apportionment
Normal	1.56	3.41
Scaled t_4	1.40	3.09
Percentage Difference	-10.7%	-9.6%

Table 7. Estimated expected sum of absolute errors in allocation and apportionment with differentdistributions of census error

4.3. Bias in state estimates

We investigated evidence for bias in census population numbers by examining estimates of net undercount rates for the 2000 and 2010 censuses.¹ If bias exists, then we expect to see that the 2000 net undercount estimates for each state predict their 2010 net undercount estimates well. Figure 5 presents the net undercount estimates in a scatterplot with the diameter of each circle proportional to the state's population. The seven largest states are labeled. The relationship is stronger for large states

¹ Data from <u>www.census.gov/dmd/www/pdf/pp-60r.pdf</u> and <u>www.census.gov/coverage_measurement/pdfs/g01.pdf</u>, accessed February 12, 2015

than for small states. While the correlation between the 2000 and 2010 net undercount rates is 0.19, the correlation weighted by the population is 0.54. This suggests more evidence for bias in census population numbers for large states than for small states.



2000 and 2010 Census Net Undercount Rates

Figure 5. Comparison of estimated 2000 and 2010 census net undercount rates by state. *Diameter of each circle proportional to state population.*

To study the sensitivity of our findings to bias in state census numbers, we considered models where the largest 25 states could have biases, but the smallest 25 states did not. For all scenarios, mean squared error is the same and equivalent to the mean squared error for the 1.0% c.v. scenario. As mean squared error can be decomposed into bias squared and variance, we varied the weight on bias squared across the scenarios. We examined weights on bias squared for the 25 largest states of 0%, 10%, 25%, 50%, 75%, 90% and 100%. The 0% weight on bias squared scenario corresponds to the 1.0% c.v. scenario used in our main analysis. We chose the signs of the state biases to have the same signs as the biases estimated for the 2010 census.

Results are presented for Medicaid in Figure 6 and for apportionment in Figure 7. We do not find much sensitivity of either Medicaid misallocation or malapportionment to bias in large states when the weight

on bias squared varies between 0% and 50%. In this range, we expect between \$1.5 and \$1.6 billion misallocated and between 3.4 and 3.5 seats malapportioned. However, error in uses increases with larger amounts of bias relative to variance. When the errors are only due to bias and not variance, we expect \$1.8 billion in funds to be misallocated and 3.9 House seats to be malapportioned. These analyses show that modeling census error with only variance and not bias could lead to understating the effects of census error, but the understatement will only be large if bias is large relative to variance.



Figure 6. Expected sum of absolute errors in Medicaid allocations FY 2012, as related to the fraction of mean square error for 25 largest states due to bias squared.



Figure 7. Expected sum of malapportioned seats in the U.S. House of Representatives, as related to the fraction of mean square error for 25 largest states due to bias squared.

4.4. Unequal coefficients of variation among states

The main analyses use equal coefficients of variation across all states. We investigated the sensitivity of findings to differing coefficients of variation across states. We considered five sets of scenarios. In all scenarios, ten states were assigned a c.v. of 5.0% and forty states were assigned a 0.0% c.v. Thus, in all scenarios, the average c.v. is 1.0%. We compare scenarios where the ten largest states have the high

c.v., the 11th through 20th largest states have the high c.v., the 21st through 30th largest states have the high c.v., the 31st through 40th largest states have the high c.v. and where the ten smallest states have the high c.v.

Results are presented in Figures 4.6 and 4.7. Findings should be compared to results from the constant 1.0% c.v. scenario, where we expect \$1.6 billion in funds to be misallocated for Medicaid in FY 2012 and 3.4 House seats to be malapportioned. When the ten largest states have the high c.v., sums of absolute errors are very large. We expect \$4.9 billion in funds to be distributed to the wrong state and 10.5 seats to be apportioned to the wrong state. On the other hand, when the ten smallest states have the high c.v., we expect only \$0.3 billion to be misallocated for Medicaid and 1.1 House seats to be malapportioned. This investigation shows that results are highly sensitive to whether large or small states have high c.v.'s.



Figure 8. Expected sum of absolute errors in Medicaid allocations FY 2012, for scenarios where 40 states have no error and 10 states designated have 5% relative RMSE.



Figure 9. Expected sum of malapportioned seats in the U.S. House of Representatives, for scenarios where 40 states have no error and 10 states designated have 5% relative RMSE.

5. Conclusions

Figures 1 and 2 show how that census inaccuracy can have a large effect on apportionment of the U.S. House of Representatives and on the distribution of more 4 trillion dollars in federal funds over the decade. Our analysis suggests that, if the average percentage error in census numbers for states is 4%, the expected number of House seats going to the wrong state (relative to having perfect data) is in the range of 10 to 14, and between \$60 billion and \$80 billion dollars in federal grants in aid will go to or from the wrong states.

It is important to understand the limitations of the scope of the current analysis. In this paper we study only direct and specific "instrumental" uses census statistics for allocating funds and House seats. We have not studied the effect of census quality on "conceptual" uses of census data for scientific research or for policy-making, uses which are vastly more difficult to identify and describe (Beyer 1977). Part of the reason conceptual uses resist study is that they are hidden in chains of analysis. For example, policy X is adopted or theory Y is accepted on the basis of cited research that depended in part on supporting research that depended on past census data, but the role of the census is not apparent. For another example, former OMB Director OMB Director Peter Orszag (2009, 40) noted that the educational policy goal of increasing the number of postsecondary education was developed to reduce social inequality, based on empirical research of Goldin and Katz (2008, 2007) that relied in key ways on decennial census data from 1940-1980 and on Iowa State Census data from 1915. Not only is it difficult to identify such uses of census data after they have occurred, but it is even more difficult to anticipate them ahead of time. As noted by J. G. March (1994, 246),

Having knowledge when it is needed often requires an investment in knowledge that is not known to be needed at the time it is acquired. The returns from knowledge may occur in a part of the system quite different from the part where the costs are paid.

Another kind of use statistics is for window dressing, or "using research results to legitimate and sustain predetermined positions." (Beyer 1977, 17) Symbolic uses of data can be sensitive to data quality, as explained by Boruch (1984) among others. Suppose that a decision maker simply wants to use data as window dressing to defend a decision already made. If the data are high quality, then they will more accurately describe the true state. If the decision maker needs false information to justify the decision, which by itself raises questions about the validity of the decision, then that will be more difficult with high quality data.

Uniquely, the constitutionally mandated census is also important as a national ceremony (Kruskal 1984, 49-50).

The decennial census is a national ceremony and a symbol of the relationship between citizen and government. Whatever one's view of the census, whatever one's philosophical

26

position about the Federal Government, it may be argued that the census is one of our relatively few national, secular ceremonies. It provides a sense of social cohesion, and a kind of nonreligious communion: we enter the census apparatus as individual identities with a handful of characteristics; then later we receive from the census a group snapshot of ourselves at the ceremony date. Like many family pictures, the snapshot is a little blurry in spots, but recognizable and fascinating to compare across the decades...

Kruskal's comment provides background for understanding the observation made 4 decades ago by Representative William Lehman (Congress 1973, p.2812):

... confidence of the people in the census data is transferred to the confidence of the people in their political process.

In addition, the census is used to adjust or calibrate the results of virtually all national sample surveys of the U.S. population in the public and private sectors.

In conclusion, the effects of inaccuracy in the 2020 census on apportionment and allocation of federal funds are appreciable, but there are also important effects of census accuracy.

References

Alho, J. M. and Spencer, B. D. (2005) Statistical Demography and Forecasting. New York: Springer.

- Anderson, M. J. and Fienberg, S. E. (1999) *Who Counts? The Politics of Census-Taking in Contemporary America.* New York: Russell Sage Foundation.
- Balinski, M. and H. Young (1975) The quota method of apportionment, *American Mathematical Monthly* 82, 701-730.
- Balinski, M. L. and Young, H. P. (1977) On Huntington methods of apportionment. *Siam Journal on Applied Mathematics*, 33, 607-618.
- Balinski, M. L. and Young, H. P. (1982) *Fair Representation: Meeting the Ideal of One Man, One Vote*. New Haven: Yale University Press.
- Beyer, J. M (1977) Research utilization: bridging the gap between communities. Journal of Management Inquiry 6, 17-22.
- Blumerman, L. M. and Vidal, P. M. (2009) Uses of Population and Income Statistics in Federal Funds
 Distribution With a Focus on Census Bureau Data. Governments Division Report Series,
 Research Report #2009-1. Washington, D.C.: U.S. Census Bureau.
- Boruch, R. F. (1984) Research on the use of statistical data. *Proceedings of the Social Statistics Section, American Statistical Association*, 52-57.
- Census Bureau (2012) Press release for May 22, 2012.

http://www.census.gov/newsroom/releases/archives/2010 census/cb12-95.html

Christy, J. (2012) Census Update, 2012 Annual State Data Center Network Meeting, May 30-31, 2012, Sacramento, CA.

http://www.dof.ca.gov/research/demographic/state_census_data_center/meetings/documents /CASDC AnnualMtg2012 Christy CensusUpdate.pdf

- Cohen, M. L. (1990) Adjustment and reapportionment analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.
- Congress, U.S. (1973) Bills to Extend and Amend the Elementary and Secondary Education Act of 1965, and for Other Purposes. Hearings before the House Committee on Education and Labor, 93rd Congress, 1st Session, 3 May 1973.
- GAO (2012a) *2020 Census: Additional Steps are Needed to Build on Early Planning*. Report GAO-12-626. Washington, D.C.: U.S. Government Accountability Office.
- GAO (2012b) 2020 Census: Sustaining Current Reform Efforts Will Be Key to a More Cost-Effective Enumeration. Statement of Robert Goldenkoff, Director, Strategic Issues, GAO. Testimony

Before the Subcommittee on Federal Financial Management, Government Information, Federal Services, and International Security, Committee on Homeland Security and Governmental Affairs, U.S. Senate. July 18, 2012. GAO-12-905T, Washington, D.C.: U.S. Government Accountability Office.

- Gilford, L., Causey, B. D., and Rothwell, N. D. (1982) How adjusting census counts could affect congress, *American Demographics*, 4, 30-44.
- Goldin, C. and Katz, L. F. (2007) The race between education and technology: the evolution of U.S.
 educational wage differentials, 1890 to 2005. NBER Working Paper No. 12984. Cambridge:
 National Bureau of Economic Research.
- Goldin, C. and Katz, L. F. (2008) *The race between education and technology*. Cambridge: Harvard University Press.
- Kadane, J. B. (1986) A Bayesian approach to designing U.S. census sampling for reapportionment, *Journal of Official Statistics*, 12, 85-106 (with discussion).
- Keyfitz, N. (1979) Information and allocation: two uses of the 1980 census, *The American Statistician*, 33, 45-50.
- Kruskal, W. H. (1984) Research and the census. In *Federal Statistics and National Needs* prepared for the Subcommittee on Energy, Nuclear Proliferation and Government Processes, an arm of the Committee on Government Affairs of the United State Senate, by the Congressional Research Service of the Library of Congress, 49-57.
- March, J. G. (1994) A Primer on Decision Making: How Decisions Happen. New York: The Free Press.
- Melnick, D. (2002) The legislative process and the use of indicators in formula allocations. *Journal of Official Statistics* 18, 353–369.
- National Research Council (1980) *Estimating Population and Income of Small Areas*. Panel on Small-Area Estimates of Population and Income, Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D.C.: National Academy Press.
- National Research Council (2010) *Envisioning the 2020 Census*. Panel on the Design of the 2010 Census
 Program of Evaluations and Experiments, L. D. Brown, M. L. Cohen, D. L. Cork, and C. F. Citro,
 eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education.
 Washington, DC: The National Academies Press.
- National Research Council (2011) Change and the 2020 Census: Not Whether But How. Panel to Review the 2010 Census, T. M. Cook, J. L. Norwood, and D. L. Cork, eds., Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The

National Academies Press.

- Orszag, P. R. (2009) Federal statistics in the policy process. *The Annals of The American Academy of Political and Social Science* **631**, 34-42.
- Savage, I. R. (1980) Modifying census counts. Pp. 62-75 in U.S. Bureau of the Census. Conference on Census Undercount: Proceedings of the 1980 Conference. Washington, D.C.: U.S. Department of Commerce.
- Schirm, A. L. (1991) The effects of census undercount adjustment on congressional apportionment. Journal of the American Statistical Association, 86, 526-541
- Spencer, B. D. (1980a) *Benefit-Cost Analysis of Data Used to Allocate Funds: General-Revenue Sharing.* New York: Springer.
- Spencer, B. D. (1980b) Models for error in postcensal population estimates. Pp. 217-228 in National Research Council, *Estimating Population and Income of Small Areas*. Report of Panel on Small-Area Estimates of Population and Income. Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D. C.: The National Academies Press.
- Spencer, B. D. (1985) Statistical aspects of equitable apportionment. *Journal of the American Statistical Association*, 80, 815-822.

Wald, A. (1947) Sequential Analysis. New York: Wiley.

Weldon Cooper Center for Public Service (2013a) National and State Population Projections.

http://www.coopercenter.org/demographics/virginia-population-projections. Demographic Research Group, University of Virginia, Charlottesville.

Weldon Cooper Center for Public Service (2013b) *State and National Projections Methodology*. <u>http://www.coopercenter.org/sites/default/files/node/13/National_Projections_Methodology</u>. <u>pdf</u>. Demographic Research Group, University of Virginia, Charlottesville.

Appendix 1: Approximation of Expected Absolute Errors in Allocation in the Presence of Errors in Non-Census Statistics

We are interested in the difference in expected sums of absolute values of misallocations when the census is conducted under one design versus another. A difficulty in studying the effect of census accuracy on allocations of federal funds is the dependence of allocations on non-census statistics as well as census population statistics. Sometimes there are too many different non-census statistics to want to model the accuracy for all of them, and other times, as in the case of BEA personal income, there is not even systematic information available about accuracy (aside from the extent of revisions). It is convenient, then, to analyze the effects of census inaccuracy conditional on the observed values of the non-census statistics. In this appendix we consider the implications of relying on this conditional analysis.

Let f denote an allocation formula, and let the allocation to a state be denoted by f(X,Y), where Y is a census statistic measuring true value Y^* and X is a non-census statistic measuring true value X^* . The difference in expected absolute value of the error in allocation is

$$E_{1}|f(X,Y) - f(X^{*},Y^{*})| - E_{2}|f(X,Y) - f(X^{*},Y^{*})|$$
(1.1)

where E_i denotes expectation when census design *i* is used. The expectation E_i in (1.1) also takes account of the distribution of *X*. To avoid working with the distribution of *X*, we condition on the observed value *x* of the non-census statistic. Thus, we study

$$E_{1|X=x} \left| f(x,Y) - f(x,Y^*) \right| - E_{2|X=x} \left| f(x,Y) - f(x,Y^*) \right|, \qquad (1.2)$$

where $E_{i|X=x}$ denotes the conditional expectation under census design *i* given the observed value *x* of the non-census statistic.

We conjecture that (1.2) tends to overstate (1.1), and that when it does not overstate then the magnitude of understatement is not severe. To support this conjecture, consider a linear approximation to the allocation, namely f(X,Y) = a + bX + cY. Assume that the true values are fixed at x^* and y^* and that given the true values the non-census statistic is normally distributed, unbiased, and with variance τ^2 , irrespective of which census design is used. Under census design *i*, the census statistic is

stochastically independent of the non-census statistic and it is normally distributed, unbiased, with variance σ_i^2 . Under these conditions, expression (1.2) evaluates to

$$\sqrt{\frac{2}{\pi}}|c|(\sigma_1 - \sigma_2). \tag{1.3}$$

Expression (1.1) is equivalent to

$$\sqrt{\frac{2}{\pi}} \left(\sqrt{b^2 \tau^2 + c^2 \sigma_1^2} - \sqrt{b^2 \tau^2 + c^2 \sigma_2^2} \right).$$
(1.4)

It is readily seen that $|c|(\sigma_1 - \sigma_2) \ge \sqrt{b^2 \tau^2 + c^2 \sigma_1^2} - \sqrt{b^2 \tau^2 + c^2 \sigma_2^2}$, with equality when b = 0 or $\tau = 0$. Thus, expression (1.3) overstates expression (1.4). The overstatement increases with |b| and with τ . Thus, in this model, conditioning on the value of the non-census statistic leads to overstatement of the true difference in expected absolute loss. Work is underway to compare (1.1) and (1.2) under more general models.

Appendix 2: Details on Sampling of Allocation Programs

The $N = 140$ allocation programs identified by Blumerman and Vidal (2009) were partitioned into H	=6
strata based on size of FY 2007 obligation, as shown in Table A2.1.	

	FY 2007 Obliga	tion (\$Billions)			Total FY 2007
Stratum			Number	Sample Size	Obligations
h	At least	Not	N _b	n _h	(\$Billions)
		exceeding			
1	\$10.780		8	8	\$357.6
2	\$5.940	\$7.837	3	2	\$20.6
3	\$2.592	\$5.478	6	2	\$24.2
4	\$1.190	\$2.494	12	2	\$22.8
5	\$0.436	\$1.163	16	2	\$11.3
6		\$0.432	95	2	\$10.0
All			140	18	\$446.4

Table A2.1. Stratification of fund allocation programs and allocation of sample.

The sampling strata are denoted by h h=1,...,H, with N_h programs in stratum h, h=1,...,H, from which n_h are selected with SRS without replacement. The total number of allocation programs in the sampling frame is $140 = N = \sum_{h} N_h$. Simple random sampling within strata was performed to choose n_h out of N_h programs without replacement. The sampling weight for each selected program in stratum h is the reciprocal of the sampling fraction, $w_h = N_h / n_h$. The sample may be represented by the collection of indicator variables I_{hi} taking the value 1 if program i in stratum h was selected and 0 if not selected. Alternative levels of census data quality are denoted by j. The expected error sum of absolute errors in in allocations from program i in stratum h attributable to census data quality level j is denoted by y_{hij} .

The purpose of the sample is to support inferences about the effect of census data quality level j on allocations to all N programs,

$$\boldsymbol{y}_{..j} = \sum_{h} \sum_{i=1}^{N_h} \boldsymbol{y}_{hij}.$$

If we were we able to observe y_{hij} directly for sampled programs, we would estimate $y_{..j}$ by the conventional unbiased estimator in stratified sampling, $\tilde{y}_{..j} = \sum_{h} \sum_{i=1}^{N_h} w_{hi} I_{hi} y_{hij}$. We do not observe y_{hij} directly for sampled programs, however, but instead we estimate it by simulation. Census data with quality level j are independently replicated R times, leading to simulated values y_{hijr} , r = 1, ..., R. The

estimate of $\tilde{y}_{..j}$ based on replicate r is denoted by $\tilde{y}_{..jr} = \sum_{h} \sum_{i=1}^{N_h} w_{hi} I_{hi} y_{hijr}$. We use the average of the replicated estimates, $\hat{y}_{..j} = R^{-1} \sum_{r=1}^{R} \tilde{y}_{..jr}$ to estimate $y_{..j}$.

There are two components of variance in $\hat{y}_{..j}$. One arises because allocation programs are sampled, and we will refer to this as random component 1. The second arises from the replications in the simulations, conditional on the set of programs that were sampled, and we refer to this as random component 2. The simulations are conditionally independent of the set of programs that were selected, and If the modeling of the effect of census data quality is perfect, then each $\tilde{y}_{..jr}$ is a conditionally unbiased estimate of $\tilde{y}_{..jr}$ given the selection of programs.

It follows from standard results that

$$E(\hat{y}_{..j}) = E_1 E_2 \left(R^{-1} \sum_{r=1}^{R} \tilde{y}_{..jr} | I_{11}, ..., I_{H,N_H} \right) = E_1 \left(\sum_{h=1}^{N_h} w_{hi} I_{hi} y_{hij} \right) = Y_{..j}.$$

Similarly,

$$V(\hat{y}_{..j}) = E_1 V_2 \Big(R^{-1} \sum_{r=1}^{R} \tilde{y}_{..jr} | I_{11}, ..., I_{H,N_H} \Big) + V_1 E_2 \Big(R^{-1} \sum_{r=1}^{R} \tilde{y}_{..jr} | I_{11}, ..., I_{H,N_H} \Big)$$
$$= E_1 \sigma_{2j}^2 + V_1 \Big(\sum_h \sum_{i=1}^{N_h} w_{hi} I_{hi} y_{hij} \Big)$$

with $\sigma_{2j}^2 = V_2 \left(R^{-1} \sum_{r=1}^{R} \tilde{y}_{..jr} | I_{11}, ..., I_{H,N_H} \right)$. A conditionally unbiased estimator of σ_{2j}^2 is $\hat{\sigma}_{2j}^2 = R^{-1} \sum_{r=1}^{R} (\tilde{y}_{..jr} - \hat{y}_{..j})^2 / (R-1)$, and we may also use $\hat{\sigma}_{2j}^2$ as an unbiased estimator of $E_1 \sigma_{2j}^2$. Additionally, $V_1 \left(\sum_h \sum_{i=1}^{N_h} w_{hi} I_{hi} y_{hij} \right) = \sum_{h=1}^{H} N_h^2 (1 - n_h / N_h) \sigma_{1hj}^2 / n_h$, with $\sigma_{1hj}^2 = \sum_{i=1}^{N_h} (y_{hij} - \overline{y}_{Uh,j})^2 / (N_h - 1)$ and $\overline{y}_{Uh,j} = \sum_{i=1}^{N_h} y_{hij} / N_h$. In estimating σ_{1hj}^2 we use replicate means $\overline{y}_{hij.} = R^{-1} \sum_{r=1}^{R} y_{hijr}$ to estimate y_{hij} . Then, we estimate σ_{1hj}^2 by $\hat{\sigma}_{1hj}^2 = \sum_{i=1}^{N_h} I_{hi} (\overline{y}_{hij.} - \overline{y}_{h.j.})^2 / (n_h - 1)$ with $\overline{y}_{h.j.} = \sum_{i=1}^{N_h} N_h^2 (n_h$. It is straightforward to show that $E\hat{\sigma}_{1hj}^2 > \sigma_{1hj}^2$ and $E\hat{\sigma}_{1hj}^2 - \sigma_{1hj}^2 = O(R^{-1})$, so the overestimation is negligible if the number of replications R is large enough. Thus, we estimate $V(\hat{y}_{..j})$ by

$$\hat{V}(\hat{y}_{..j}) = \hat{\sigma}_{2j}^{2} + \sum_{h=1}^{H} N_{h}^{2} (1 - n_{h} / N_{h}) \hat{\sigma}_{1hj}^{2} / n_{h}$$

Appendix 3: Use of Census Statistics in 18 Sampled Programs Using Population Statistics for Allocating Federal Funds

In this section, for statistic X_i in state *i*, let $\breve{X}_i = X_i / \sum_j X_j$ denote state *i*'s national share for the

statistic. The number in parentheses following the name of the allocation program is the Catalog of Federal Domestic Assistance (CFDA) number.

Medical Assistance Program (93.778)

The Medical Assistance Program pays a fraction of state medical expenditure as determined by the state's Federal Medical Assistance Percentage (FMAP), which is based on a state's relative per capita income. The FMAP has a minimum of 0.50 and a maximum of 0.83. The grant amount G_i for state *i* is defined as $G_i = E_i [\max(0.50,\min(0.83,1-0.45(\tilde{I}_i / \tilde{P}_i)^2))]$. Here, E_i is medical expenditures, I_i is BEA personal income, and P_i is population. Note that this formula does not apply to DC, which has a fixed FMAP of 0.70.

Unemployment Insurance (17.225)

When state unemployment rates are elevated, unemployment insurance recipients are eligible for extra weeks of compensation through Extended Benefits and, during the recent recession, Emergency Unemployment Compensation. The federal government provides part of the funding for these two programs. State unemployment rates are estimated by the Current Population Survey, which uses the decennial census for its sampling frame.

Highway Planning and Construction (20.205)

The Moving Ahead for Progress in the 21st Century Act passed in 2012 changed the funding formulas for programs administered through Highway Planning and Construction. Apportionment for the various programs is fixed to proportions states received in previous years and does not depend on new population statistics.

Supplemental Nutrition Assistance Program (10.551)

A recipient's eligibility for SNAP benefits and amount received are based on poverty threshold which are revised annually based on CPI-U. CPI-U is estimated in part using the decennial census for a sampling frame.

Temporary Assistance for Needy Families (93.558)

The grant amounts to states are fixed to the proportions of the grants in 2002. New census statistics are not used in the determination of grant amounts.

Federal Pell Grant Program (84.063)

A student's Pell Grant amount is the cost of attendance minus the Expected Family Contribution subject to a maximum. The formula to determine the Expected Family Contribution is revised each year based on measures of inflation. The inflation measures are estimated in part using the decennial census for a sampling frame.

Title I Grants to Local Educational Agencies (84.010)

Title I funding consists of four sets of grants to Local Educational Agencies (LEAs): Basic Grants, Concentration Grants, Targeted Grants and Education Finance Incentive Grants (EFIG). All four grant programs depend on Small Area Income and Poverty Estimates (SAIPE) data for the age 5 to 17 population in poverty for each LEA. In our modeling, we treat SAIPE LEA errors the same way that we treat the corresponding state-level errors in census counts.

The Basic Grant amount G_i^{Basic} for LEA *i* is specified by

$$G_{i}^{\text{Basic}} = N_{i} \Big[\min \Big(0.48 E_{US}, \max (0.32 E_{US}, 0.4 E_{i}) \Big) \Big] \chi_{\{N_{i} \ge 10\}} \chi_{\{N_{i}/T_{i} \ge 0.02\}}$$

Here, N_i is a measure of the number of children at-need in the LEA, including the SAIPE age 5—17 population in poverty; T_i is the total number of school-aged children; E_i is the per-pupil expenditures in the state that includes the LEA; E_{US} is the national per-pupil expenditure; $\chi_{\{A\}}$ is the indicator function taking the value 1 if A is true and 0 otherwise.

The Concentration Grant amount $G_i^{\text{Concentration}}$ for LEA *i* is specified by

$$G_i^{\text{Concentration}} = N_i \Big[\min \big(0.48 E_{US}, \max \big(0.32 E_{US}, 0.4 E_i \big) \big) \Big] \chi_{\{N_i \ge 6500 \text{ OR } N_i/T_i \ge 0.15\}}$$

with all variables having the same definitions as they do for Basic Grants.

Title I Grants to Local Educational Agencies (84.010) (continued)

The Targeted Grant amount G_i^{Targeted} for LEA *i* is specified by

$$G_{i}^{\text{Targeted}} = W_{i}N_{i}\left[\min(0.48E_{US},\max(0.32E_{US},0.4E_{i}))\right]\chi_{\{N_{i}\geq10\}}\chi_{\{N_{i}|T_{i}\geq0.05\}}$$

with all variables having the same definitions as they do for Basic Grants, except that W_i is a weight between 1.0 and 4.0 that increases with N_i and N_i / T_i and depends on county or school district administration.

The EFIG amount G_i^{EFIG} for LEA *i* is specified by

 $G_i^{\text{EFIG}} = N_i \left[\min(0.46E_{US}, \max(0.34E_{US}, 0.4E_i)) \right] \times \text{Effort}_i \times (1.3 - \text{Equity}_i) \chi_{\{N_i \ge 10\}} \chi_{\{N_i / T_i \ge 0.05\}},$

where all variables have the same definitions as they do for basic grants, except that Effort_i is the effort factor and Equity_i is the equity factor for the state to which LEA *i* belongs. The effort factor depends upon per capita personal income (BEA) and thus indirectly on census population statistics. Specifically, Effort_i = $(E_i / E_{US})(\vec{P}_i / \vec{I}_i)$, where E_i and E_{US} are as defined above and $\vec{P}_{state} / \vec{I}_{state}$ is algebraically equal to the ratio of US per capita income to state per capita income. The equity factor for the state Equity_i depends on a weighted coefficient of variation of LEA per-pupil expenditure within the state to which LEA *i* belongs, with the weighting depending on N_i and T_i .

All four grants are ratably reduced to sum to the total amount allocated for each grant program. Adjustments are made so each state receives a minimum amount for each of the four grants.

Special Education Grants to States (84.027)

Grants to states use measures of states' Free Appropriate Public Education (FAPE) age population (usually age 3-21 population) and states' FAPE age population in poverty. The FAPE age population is taken from postcensal population estimates by single year of age. The FAPE age population in poverty is determined by combining the postcensal single year of age data with ACS Public Use Microdata Sample poverty data. Specifically, state *i* receives grant amount G_i , where $G_i = G_i^{99} + R[0.85\breve{P}_i + 0.15\breve{N}_i]$, G_i^{99} is the grant amount in 1999, *R* is the total amount available for the program in excess of 1999 total amount, P_i is the FAPE age population, and N_i is the FAPE age population in poverty.

Head Start (93.600)

Head Start agencies must have a certain percentage of children they serve be from families who meet poverty thresholds. The poverty thresholds are revised based upon CPI-U, which is estimated in part using the decennial census for a sampling frame. Because some Head Start agencies are not fully enrolled, and agencies can respond to changes in poverty threshold eligibility by either increasing or decreasing their effort to recruit students, we estimate that 2020 census error will have no effect on Head Start funding.

State Children's Insurance Program (93.767)

The State Children's Insurance Program (CHIP) pays a fraction of state CHIP expenditure as determined by the state's Enhanced Federal Medical Assistance Percentage (eFMAP), which is based on a state's relative per capita income. The eFMAP has a minimum of 0.65 and a maximum of 0.85. The formula for the grant amount G_i for state *i* is $G_i = E_i \left[\max \left(0.65, \min[0.85, 1 - 0.315(\overline{I}_i / \overline{P}_i)^2] \right) \right]$, where E_i is medical expenditures, I_i is BEA personal income, and P_i is population. This formula does not apply to D.C., which has a fixed FMAP of 0.79.

Special Supplemental Nutrition Program for Women, Infants and Children (10.557)

WIC uses two grant programs, one for food costs and another for nutrition services and administrative (NSA) costs. Food grant amounts are proportional to a state's model-based estimate of the number of children age 1 to 4 below 185% of the poverty line. In our modeling, we treat the errors in model-based estimates the same way that we treat errors in census counts. NSA grants use BEA measures of food cost inflation. We treat the NSA grants as unaffected by census inaccuracy.

Child Care Mandatory and Matching Funds of the Child Care and Development Fund (93.596)

Mandatory funds are allocated based on a state's share of expenditures for the now-repealed AFDC child care programs during the years 1992-1995, and thus not based on government statistics. Matching funds are allocated proportional to a state's population under age 13, which is determined from postcensal single year of age estimates.

Child Care and Development Block Grant (93.575)

Grants to states depend on a variety of statistics. Census statistics include state population and state population under age 5 (from postcensal single year of age estimates). Non-census statistics include BEA personal income and the number of children receiving free or reduced school lunch from the Department of Agriculture. Specifically, state *i* receives grant G_i where

$$G_{i} = 0.5A \left(\frac{\breve{Y}_{i}H_{i}}{\sum_{j}\breve{Y}_{j}H_{j}} + \frac{\breve{L}_{i}H_{i}}{\sum_{j}\breve{L}_{j}H_{j}} \right)$$

Here, *A* is the total amount allocated for the program, Y_i is population under age 5, L_i is free or reduced lunch population, I_i is personal income, P_i is population, and $H_i = \min(1.2, \max(0.8, \vec{P}_i / \vec{I}_i)))$ is the reciprocal of per capita personal income, constrained to lie between 0.8 and 1.2.

Social Services Block Grant (93.667)

Grants to states are proportional to state population statistics.

English Language Acquisition Grants (84.365)

Grants to states use measures for the number of Limited English Proficiency (LEP) children and the number of immigrant children and youth, both estimated from the American Community Survey. Specifically, state grants are proportional to the sum of 80% of the state's national share of LEP children plus 20% of the state's national share of immigrant children and youth.

Special Education-Grants for Infants and Families (84.181)

Grants to states are proportional to state population age 0 to 2, which is determined from postcensal single year of age estimates. Each state receives a minimum of 0.5% of all funding allocated for the grant program.

Nonpoint Source Implementation Grants (66.460)

Grants to states are determined using a formula based on a variety of government statistics, including 1990 census state population and 1987 postcensal state population estimates. New census statistics are not used in the determination of grant amounts.

Title V Delinquency Prevention Program (16.548)

Grants to states are proportional to a state's youth population under the maximum age of original juvenile court delinquency jurisdiction, which is estimated from postcensal single year of age estimates. The maximum age of juvenile court delinquency jurisdiction varies by state and is available at http://www.ojjdp.gov/ojstatbb/structure_process/qa04101.asp for 2013.

Appendix 4: Specification of Population Sizes of States in 2020

For the analysis of the sensitivity of House apportionment to census data quality, we developed a joint probability distribution for the true and census-estimated population sizes. The means of the true population sizes were set equal to the state population projections for 2020 prepared by the University of Virginia's Weldon Cooper Center for Public Service (2013) based on the 2010 census results. We did not use projections from the Census Bureau because the Bureau no longer produces state population projections. The variances of the 2020 population sizes were specified to be consistent with the observed levels of error in state population projections prepared a decade earlier by the Census Bureau. Specifically, in 2005 the Census Bureau used 2000 census results to project state populations for July 1, 2010. The error in those projections was estimated by the difference between the projection, say Y, and the Census Bureau's population estimates for July 1, 2010, say X, which are equal to the 2010 census enumeration adjusted for births, deaths, and net migration over the 3 month interval from April 1 to July 1. The relative error was computed as the error Y - X divided by x, or equivalently Y/X - 1. The relative errors were observed to be approximately normally distributed about zero, and the relative errors tended to be closer to zero for the larger states than the smaller states. To model the squared relative error as a function of the true population size, we used Stata® 11 to carry out a lowess fit of $(Y/X-1)^2$ against X using a bandwidth of 0.8 and preserving the mean. The lowess fitted values were taken as the relative variances of the population projections for 2020 and as the relative variances of the future 2020 population sizes of the states. As specified in Section 2, the joint conditional distribution of the census estimates given the true state population sizes was modelled as independent across states, with mean equal to the true size and constant coefficient of variation.

The assumption of independence for the distribution of true population sizes of states was motivated by the following considerations. State population projections typically will be controlled to sum to national forecasts, which account for births, deaths, and net immigration since the last census. The latter likely induce a source of positive covariance among state population projections (if we treat the projections as random variables, as in Alho and Spencer 2005). However, the dominant source of error in forecasts of 10 years or shorter will be uncertainty about interstate migration. Since the interstate migration flows must sum to zero, the covariances cannot all be positive, but will have a more complex pattern. For simplicity, we take the 2020 population sizes to be independent, knowing that only the population shares matter for apportionment, and that the shares implicitly include some negative covariances because the sum of shares is always 1.

41

State	Mean	Coefficient of Variation (%
Alabama	5,066,866	2.6
Alaska	811,718	4.4
Arizona	7,604,382	2.5
Arkansas	3,120,724	3.1
California	41,715,522	2.0
Colorado	5,733,049	2.5
Connecticut	3,723,612	2.8
Delaware	997,528	4.2
Florida	21,784,582	1.8
Georgia	11,078,010	2.3
Hawaii	1,489,774	3.8
Idaho	1,772,613	3.7
Illinois	13,277,307	2.1
Indiana	6,804,046	2.5
lowa	3,085,572	3.0
Kansas	3,011,419	3.1
Kentucky	4,558,229	2.7
Louisiana	4,635,071	2.7
Maine	1,394,018	3.9
Maryland	6,282,303	2 5
Massachusetts	6.806.874	2.3
Michigan	10,074,617	2.3
Minnesota	5.704.065	2.5
Mississinni	3,111,177	2.5
Missouri	6.336.145	2.5
Montana	1.055.292	2.5
Nobraska	1 908 775	4.1
Novada	3,328,548	2.1
Now Hampshiro	1,446,097	3.1
	9 252 696	5.9
New Movico	2 307 561	2.5
New Verk	19 952 674	5.5
New YOR	10 736 114	1.8
North Dakata	678 125	2.3
	11 762 965	4.5
Olilohama	3 006 056	2.2
Orianoma	2,200,930 A 222 601	2.7
Deenenhani	12 061 010	2.7
Perinsylvania	1 095 057	2.1
Knode Island	I,UQD,AD/	4.1
South Carolina	3,118,310	2.7
South Dakota	۵۵۵,۶43 ۵۵۵,۵۵۵	4.3
Tennessee	0,919,966	2.5
lexas	20,/30,112	1.8
Utah	3,193,030	3.1
Vermont	662,770	4.7
Virginia	8,871,484	2.3
Washington	7,576,478	2.3
West Virginia	1,817,852	3.5
Wisconsin	6,004,398	2.5
Wyoming	594,027	4.8

Table A4. Specifications of moments of state populations in 2020.

Appendix 5: Modeling Effects of Census Error for Unemployment Rates and CPI-U

CPS Unemployment Rates

From 2008 through 2013, two programs were available in the U.S. that provided additional funding for unemployment insurance (UI) for states with elevated unemployment rates, Extended Benefits (EB) and Emergency Unemployment Compensation (EUC). We modeled the effect of census error on UI for these two programs.

We used estimates of differential net undercount by age, race and sex² as well as unemployment rates by these groups³, to estimate error in unemployment rates. We obtained unemployment rate estimates for cross-tabulations by age, race and sex via raking. We then developed an adjusted estimate of the 2010 U.S. unemployment rate, adjusting the original estimate by age/race/sex groups. We estimated that adjusting the U.S. unemployment rate for undercount would increase the 2010 estimate from 9.63% to 9.66%.

We determined how EB and EUC benefits would change for each year 2008 to 2013 if unemployment rates had been increased by 0.1% for census error. We found that EB would have increased by \$18.6 million per year and EUC would have increased by \$724.4 million per year. As these reflect results occurring near a major recession when two programs for extra benefits were available, we estimated the effect of census undercount by group on UI distribution each year by multiplying these results by 0.1. We suggest 0.1 as a plausible subjective probability of a specific year having a program like EUC available due to a severe recession. The program user can adjust this subjective probability in our estimation programs and study the sensitivity of estimates.

To relate errors in state population counts to errors in state unemployment rates, we used the simplifying assumption that the estimated effect of undercount by age, race and sex on state unemployment rates is proportional to the effect of error in the state population counts on errors in estimated state unemployment rates.

² Data from <u>www.census.gov/coverage_measurement/pdfs/g01.pdf</u>, accessed February 12, 2015.

³ Data from <u>www.bls.gov/cps/aa2010/cpsaat3.pdf</u> accessed February 20, 2015.

CPI-U

Estimating the bias in CPI-U due to census undercount is more difficult than for CPS unemployment rates. The census serves as a sampling frame for the Consumer Expenditure Survey, which determines the weights on different consumption component categories used to calculate CPI-U. Separate surveys estimate the price indexes within a consumption component category (Bureau of Labor Statistics 2007). Correspondence with the Bureau of Labor Statistics led us to find that, for the most part, the surveys used for the consumption component indexes rely on the 1990 census for a sampling frame. The exception is for rent and owners' equivalent rent, which are based on the CPI Housing Survey and rely on more frequently updated sampling frames.⁴

To model the effects of census on error on CPI-U, we examine how changes in owner and renter census undercount affect the weights on the consumption component indexes used to calculate CPI-U.⁵ We estimated a historical series of error in CPI-U due to census error and forecasted the error with confidence intervals through 2029 with an AR(2) model. We found the error to be negative, with the lower bound of the 95% prediction interval for each forecasted year through 2029 never less than -0.015%. Thus, we chose -0.015% as a lower bound for the forecasted relative error each year.

Grant distributions for three programs depend on poverty thresholds using CPI-U: SNAP, Head Start and Pell Grants. For SNAP, we used data on SNAP recipients and amounts and modeled how much SNAP benefits would change with a relative error of -0.015% in CPI-U.

Head Start preschools are eligible for grants if a certain percentage of students are below poverty thresholds. We assume CPI-U error has no effect on Head Start funding as preschools can always recruit more or fewer students to their schools as poverty thresholds change.

We estimated the effect of census error on Pell Grant funding as -0.015%, recognizing this would be a bound on the error due to caps on grant amounts.

To relate errors in state population counts to error in estimating CPI-U, we used the simplifying assumption that the estimated effect of undercount by owners and renters on CPI-U is proportional to the effect of error in the state population counts on errors in the estimated CPI-U.

⁴ See <u>www.bls.gov/opub/mlr/2013/article/updating-the-rent-sample-for-the-cpi-housing-survey.htm</u>, accessed February 20, 2015.

⁵ Data downloaded from <u>http://research.stlouisfed.org/fred2/</u> on August 31, 2014.

Acknowledgements

We are grateful for the assistance and comments of many people. While noting that responsibility for all views expressed, and any errors in the report, are solely those of the authors, we do gratefully acknowledge and thank the following individuals.

Wesley Basel (Social, Economic, and Housing Statistics Division, Census Bureau), William R. Bell (Associate Director for Research and Methodology), Patrick Cantwell (Decennial Statistical Studies Division), Robert W. Colosi (2020 Research and Planning Office, Census Bureau), Mark Cooley (Center for Medicaid and State Operations, DHHS), Richelle Davis (Office of Special Education Programs, DE), Dennis Finney (Grants Administration Division, EPA), Gay Gilbert (Office of Unemployment Insurance, DOL), Mark Glander (Office of Elementary and Secondary Education, ED), Robert M. Groves (Georgetown University), Howard R. Hogan (Director's Office, Census Bureau), Al Jones (Office of Special Education Programs, DE), Christa D. Jones (Director's Office, Census Bureau), Edward L. Kobilarcik (Decennial Management Division, Census Bureau), Thomas A. Louis (Director's Office, Census Bureau), Charles F. Manski (Northwestern University), Mary H. Mulry (Center for Statistical Research Methodology, Census Bureau), Laura Nixon (Governments Division, Census Bureau), Nancy A. Potok (Director's Office, Census Bureau), Burton H. Reist (2020 Research and Planning Office, Census Bureau), Todd Stephenson (Office of Elementary and Secondary Education, ED), Richard Strauss (Center for Medicaid and State Operations, DHHS), John H. Thompson (Director, Census Bureau), Robert Trombley (Office of Planning, Evaluation and Policy Development, ED), Philip Vidal (Governments Division, Census Bureau), Frank A. Vitrano (2020 Research and Planning Office, Census Bureau), Daniel H. Weinberg (Associate Director for Research and Methodology, Census Bureau), Patrick Wells (Administration for Children and Families, DHHS), Everett G. Whiteley (Budget Office, Census Bureau), Ron Wilus (Office of Unemployment Insurance, DOL).