Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 519–549

Missing the Mark? Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates

Michael Davern[1], *Lynn A. Blewett*[2], *Boris Bershadsky*[3], *and Noreen Arnold*[4]

The Demographic Supplement to the U.S. Current Population Survey (CPS) is used to produce state estimates of health insurance coverage and income. These estimates are used in federal allocation formulas that distribute $10–11 billion annually to states for the State Children's Health Insurance Program (SCHIP) and the Elementary and Secondary Education Act. The purpose of this article is to examine the CPS for evidence of bias in state estimates due to missing data imputation and estimate the extent of the bias for each of the fifty-one states and Washington DC. Comparing three years of CPS (1998–2000), to the Census 2000 Supplementary Survey and 1990 Decennial Census data benchmarks, we find evidence of bias in state estimates of earned income. We also extend the technique to the CPS state health insurance coverage estimates and find even more evidence of bias. In general, the "better off" states (those with higher insurance coverage rates or more income) tend to be even "better off" (have higher estimates of average income and coverage rates) after correcting for bias (and vice versa). We conclude by considering alternative strategies for the U.S. Census Bureau to alter its current imputation procedures.

Key words: Hot deck; allocation formulas; item nonresponse; missing data; 1990 Decennial Census; Census 2000 Supplementary Survey; American Community Survey.

1. Introduction

The U.S. Census Bureau produces annual estimates of health insurance coverage and household income for each of the 50 states and the District of Columbia using data from the Annual Demographic Supplement to the Current Population Survey (CPS) (Mills 2001; DeNavas-Walt, Cleveland, and Roemer 2001; Dalaker 2001). These data widely

[1] University of Minnesota, School of Public Health, State Health Access Data Assistance Center, 2221 University Avenue, S.E., Minneapolis, MN 55414, U.S.A. Email: daver004@umn.edu
[2] University of Minnesota, School of Public Health, State Health Access Data Assistance Center, Health Services Research and Policy, 420 Delaware Avenue, S.E., MMC 729, U.S.A.
[3] University of Minnesota, School of Public Health, State Health Access Data Assistance Center, Health Services Research and Policy, 420 Delaware Avenue, S.E., MMC 729, U.S.A.
[4] University of Minnesota, School of Public Health, State Health Access Data Assistance Center, 2221 University Avenue, S.E., Minneapolis, MN 55414, U.S.A.
Acknowledgments: The authors wish to thank Linda Bilheimer, Gestur Davidson, John Czajka, Steve Ruggles, Steve Zuckerman, Paul Fronstin, Chuck Nelson, Tim Beebe, Paul Siegel, and Scott Susin for their helpful comments. We would also like to thank those people who gave us feedback during presentations of this article at the Minnesota Population Center's presentation series in September 2001, the U.S. Census Bureau in October 2001, and during an informal presentation in Washington DC in November. All remaining problems are the fault of the authors only. Preparation of this manuscript was funded by Grant no. 038846 from The Robert Wood Johnson Foundation to the State Health Access Data Assistance Center at the University of Minnesota, School of Public Health.

cited in the media and used to allocate federal dollars to states. The state income and health insurance data are used in a formula to distribute $3-4 billion in federal allocations to states for the State Children's Health Insurance Program (SCHIP) (Federal Register 2000). State level income data are also used as one component of a complex statistical model that distributes seven billion dollars to educationally disadvantaged children under Title I of the Elementary and Secondary Education Act (National Research Council 2000). Concern about bias in the development of these estimates is critical from a technical as well as policy perspective as bias may lead to misallocation of funds for the SCHIP program and for federal education funding. In this article we examine the U.S. Census Bureau's process for dealing with item nonresponse in the CPS state estimates of health insurance coverage and income to determine whether bias exists and, if so, whether this bias is significant.

The article is organized in the following manner. First, we present the background on the hot deck procedure used to impute values for item nonresponse in the Census Bureau's demographic surveys. Second, we introduce our model and hypotheses for testing whether the state estimates are biased because of the hot deck procedure. Third, we present the results of our analysis, including whether the bias makes a significant difference in the state estimates. And, finally, we discuss four potential models to correct the state health insurance coverage and income estimates.

2. Missing Data and Imputation

Missing data in the form of item nonresponse is a common problem in survey research (Groves, Dillman, Eltinge, and Little 2001). Item nonresponse happens when a respondent terminates an interview after it has begun, accidentally skips an item, refuses to answer an item, or does not know the answer to the specific question. Regardless of the reason, the result is missing data for the particular item(s). Questions that are considered sensitive (e.g., income), and recodes that use several source variables (e.g., health insurance coverage), tend to have higher rates of missing data. Approximately 17.0 percent of the people who have earned income do not report it in the Current Population Survey. Health insurance coverage is a recode derived from several items and its overall rate of missing values is equal to 11 percent in the Current Population Survey. (Imputation rates for the CPS are available in Table 4.)

Demographers and statisticians have developed a wide range of techniques for dealing with item nonresponse (e.g., Kalton 1983; Kalton and Kasprzyk 1986; Little and Rubin 1987; Rubin 1996; Marker, Judkins, and Winglee 2001; Heeringa, Little, and Raghunathan 2001). Most of the techniques use completed cases to impute some kind of model-based estimate for the cases with missing data. Multiple imputation represents the most statistically sound variation of model-based imputation (Rubin, 1996). Nevertheless, a combination of historical inertia, several studies affirming that hot deck estimates are relatively valid (e.g., Marker, Judkins, and Winglee 2001; Mason, Lesser, and Traugott 2001; David, Little, Samuhel, and Triest 1986), and relative simplicity of understanding hot deck imputation make the latter the most widespread technique in demographic surveys. The U.S. Census Bureau uses hot deck imputation to correct for

item nonresponse in all its demographic surveys (CPS, Decennial Census, Survey of Income and Program Participation, Survey of Program Dynamics, etc.).

Hot deck is a type of model-based imputation by which a respondent's valid value for a specific variable is assigned to another respondent who does not have a valid value for the variable. The respondent with the valid value is called a "donor" and a person with a missing value is called a "recipient." For example, if the donor is 35 years old, then the recipient (respondent with missing age) is given a value of 35 and the donor maintains the age of 35.

The process of selecting a donor is the most important component of the hot deck model. Potential donors are sectioned into homogeneous groups called "cells" and the cells are defined by many parameters. For example, all white, unemployed, college educated males over the age of 65 with a valid value for the specific variable can be placed into one cell and nonwhite, unemployed, college educated males over 65 can be placed into another cell. Recipients are matched to these homogeneous cells of donors on the basis of their characteristics. A donor is selected from within a cell and supplies his or her value to the recipient.

The characteristics used to group the respondents should be highly correlated with the variable being imputed. For example, when imputing income, donors are matched with recipients on the basis of highest educational level and current occupation because education and occupation are highly correlated with income. The variables chosen to match the donors and the recipients form the basis of a "model" for predicting the imputed variable. A good imputation procedure should provide unbiased estimates of the mean and variance of the variable by correcting for potential distributional differences between people with and without reported data. The basic underlying assumption is that the value of the variable being estimated (e.g., health insurance coverage) is not conditional on (i.e., moderated by) the missing data mechanism (Little and Rubin 1987). For example, all the respondents with missing health insurance data do not have a different relationship between health insurance coverage and age than all the respondents with reported data.

Although properly specified imputation can alter basic distributional summary statistics (means and variances) from the statistics calculated using complete cases only, it should not transform the relationships among variables. If there was a relationship between two variables in the reported data it should be the same in the imputed data, and no new relationships should appear after the imputation. The basic idea of model-based (and particularly, hot deck) imputation is to use the existing relationships within the reported data to adjust for distributional differences among those who are likely to report data and those who are less likely.

The hot deck is limited in the number of variable levels it can have. For example, the variable "highest degree attained" can be broken down into three variable levels (or cells) for the hot deck: less than high school, high school diploma, and college degree. The number of total hot deck cells is equal to the product of the number of variable levels of each varible used to match donors with recipients. If there are too many variable levels used in the hot deck, then many of the cells will not be populated with donors. The more variable levels that are used (i.e., the more hot deck cells), the more donors are needed for the hot deck to work.

In assessing accuracy of datasets with imputed hot deck values for estimating demographic characteristics, the critical question is whether the relationship among the variables used to compute the estimate was maintained in the hot deck procedure, which in turn depends on the particular analysis being run. The same hot deck procedure can be accurate when estimating health insurance coverage at the national level, and be biased when comparing health insurance coverage across states. Thus, the ability of the hot deck to have the desired properties of high-quality imputation depends on both the procedure itself and the specific analysis the analyst is interested in conducting.

Because the state estimates of health insurance coverage and income are used in federal allocation funding formulas we have undertaken an examination of whether the state estimates are biased because of the hot deck procedure that the U.S. Census Bureau currently employs. With this in mind, there are three questions we attempt to answer in this analysis: 1) Does hot deck imputation create a significant bias in comparing health insurance coverage and household income across states? 2) Does this bias make a difference in the overall state estimates? 3) If there is significant bias, then how can the process be improved?

3. Evaluation Model and Hypotheses

Our evaluation of the CPS for bias consists of two components. In order to infer that bias exists in the CPS state estimates of income and health insurance coverage, the statistical model needs to be combined with our expectation of what would happen to the model's coefficients if bias existed in the state estimates. The general model we use to evaluate possible imputation biasis:

$$Y_{j,i} = \alpha + \beta_1 * (\text{state}_j)_i + \beta_2 * (\text{impute})_i + \beta_3 * (\text{state}_j)_i * (\text{impute})_i + \varepsilon_i \qquad \text{(Model 1)}$$

The dependent variable (Y) is earned income or health insurance coverage and different estimation techniques are used for the continuous income variable and the binomial coverage variable. The variable (state_j) is a binary indicator of whether respondent i lives in the state j being evaluated in the current model. The model describes 51 equations ($j = 1$ to 51, 50 states and the District of Columbia) that compare the individual states with the rest of the country. The "impute" variable is a binary indicator of whether the dependent variable Y for respondent i was imputed or not. For the outcome variables the model predicts the following:

(1) Earned income/insurance for reported cases in the state of interest $= \alpha + \beta_1$
(2) Earned income/insurance for imputed cases in the state of interest $= \alpha + \beta_1 + \beta_2 + \beta_3$
(3) Earned income/insurance for reported cases in the remaining states $= \alpha$
(4) Earned income/insurance for imputed cases in the remaining states $= \alpha + \beta_2$
(5) Interaction effect of state of interest and imputed earned income/insurance $= \beta_3$

If there is no bias in a specific state's imputations we expect $\beta_3 = 0$ (5) and earned income/insurance for imputed cases in the state of interest to be reduced to $\alpha + \beta_1 + \beta_2$ (2). In other words, we expect that after controlling for the general state effect associated with reported scores, $\alpha + \beta_1$ (1) and the general effect of being imputed, $\alpha + \beta_2$ (4), that the interaction between being imputed and in the state of interest (estimated by β_3) should be zero. Every case of rejecting the null hypothesis that the interaction effect is zero ($\beta_3 = 0$) is

considered as an indication of *possible* bias in imputation. Furthermore, by assuming that our model of "unbiased" imputation is correct, we can simulate unbiased imputation by forcing the interaction effect to equal zero, $\beta_3 = 0$ (5), and comparing the outcome with the predicted score when the interaction effect is not constrained to equal zero.

In order to infer the existence of bias, not only should β_3 be significantly different from 0, but the pattern emerging from the analysis should follow a pattern of bias across the states. (There is the possibility that β_3 does not actually equal zero and the imputation is not biased. If there is a relationship that is taken into account by the hot deck that is not taken into account by this simple model, then it is possible that unbiased imputation occurs when β_3 does not equal 0. This is why it is essential that the "bias pattern" laid out in the hypotheses 1 and 2 be followed to infer the presence of bias.) In addition to a statistically significant interaction effect, β_3, the bias pattern will be evidenced by parameter estimates for β_3 and β_1 being in opposite directions. In other words, in a state with higher than average (as compared with the country or census region) income (i.e., a positive β_1) we expect the imputed cases not to reflect the higher than average income pattern for the state. The imputed cases within the state should, in general, have a negative parameter estimate associated with the interaction effect because imputation is drawing the imputed state residents back toward the overall average for the region or country and not reflecting the state average. If there is a significant effect on the incomes or coverage rate of residents of a particular state over living in other parts of the United States (measured by β_1), then the imputed cases will not reflect that relationship. The following hypothesis sums up our expectation for inferring that bias exists:

> **Hypothesis 1:** To the extent that the magnitude of the state effect β_1 is different from the lowest level of geography (e.g., census region or country) specified in the hot deck, the larger (and in the opposite direction) the magnitude of the imputation bias estimate β_3 will be.

The bias pattern should also be more likely to emerge in states with smaller sample sizes relative to others. People in smaller states with missing data are more likely to end up with the donated value from someone who is not in their state if the selection of donors is done randomly within the homogeneous hot deck cells. On the other hand, people in states with larger than average samples will be more likely to end up with the donated value from someone within their state.

> **Hypothesis 2:** The smaller the state sample size relative to others, the less likely it is to maintain significant state effects through imputation.

Bias in the state estimates is a function of the magnitude of the difference between the state effect β_1, the lowest level of geography used in the hot deck, and the sample size from within the state relative to others.

4. Data and Model Implementation

In the following analyses we use three data sources: the 1990 Decennial Census, the Census 2000 Supplementary Survey (C2SS), and the CPS. (Census 2000 would be preferable to 1990 Census data, but since the microdata were not publicly available at the time of our analysis, we used 1990 Census data. Although our main concern is with

bias in the CPS state estimates, it is important to compare the results from the CPS with another data source collected by the U.S. Census Bureau. The Decennial Census measures several of the same concepts as CPS, such as labor force participation and income. For the most part they are processed using similar hot deck imputation strategies and editing procedures.

There are, however, three significant differences between the CPS, the C2SS and the Decennial Census that impact the potential for bias in the state estimates. The first major difference between the data sources is that the 1990 Decennial Census data are processed state by state, whereas the C2SS and the CPS are processed with all states together. Another difference, although not a major one, is that the Decennial Census uses joint item imputation procedures and the C2SS and CPS use single item imputation procedures. This means that if one key item is missing from the income data of a respondent's record, all the data from that module are replaced with data from a single donor. This practice can wipe out reported data and replace it with imputed data. The CPS only imputes specific missing items. This difference accounts for the higher imputation rates in the Decennial Census than the CPS (see Table 4) (U.S. Census Bureau 2001; U.S. Census Bureau 1993; and U.S. Census Bureau 1987). Thus, even though state of residence is not explicitly a factor in the Decennial Census hot deck, in practice it actually is a factor because the data are processed by state and most donors are taken from within each state. The third major difference is that the 1990 Census and Census 2000 Supplementary Survey (C2SS) select a donor from the appropriate cell based on geographic proximity to the recipient. All things being equal, this proximity preference makes someone in a state more likely to receive a donor's value from someone else within their state regardless of whether the variable of state is used on the hot deck procedure to form allocation cells. This makes the likelihood of maintaining any state effect higher when the geographic proximity component is added to the hot deck procedure. These processing differences should result in less biased state income estimates from the Decennial Census and C2SS relative to the CPS.

For this article we use three years of CPS Demographic Supplement data from 1998-2000 and we evaluate the state estimates of income and health insurance coverage. There is one important difference between the respective hot deck procedures regarding CPS income and health insurance items that should effect the amount of bias in the state estimates. The income hot deck procedure uses Census Region as one of the variables to define hot deck cells and the health insurance hot deck does not. Thus the lowest level of state aggregated geography used in the hot deck for income is Census Region and for Health insurance it is the entire country.

The U.S. Census Bureau uses three-year estimates from the CPS for making comparisons among states (Mills 2001). The Demographic Supplement to the CPS is conducted annually and is used to make estimates of household income, health insurance status, and poverty for the previous calendar year (Mills 2001; DeNavas-Walt, Cleveland, and Roemer 2001; Dalaker 2001). For example, the data collected in 2001 asks respondents about health insurance coverage and income from the previous calendar year (2000). The CPS sample size for this period was roughly 65,000 households per year (U.S. Census Bureau 2000).

In this article we use the Integrated Public Use Microdata Series (IPUMS) 1 percent sample from the 1990 Decennial Census (Ruggles et al. 1997). The Decennial Census data

is used to construct labor force, income and poverty statistics for the country as well as a host of other basic demographics. It is conducted once every ten years with most of the households in the United States receiving a "short form" containing basic demographics, and roughly one in six households receiving the "long form" containing questions asking more detailed information regarding ancestry, housing characteristics, income, commuting patterns, and labor force participation (Alexander 1998). The long form data are used to compile the earnings data evaluated in this article.

The C2SS data are drawn from the Public Use Micro (PUM) data available from the U.S. Census Bureau. The PUM data represent roughly 1 in 4 of the households sampled as part of the C2SS and the total sample size for the C2SS was 890,698 households throughout the U.S. (U.S. Census Bureau 2002). The C2SS was the first nation-wide test of the American Community Survey (ACS). The ACS, if fully funded, will be an annual survey with very similar content to the Decennial Census long form of approximately 3,000,000 households (Alexander 1998), and will completely replace the Decennial Census long form operation in the future.

4.1. Earned income

The Decennial Census, CPS, and C2SS measure total earned income as a combination of self-employment income (including farm income) and wage-salary income. (In the CPS, earned income accounts for approximately 90% of the typical families' total income. Other sources include transfer program income and asset income.) In all three surveys we consider respondents with all three nonmissing components as not being imputed. For the CPS and C2SS several respondents have at least one component of income imputed and one that is not imputed. If the imputed portion accounted for less than 50 percent of total earned income, we classify them as not imputed. (For the most part people fall at either extreme, with almost all of the income either imputed or not imputed.) The Decennial Census, on the other hand, uses a process of joint item imputation, which means that if any one item is missing, the entire income record is imputed from a donor (U.S. Census Bureau 1993). In this process reported data can be discarded and replaced with imputed data from the donor's record. Because of this fact the 50 percent distinction is not necessary with the Decennial Census data, and this also accounts for the higher imputation rates in the Decennial earned income data than the CPS earned income data. See Table 4 for details on CPS imputation rates. (1990 Census and C2SS income imputation rates are available from the authors upon request.)

For the following analysis the income data were adjusted to deal with three problems: (1) The Consumer Price Index for all Urban (CPI-U) areas was used to convert the 1990 Census income data (that corresponds to the 1989 calendar year) to 1998 dollars (the mid year of the three CPS calendar years used for this analysis). (2) A single constant was added to all the CPS and Census data to bring all the negative income amounts above zero. (3) The natural logarithm of each income value was taken. The individual state models were run using the log-transformed data.

We use the ordinary least squares estimation and normalized person survey weights to obtain parameter estimates. (To normalize the weights, the person weight is divided by the average person weight.) In order to develop predicted scores of actual dollars we use a

smearing estimate to transform the logged scale values to dollars for Tables 1 and 2 (Duan 1983). The standard errors are adjusted to account for the sampling design of the Decennial Census, C2SS, and the CPS by using the design effect parameters provided by the U.S. Census Bureau. The 1990 Census estimates are adjusted using the estimated 1990 design effect for income sources (U.S. Census Bureau 1993). The C2SS estimates are adjusted using the C2SS design effect for earned income data (U.S. Census Bureau 2002). CPS design effect is estimated using the state appropriate adjustment and the appropriate adjustment parameter for individual earned income (U.S. Census Bureau 1998–2000). Furthermore, the observations from successive years of CPS data are not independent because of the household rotation schedule (U.S. Census Bureau 2000). For example, roughly half of the households (and many of the people) in the March 2000 CPS are also in the March 2001 CPS. Therefore, an additional adjustment was made to correct the standard errors for this nonindependence (U.S. Census Bureau 1998–2000).

4.2. Health insurance coverage

Health insurance coverage bias is only evaluated using the CPS because the 1990 Decennial Census and the C2SS did not collect information on health insurance coverage.

CPS health insurance coverage is determined using several variables. Respondents are asked if they have employer or union based insurance coverage, Medicaid coverage, Medicare coverage, military/VA health insurance, or some other form of health insurance. If a respondent has a "yes" on their record for any of these questions, the person is considered insured. If the respondent has a "no" on his or her record for all of them, then the respondent is considered uninsured (Mills 2001). For our purposes we define a global health insurance coverage imputation flag by using the following logic: (1) if the person has coverage, all positively-identified sources of coverage will have to be imputed, and (2) if the person does not have coverage, at least one of the sources of coverage needs to be imputed.

There is a limitation in the Public Use File in that the imputation of a "family" policy, and therefore the imputation of coverage for all dependents in the family, is not known. We have alerted the U.S. Census Bureau to this problem. What we do know is when someone has health insurance coverage from an imputed family coverage policy. What we do not know, and would need to know, is whether imputation of "noncoverage" occurred as well. In other words all the dependents with the dependent imputation flag set to imputed, have coverage (there are none that were imputed to not have coverage). Those who were allocated to not have a family policy, did not have their imputation flag set to "imputed." We decided not to treat those cases as imputed.

The regression model used to evaluate the health insurance imputations used the binomial variable (covered, not covered) of respondents' health insurance coverage status as the dependent variable. We ran a logistic regression to estimate Model 1 using normalized person weights (see above) and a maximum likelihood estimator. The standard errors were adjusted to account for the sampling design of the CPS by calculating a design effect for each state based on the state appropriate adjustment and the appropriate adjustment parameter for individual health insurance coverage (U.S. Census Bureau 1998–2000). As with the CPS earned income data, the observations from successive years of CPS data are not independent because of the household rotation schedule. Therefore, an

Table 1. Direct State effect, interaction effect and percent change in the predicted scores with and without the estimated interaction effect using earned income data from the 1990 Decennial Census

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Alabama	−3,497**	373	22,752	23,125	1.6
Alaska	4,300**	1,061	30,485	31,547	3.5
Arizona	−1,733**	−314	24,495	24,181	−1.3
Arkansas	−5,990**	756	20,254	21,010	3.7
California	3,497**	403*	29,231	29,634	1.4
Colorado	−928**	−201	25,284	25,083	−0.8
Connecticut	7,527**	−691	33,625	32,934	−2.1
Delaware	1,646*	−1,483	27,843	26,360	−5.3
Dist. of Columbia	5,535**	−2,080	31,722	29,642	−6.6
Florida	−1,899**	1,026**	24,344	25,370	4.2
Georgia	−628*	−196	25,592	25,396	−0.8
Hawaii	1,295*	500	27,485	27,986	1.8
Idaho	−5,160**	461	21,053	21,515	2.2
Illinois	1,577**	−229	27,714	27,485	−0.8
Indiana	−1,859**	−335	24,387	24,051	−1.4
Iowa	−4,131**	−363	22,116	21,754	−1.6
Kansas	−2,022**	−354	24,199	23,845	−1.5
Kentucky	−4,083**	−153	22,171	22,019	−0.7
Louisiana	−3,841**	−21	22,420	22,399	−0.1
Maine	−3,089**	51	23,124	23,175	0.2
Maryland	5,161**	−1,245	31,278	30,033	−4.0
Massachusetts	3,759**	576	29,847	30,423	1.9
Michigan	721**	−403	26,906	26,504	−1.5
Minnesota	−642*	−434	25,575	25,140	−1.7

Table 1. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Mississippi	−5,700**	160	20,552	20,712	0.8
Missouri	−2,465**	−773	23,799	23,026	−3.2
Montana	−6,771**	1,377	19,442	20,819	7.1
Nebraska	−4,507**	−1	21,719	21,718	0.0
Nevada	−122	973	26,071	27,045	3.7
New Hampshire	2,095**	−17	28,281	28,264	−0.1
New Jersey	6,699**	99	32,662	32,761	0.3
New Mexico	−4,432**	−1,046	21,794	20,748	−4.8
New York	4,423**	−247	30,281	30,034	−0.8
North Carolina	−2,951**	227	23,323	23,551	1.0
North Dakota	−6,306**	1,138	19,904	21,042	5.7
Ohio	−1,110**	−111	25,133	25,022	−0.4
Oklahoma	−3,684**	−793	22,573	21,780	−3.5
Oregon	−2,913**	351	23,313	23,664	1.5
Pennsylvania	−256	259	25,941	26,201	1.0
Rhode Island	644	557	26,836	27,393	2.1
South Carolina	−3,308**	186	22,932	23,118	0.8
South Dakota	−7,295**	1,814	18,916	20,730	9.6
Tennessee	−3,109**	612	23,139	23,751	2.6
Texas	−1,714**	−908**	24,660	23,752	−3.7
Utah	−4,318**	775	21,898	22,673	3.5
Vermont	−2,061*	203	24,141	24,344	0.8

Table 1. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Virginia	1,604**	−870	27,782	26,912	−3.1
Washington	−169	152	26,029	26,182	0.6
West Virginia	−4,605**	373	21,616	21,989	1.7
Wisconsin	−2,303**	−280	23,942	23,662	−1.2
Wyoming	−3,706**	524	22,497	23,021	2.3

Source: IPUMS 1% Sample of the 1990 Decennial Census
$N = 1{,}290{,}797$
$*p < .01$
$**p < .001$

Table 2. Direct State effect, interaction effect and percent change in the predicted scores with and without the estimated interaction effect using earned income data from the Census 2000 Supplementary Survey

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Alabama	−3,452**	860	26,356	27,216	3.3
Alaska	2,177	3,710	31,932	35,642	11.6
Arizona	−1,016	−1,376	28,794	27,418	−4.8
Arkansas	−5,131**	439	24,683	25,122	1.8
California	2,538**	−2,371**	32,278	29,907	−7.3
Colorado	959	−452	30,722	30,270	−1.5
Connecticut	7,337**	781	37,003	37,785	2.1
Delaware	2,958	1,538	32,713	34,252	4.7
Dist. of Columbia	9,550**	−6,342	39,310	32,968	−16.1
Florida	−2,311**	889	27,535	28,424	3.2
Georgia	88	1,465	29,808	31,272	4.9
Hawaii	364	2,130	30,122	32,252	7.1
Idaho	−6,305**	1,517	23,480	24,997	6.5
Illinois	2,070**	−423	31,764	31,341	−1.3
Indiana	−1,050	−639	28,750	28,111	−2.2
Iowa	−4,369**	−569	25,443	24,875	−2.2
Kansas	−3,561**	2,225	26,218	28,442	8.5
Kentucky	−5,206**	483	24,633	25,115	2.0
Louisiana	−4,050**	2,158	25,753	27,911	8.4
Maine	−4,401**	1,515	25,378	26,894	6.0
Maryland	4,738**	−3,391	34,477	31,086	−9.8
Massachusetts	6,006**	−149	35,637	35,488	−0.4
Michigan	1,040*	838	30,742	31,580	2.7
Minnesota	497	372	30,249	30,622	1.2

Table 2. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Mississippi	−4,965**	216	24,861	25,077	0.9
Missouri	−2,851**	818	26,956	27,774	3.0
Montana	−6,834**	4,245	22,940	27,185	18.5
Nebraska	−4,511**	1,711	25,270	26,982	6.8
Nevada	−642	−449	29,135	28,687	−1.5
New Hampshire	2,093	799	31,848	32,647	2.5
New Jersey	6,983**	435	36,513	36,948	1.2
New Mexico	−5,749**	−675	24,057	23,382	−2.8
New York	4,928**	−196	34,326	34,131	−0.6
North Carolina	−1,752**	−1,043	28,106	27,063	−3.7
North Dakota	−7,072**	3,002	22,705	25,707	13.2
Ohio	−1,478**	744	28,316	29,060	2.6
Oklahoma	−4,721**	564	25,106	25,670	2.2
Oregon	−2,158**	−1,053	27,643	26,590	−3.8
Pennsylvania	−722	−184	29,086	28,902	−0.6
Rhode Island	528	−45	30,294	30,249	−0.1
South Carolina	−1,485*	−559	28,317	27,758	−2.0
South Dakota	−6,042**	3,130	23,733	26,863	13.2
Tennessee	−2,854**	1,536	26,941	28,477	5.7
Texas	−1,530**	1,090	28,268	29,358	3.9
Utah	−4,104**	−1,285	25,700	24,415	−5.0
Vermont	−3,713*	−3,255	26,071	22,816	−12.5

Table 2. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Virginia	2,653**	−2,015	32,407	30,391	−6.2
Washington	1,269*	552	31,004	31,556	1.8
West Virginia	−6,323**	−590	23,482	22,893	−2.5
Wisconsin	−1,985**	−143	27,822	27,679	−0.5
Wyoming	−6,082**	2,605	23,692	26,298	11.0

Source: Census 2000 Supplementary Survey Public Use microdata file
$N = 207{,}608$
$*p < .01$
$**p < .001$

additional adjustment was made to correct for this nonindependence (U.S. Census Bureau 1998–2000).

5. Results

Tables 1-4 are formatted similarly and display the results from the regression analyses for the 1990 Census earned income, C2SS earned income, CPS earned income, and CPS health insurance respectively. The first two columns of each table include the estimates of β_1 and β_3 for each of the 50 states and the District of Columbia. The third column is the predicted value of someone who lives within the state and had his or her response imputed forcing $\beta_3 = 0$. The fourth column is the predicted value using the estimated coefficient for β_3. The final column is the percent change from the predicted score where β_3 is set to zero, to the predicted score when the parameter estimate for β_3 is used.

Table 1 provides an overview of the estimates from Model 1 using the 1990 Decennial Census for income. As expected, many of the state effects β_1 are statistically significant, demonstrating that the average earned income for reported cases in most states differs significantly from the national average. As expected with the Decennial Census data, only 3 of the 51 states had significant interaction effects β_3 and there is little support for Hypotheses 1 and 2 the bias pattern. The states with significant effects were three of the largest: Texas, California, and Florida. To check whether these effects were the result of the extremely large sample size of the Decennial Census, we ran the model through simulations using the same sample size as three years of CPS data and none of these three states had significant interaction effects β_3 in these simulations. (The results of these analyses are available from the authors upon request.) Furthermore, of the three states only Florida confirmed Hypothesis 1 of the bias pattern with a negative state direct effect β_1 and a positive interaction effect β_3.

Table 2 provides the estimates from Model 1 using the C2SS data. The results are very similar to those of the Decennial Census, with 39 of the 51 states having a significant difference from the overall national average β_1. Only California had a significant parameter estimate associated with the interaction effect that is estimating whether there is possible bias β_3. California's significant interaction effects β_3 followed the pattern predicted by Hypothesis 1. However, Hypothesis 2 was not supported because the smaller states with large direct effects β_1 did not have significant interaction effects β_3 (California had larger than average C2SS samples). Even though state is not explicitly used in the hot deck procedure, the geographic proximity preferential in donor selection within the hot deck cell keeps the bias pattern from emerging.

Table 3 provides the CPS earned income estimates. As with the Decennial Census and C2SS data in Table 1 many of the state direct effects β_1 are statistically significant. This means that many of the states' averages for reported cases differ significantly from the national average. In contrast to the Decennial Census and C2SS data, however, many more of the interaction effects are statistically significant. Nine of the 51 states had significant interaction effects ($p <= .01$). These findings support both Hypotheses 1 and 2. In each of the nine states with a significant interaction effect β_3, the effect is in the opposite direction from the state's direct effect. In line with Hypothesis 1 the states with significant estimates of β_3 have significant estimates of β_1 and in the opposite direction from

Table 3. Direct State effect, interaction effect and percent change in the predicted scores with and without the estimated interaction effect using earned income data from the 1998–2000 CPS

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Alabama	−2,697**	2,387	26,129	28,516	9.1
Alaska	2,134**	−6,107**	30,960	24,854	−19.7
Arizona	−1,674*	3,215	27,137	30,353	11.8
Arkansas	−5,715**	3,450	23,141	26,590	14.9
California	1,971**	−2,366**	30,881	28,514	−7.7
Colorado	2,135**	−3,178	30,976	27,798	−10.3
Connecticut	5,535**	−1,324	34,309	32,985	−3.9
Delaware	592	−1,120	29,420	28,300	−3.8
Dist. of Columbia	6,005**	−6,465**	34,797	28,332	−18.6
Florida	−1,341**	899	27,505	28,404	3.3
Georgia	−1,205	1,832	27,596	29,428	6.6
Hawaii	−1,380	833	27,458	28,292	3.0
Idaho	−4,882**	215	24,016	24,231	0.9
Illinois	2,312**	−2,277*	31,141	28,864	−7.3
Indiana	−2,089**	884	26,749	27,633	3.3
Iowa	−3,391**	36	25,478	25,514	0.1
Kansas	−3,135**	2,620	25,706	28,326	10.2
Kentucky	−2,466**	699	26,385	27,084	2.6
Louisiana	−1,690*	357	27,161	27,518	1.3
Maine	−3,515**	1,887	25,339	27,226	7.4
Maryland	5,722**	−6,069**	34,592	28,523	−17.5
Massachusetts	2,718**	1,791	31,376	33,167	5.7
Michigan	1,095*	−1,893	29,958	28,065	−6.3
Minnesota	330	−3,408	29,193	25,785	−11.7

Table 3. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Mississippi	−4,711**	3,473	24,134	27,607	14.4
Missouri	−1,060	−2,714	27,832	25,118	−9.8
Montana	−8,155**	3,748*	20,765	24,513	18.0
Nebraska	−4,357**	2,299	24,511	26,810	9.4
Nevada	788	1,625	29,597	31,222	5.5
New Hampshire	774	−896	29,600	28,705	−3.0
New Jersey	6,085**	−3,122*	34,781	31,659	−9.0
New Mexico	−4,940**	895	23,958	24,854	3.7
New York	2,555**	−1,031	31,257	30,226	−3.3
North Carolina	−651**	−26	28,198	28,172	−0.1
North Dakota	−7,141**	3,575*	21,760	25,335	16.4
Ohio	364	67	29,179	29,246	0.2
Oklahoma	−2,927**	1,787	25,919	27,706	6.9
Oregon	−1,484	−397	27,368	26,971	−1.4
Pennsylvania	−302	94	28,537	28,630	0.3
Rhode Island	797	3,931*	29,589	33,520	13.3
South Carolina	−1,409	−1,482	27,455	25,973	−5.4
South Dakota	−6,532**	1,078	22,375	23,452	4.8
Tennessee	−3,139**	260	25,718	25,977	1.0
Texas	−1,377**	1,473	27,427	28,900	5.4
Utah	−3,926**	2,110	24,939	27,049	8.5
Vermont	−3,368**	−115	25,495	25,380	−0.5

Table 3. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. the rest of the country	Interaction effect between living in the State and having imputed data	State income without interaction effect	State income with interaction effect	Percent change
Virginia	2,501**	1,033	31,259	32,292	3.3
Washington	1,396	−1,321	30,234	28,914	−4.4
West Virginia	−5,143**	2,325	23,725	26,051	9.8
Wisconsin	−910	2,779	27,882	30,661	10.0
Wyoming	−4,798**	1,663	24,088	25,750	6.9

Source: 1998-2000 Current Population Survey's March Supplement
$N = 211{,}546$
$*p < .01$
$***p < .001$

Table 4. Direct State effect, interaction effect and percent change in the predicted scores with and without the estimated interaction effect using health insurance coverage data from the 1998–2000 CPS

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State uninsurance rate without interaction effect	State uninsurance rate with interaction effect	Percent change
Alabama	−0.03	0.00	24.3%	24.3%	0.0
Alaska	0.18**	−0.16	28.4%	25.2%	−11.0
Arizona	0.55**	−0.37*	36.3%	28.3%	−22.0
Arkansas	0.28**	−0.31*	30.4%	24.2%	−20.2
California	0.45**	−0.22**	33.5%	28.7%	−14.1
Colorado	−0.09	0.39**	23.0%	30.6%	33.3
Connecticut	−0.50**	0.38**	16.7%	22.7%	36.0
Delaware	−0.38**	0.54**	18.4%	27.8%	51.5
Dist. of Columbia	0.07	−0.24	26.0%	21.6%	−16.9
Florida	0.25**	−0.26**	29.7%	24.5%	−17.6
Georgia	0.10*	−0.20	26.7%	22.9%	−14.2
Hawaii	−0.66**	1.16**	14.5%	35.1%	142.5
Idaho	0.22**	−0.41**	29.0%	21.3%	−26.5
Illinois	−0.21**	0.18*	21.1%	24.2%	14.6
Indiana	−0.35**	0.24	18.8%	22.7%	20.7
Iowa	−0.54**	0.00	16.1%	16.1%	−0.1
Kansas	−0.37**	−0.04	18.5%	18.0%	−2.9
Kentucky	−0.06**	−0.41**	23.8%	17.2%	−27.7
Louisiana	0.34**	−0.30*	31.5%	25.5%	−19.0
Maine	−0.23**	0.14	20.7%	23.0%	11.2
Maryland	−0.21**	0.29*	21.1%	26.3%	24.9
Massachusetts	−0.63**	0.67**	15.0%	25.7%	71.5
Michigan	−0.49**	0.55**	16.8%	26.0%	54.6
Minnesota	−0.81**	0.65**	12.8%	21.8%	70.9

Table 4. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State uninsurance rate without interaction effect	State uninsurance rate with interaction effect	Percent change
Mississippi	0.25**	−0.18	29.6%	26.0%	−12.1
Missouri	−0.52**	0.14	16.4%	18.5%	12.4
Montana	0.26**	−0.08	29.9%	28.3%	−5.2
Nebraska	−0.53**	0.21	16.2%	19.3%	19.1
Nevada	0.32**	−0.36**	31.1%	24.0%	−22.9
New Hampshire	−0.44**	0.09	17.5%	18.8%	7.2
New Jersey	−0.09*	0.12	23.2%	25.4%	9.4
New Mexico	0.53**	−0.47**	36.0%	26.1%	−27.5
New York	0.06*	0.00	25.8%	25.8%	0.3
North Carolina	−0.08	0.01	23.4%	23.6%	0.9
North Dakota	−0.16**	0.39*	21.9%	29.3%	33.8
Ohio	−0.49**	0.25**	16.9%	20.7%	22.3
Oklahoma	0.19**	−0.39**	28.5%	21.2%	−25.6
Oregon	−0.16**	0.05	22.0%	22.9%	4.2
Pennsylvania	−0.60**	0.27**	15.5%	19.3%	24.8
Rhode Island	−0.78**	0.58**	13.1%	21.2%	62.4
South Carolina	0.11*	−0.37*	26.9%	20.3%	−24.5
South Dakota	−0.31**	0.49**	19.4%	28.3%	45.8
Tennessee	−0.34**	0.19	19.1%	22.2%	16.3
Texas	0.62**	−0.39**	37.5%	28.9%	−23.1
Utah	−0.24**	0.32**	20.6%	26.4%	27.9
Vermont	−0.54**	0.37*	16.1%	21.8%	35.3

Table 4. Continued

State	State direct and interaction effects		Imputed cases only		
	Direct effect of living in the State vs. living in the rest of the country	Interaction effect between living in the State and having imputed data	State uninsurance rate without interaction effect	State uninsurance rate with interaction effect	Percent change
Virginia	−0.27**	0.38**	20.1%	26.8%	33.6
Washington	−0.29**	0.40**	19.8%	26.8%	35.6
West Virginia	0.10*	0.00	26.6%	26.6%	−0.2
Wisconsin	−0.64**	0.79**	14.7%	27.5%	86.6
Wyoming	0.05	0.09	25.7%	27.5%	7.0

Source: 1998–2000 Current Population Survey's March Supplement
$N = 397{,}618$
$*p < .01$
$**p < .001$

the lowest level of geography included in the hot deck (in this case Census region). On the face of it, Rhode Island appears to counter Hypothesis 1 because both β_3 and β_1 are positive although the direct state effect is insignificant. However, the imputed cases are much higher for Rhode Island than the modest and nonsignificant state effect because the imputed values are being drawn to the much higher Eastern Census Region average income. Supporting Hypothesis 2, 8 of the 9 states with significant interaction effects β_3 have smaller than average CPS sample sizes.

Table 4 provides the CPS health insurance coverage estimates. As with both the CPS income and the Decennial Census data presented in Tables 1 and 2, many of the direct state effects β_1 are statistically significant from the national average. This is not surprising given that we expect states to vary in their rates of health insurance coverage. Thirty-one of the 51 states have statistically significant interaction effects β_3. In thirty of the 31 states with a significant interaction effect β_3, the effect is in the opposite direction from the state direct effect β_1. The one exception of Kentucky shows a significant estimate of β_3 but an insignificant estimate of β_1. As expected in view of the lack of geographic controls for even the Census Region level, the health insurance coverage data show an even stronger bias pattern. Also small sample states with significant estimates for the state direct effect β_1 tended to have significant interaction effects β_3. This presents strong support for Hypotheses 1 and 2.

Both of our basic hypotheses received support. Few of the interaction effects were significant for the Decennial Census data because it is processed by state and therefore, state is included in the imputation process. The C2SS data are not processed by state, but donors are selected using geographic proximity preference, making it more likely that an imputed value will originate from a donor residing in the recipient's state. However, for the CPS there is strong evidence to support the predicted bias pattern due to not using state in the hot deck or using a geographic proximity preference within the hot deck cells. The analysis of the CPS health insurance coverage state estimates resulted in many more significant interaction effects β_3.

5.1. Does the bias matter?

Now that we have made the case for the existence of a bias in the CPS state estimates of health insurance coverage and earned income, we must ask whether the bias is enough to really make a difference in the overall state estimates of earned income and health insurance coverage. The results of these analyses are presented in Table 5.

Using our estimate of the total amount of bias in the state estimates for income, eight states experience a change of at least plus or minus 2 percent of the state's average earned income. Seven of the eight states had statistically significant estimates for the interaction effect β_3. Mississippi was the only state with an estimated bias over 2 percent without a significant interaction effect. States that have a statistically significant bias towards higher income of 2 percent or more are: Rhode Island (3.5 percent or $1,160), North Dakota (2.7 percent or $673), Montana (2.5 percent $606). The four states with the bias towards lower income are: Washington DC (3.2 percent or $911), Alaska (3.0 percent or $739), Maryland (3.0 percent or $866), and New Jersey (2.0 percent or $640).

Table 5. Percent of imputed cases by State and the estimated bias in the overall CPS estimates of income and health insurance

State	Percent imputed income	Percent change for imputed income	Estimated percent change in average State income due to bias	Estimated State average income bias	Percent imputed coverage	Estimated percent change from imputed coverage	Estimated percent change in State coverage due to bias	Estimated percent change in State coverage rate due to bias
U.S.	17.0	0.0	0.0	$0	11.3	0.0	0.0	0.0
Alabama	16.3	9.1	1.5	$424	10.3	0.0	0.0	0.0
Alaska	15.1	−19.7	−3.0	−$739	10.0	−14.7	−1.5	−0.3
Arizona	12.6	11.8	1.5	$451	5.4	−30.7	−1.7	−0.5
Arkansas	12.0	14.9	1.8	$475	10.2	−26.6	−2.7	−0.7
California	16.2	−7.7	−1.2	−$354	10.2	−19.8	−2.0	−0.6
Colorado	12.2	−10.3	−1.3	−$349	11.5	48.1	5.5	1.1
Connecticut	20.2	−3.9	−0.8	−$257	13.7	46.6	6.4	0.8
Delaware	17.8	−3.8	−0.7	−$191	14.0	71.3	10.0	1.6
Dist. of Columbia	17.3	−18.6	−3.2	−$911	7.9	−21.6	−1.7	−0.3
Florida	15.9	3.3	0.5	$148	10.8	−23.2	−2.5	−0.6
Georgia	15.7	6.6	1.0	$306	13.7	−18.5	−2.5	−0.5
Hawaii	19.1	3.0	0.6	$164	4.9	219.6	10.8	1.2
Idaho	13.5	0.9	0.1	$29	10.1	−33.7	−3.4	−0.8
Illinois	18.6	−7.3	−1.4	−$392	11.6	19.3	2.2	0.4
Indiana	16.8	3.3	0.6	$154	10.2	26.8	2.7	0.4
Iowa	14.8	0.1	0.0	$5	8.8	−0.1	0.0	0.0
Kansas	11.6	10.2	1.2	$335	8.5	−3.5	−0.3	0.0
Kentucky	17.8	2.6	0.5	$128	12.9	−33.4	−4.3	−0.7
Louisiana	22.2	1.3	0.3	$80	13.0	−25.5	−3.3	−0.8
Maine	15.9	7.4	1.2	$323	8.7	14.6	1.3	0.2
Maryland	17.3	−17.5	−3.0	−$866	10.5	33.7	3.5	0.6

Table 5. Continued

State	Percent imputed income	Percent change for imputed income	Estimated percent change in average State income due to bias	Estimated State average income bias	Percent imputed coverage	Estimated percent change from imputed coverage	Estimated percent change in State coverage due to bias	Estimated percent change in State coverage rate due to bias
Massachusetts	28.0	5.7	1.6	$530	14.4	96.2	13.9	1.8
Michigan	18.4	−6.3	−1.2	−$327	13.5	73.7	9.9	1.4
Minnesota	9.4	−11.7	−1.1	−$283	10.8	90.8	9.8	1.0
Mississippi	18.2	14.4	2.6	$725	9.0	−16.3	−1.5	−0.3
Missouri	16.2	−9.8	−1.6	−$397	12.3	15.2	1.9	0.2
Montana	13.7	18.0	2.5	$606	8.1	−7.3	−0.6	−0.1
Nebraska	11.8	9.4	1.1	$295	8.6	23.7	2.0	0.2
Nevada	13.5	5.5	0.7	$231	12.1	−30.2	−3.6	−0.9
New Hampshire	18.1	−3.0	−0.5	−$157	10.5	8.9	0.9	0.1
New Jersey	22.5	−9.0	−2.0	−$640	14.1	12.6	1.8	0.3
New Mexico	13.4	3.7	0.5	$125	8.2	−37.2	−3.1	−0.9
New York	23.0	−3.3	−0.8	−$229	14.5	0.4	0.1	0.0
North Carolina	19.2	−0.1	0.0	−$5	13.9	1.2	0.2	0.0
North Dakota	16.2	16.4	2.7	$673	4.9	47.8	2.3	0.4
Ohio	15.8	0.2	0.0	$11	10.8	28.2	3.1	0.4
Oklahoma	15.8	6.9	1.1	$303	11.3	−32.5	−3.7	−0.8
Oregon	15.5	−1.4	−0.2	−$61	11.1	5.4	0.6	0.1
Pennsylvania	19.9	0.3	0.1	$19	10.5	30.7	3.2	0.4
Rhode Island	26.0	13.3	3.5	$1,160	11.7	79.2	9.3	0.9

Table 5. Continued

State	Percent imputed income	Percent change for imputed income	Estimated percent change in average State income due to bias	Estimated State average income bias	Percent imputed coverage	Estimated percent change from imputed coverage	Estimated percent change in State coverage due to bias	Estimated percent change in State coverage rate due to bias
South Carolina	16.9	−5.4	−0.9	−$236	8.3	−30.7	−2.6	−0.5
South Dakota	12.7	4.8	0.6	$143	7.4	63.9	4.7	0.7
Tennessee	14.7	1.0	0.1	$39	14.3	20.9	3.0	0.4
Texas	13.8	5.4	0.7	$215	10.9	−32.4	−3.5	−1.1
Utah	13.8	8.5	1.2	$315	12.1	37.9	4.6	0.8
Vermont	16.6	−0.5	−0.1	−$19	10.4	45.2	4.7	0.6
Virginia	18.9	3.3	0.6	$201	12.0	45.9	5.5	0.9
Washington	11.9	−4.4	−0.5	−$150	9.7	48.6	4.7	0.7
West Virginia	17.7	9.8	1.7	$453	9.6	−0.2	0.0	0.0
Wisconsin	14.9	10.0	1.5	$456	9.4	119.4	11.2	1.3
Wyoming	14.2	6.9	1.0	$253	6.0	9.6	0.6	0.1

Source: 1998–2000 March Demographic Supplements to the CPS

The amount of estimated bias for health insurance coverage is more dramatic. Twentyfive states have estimated biases of at least plus or minus 3 percent from their overall estimated coverage rate. Eighteen of the twentyfive have uninsurance rates that are biased in the upward direction and 17 of these 18 have statistically significant interaction effects. The most dramatic bias is 13.9 percent in Massachusetts (an absolute change of 1.8 percent). Massachusetts is more dramatic than the other states for two reasons. First, Massachusetts has a relatively low rate of uninsurance and a percent change calculated from a base of a relatively lower rate of uninsurance makes an absolute change of 1.5 percent more dramatic. Second, Massachusetts has the second-highest rate of imputed health insurance coverage values (only New York is higher). This combination of having a relatively low uninsurance rate and having a relatively high number of imputations makes Massachusetts more susceptible to the bias. Delaware, Hawaii, Wisconsin, Michigan, Minnesota, and Rhode Island are also low uninsurace states and they all follow closely behind Massachusetts with estimated biases of over 9 percent of the original estimate.

On the other hand, the states with higher rates of uninsurance such as Kentucky, Oklahoma, Texas, New Mexico, Idaho, Louisiana and Nevada have estimated biases in the opposite direction with lower percent changes. This is because these states begin with a higher base uninsurance rate percentage and biases of 1 percent do not make as much of a relative effect on these states. All 7 of these states have an estimated bias of at least 3 percent towards lower uninsurance rates than they actually have. The percentage change ranges from 3.1 percent in Florida (an absolute change of .9 percent) to 4.3 percent in Kentucky (an absolute change of .7 percent). All of these states also had significant interaction effects.

6. Discussion

Hot deck imputation procedures can lead to substantially biased parameter estimates when three conditions are met simultaneously. First, the parameter being estimated is not explicitly considered in the imputation process. For example, the state rate of health insurance coverage is estimated without using the variable "state of residence" in the hot deck imputation process. Second, there is a considerable amount of missing data for a variable used in the estimate (e.g., over 10 percent). And third, the association between the components of the parameter being estimated is not completely explained by some combination of the variables used in the imputation process. For example, the relationship between state and health insurance coverage is not explained by some combination of factors used in the hot deck imputation process.

The CPS-based estimates of health insurance coverage and income by state meet the first two out of the three criteria. First, the variable state of residence is not used in the imputation procedure for either income or health insurance coverage, despite the fact that the variable is related to both income and health insurance coverage (Mills 2001; U.S. Census Bureau 1998). Second, over 10 percent of the income and health insurance coverage data items are imputed (see Table 4). Our assumption for evaluating the third criterion is that if the relationship between state and income/coverage is accounted for by the other variables used in the hot deck model, then we would expect that the estimates of

the interaction coefficients β_3 to be equal to zero and the bias pattern should not emerge. As our analysis demonstrated, this was not the case with the CPS data.

With the three conditions for bias being met, we have demonstrated that there is evidence of bias in the CPS state health insurance coverage and income estimates and that the bias makes a significant impact on the estimates for some states. The bias has a larger effect on the coverage estimates than on the income estimates. This is partly due to the fact that Census region (Northeast, Midwest, South, and West) is used in the income imputations but region is not used in the health insurance procedures. The income imputations, in general, are much more sophisticated than the health insurance imputations as judged by the number of variables and levels used (U.S. Census Bureau 1987; U.S. Census Bureau 1998).

There are nine states that would experience a statistically significant decline or increase in the average state income estimate from CPS if the interaction effect were actually zero. Montana, North Dakota, and Rhode Island have incomes that are biased upward. Alaska, California, the District of Columbia, Illinois, Maryland, and New Jersey have average incomes that are significantly biased downward by the current imputation practices. All nine of these states tend to be at the extreme end of the income scale relative to the rest of their Census region and/or they are small in population size. The CPS imputation procedure uses the four basic Census regions in its hot deck. Thus, we would expect that the states at the extreme of their Census region will be the ones that are most affected by the bias. Also, we would expect smaller population states with average income toward the extreme end of the Census region's income distribution to be the most affected. In small population states, the odds of receiving an imputed score from a donor within your own state is the lowest (given their relatively small population compared to the census region population). The states with the upward bias (Montana, North Dakota, and Rhode Island) are all small population states at the low end of their Census region's income distribution. Whereas the states with the downward biased income (Alaska, California, Washington D.C., Illinois, New Jersey, and Maryland) are all at the high end of their Census region's income distribution (DeNavas-Walt et al. 2001).

Examining health insurance coverage, 31 states have significant interaction effects and 25 states experience at least a three percent change in their overall estimate of uninsurance rates. In all the cases where the state direct effect is statistically significant and the interaction effect is statistically significant, the two estimates have opposite signs (see Table 3). Unlike the CPS income data, the CPS health insurance imputation procedure does not include Census region (U.S. Census Bureau 1998). If the relationship between state and health insurance coverage is not maintained through the imputation process, we expect that states differing significantly from the national average of uninsurance in their reported scores (e.g., Massachusetts, Rhode Island, and Texas) will experience a more significant change toward the national average when including the imputed scores. This happens because the imputed scores are drawn from the country as a whole and not just from people within the state. Thus we would expect a strong regression toward the mean effect for states that are above and below the national average. As can be seen in Table 3, this is the pattern that emerges from the data.

6.1. Possible solutions for bias

There are at least three solutions for fixing the current bias in the CPS estimates, each of which has its own strengths and weaknesses: (1) the bias could be modeled as done in this article and the estimates could be adjusted accordingly; (2) the U.S. Census Bureau could change its current CPS hot deck procedure to capture more of the between state variation for both income and health insurance; and (3), the U.S. Census Bureau could begin to use a multiple imputation procedure (Rubin 1996) for several of its key items.

When analysts inside and outside of the U.S. Census Bureau use Census data they could evaluate the extent to which there is possible imputation bias by using an adaptation of Model 1 used to evaluate bias in this article. The application of Model 1 would not need to be limited to state level estimates. For example, an analyst may be more interested in whether there is a bias in racial or educational estimates. The 51 states could be replaced with five racial categories. This would be especially relevant if the racial group of interest (e.g., Asian/Pacific Islanders) was not explicitly included in the hot deck. After estimating the extent of the bias, the analyst could then make the appropriate adjustments to various parameter estimates of interest. For example, an analyst could lower the Massachusetts uninsurance rate by 13.9 percent, which is the amount of the estimated bias according to Model 1.

We feel, however, that the model developed in this article is appropriate for examining whether bias exists, but that it is not the best option for correcting bias. Running these adjustments on top of doing a regular analysis will be clumsy and probably lead people to different adjustments for similar estimates. We think it is best to make the change where the bias occurs and not ex-post-facto. The other two solutions discussed involve changing the actual imputation procedures themselves.

The second possible solution to correct imputation bias is that the U.S. Census Bureau could revise its current hot deck method for items with a relatively high rate of imputation (e.g., coverage and income). The U.S. Census Bureau could alleviate the bias by using all 50 states and the District of Columbia in its imputation matrix. Although this process is used for the Decennial Census it is not feasible for the CPS because it would increase the current number of unique matching combinations by a factor of 51 for the health insurance imputation procedure and by 47 for the income imputation (50 states and DC minus the 4 regions already in the hot deck procedure). This would make the number of unique combinations or cells too large for the number of cases in the CPS. A feasible solution to reduce the bias in the state estimates is to enter more empirically based groupings of states (3-4 groups in total) into the hot deck procedure.

Both the coverage and income variable imputation procedures could use some aggregation of states along with the geographic proximity preference similar to the one used in the C2SS. Currently the CPS income imputation procedure uses Census region, but this could be modified. Instead of putting Rhode Island and New Jersey together, or Maryland and Mississippi together, aggregations could be based on trying to minimize the state variance within a grouping while maximizing the between group differences. In other words, put the high-income states into one group, the medium income states into another, and the low income states into a third group. This would also work for health insurance coverage. The Census could put all the high coverage states into one group, the medium

coverage states into another group, and the low coverage states into a third group. This would *reduce* the amount of bias in the overall state estimates while only making the matrix 3-4 times larger depending on how many state groupings are added. For income it would not increase the size of the current matrix because a four Census region grouping is already used. It would only cause the groupings to be more empirically based with high income states being put with other high income states and low income states being put with other low income states.

The U.S. Census Bureau could also consider adding a geographic proximity preference to its CPS hot deck procedure. This kept the bias to a minimum in our analysis of the C2SS data and could work in the case of the CPS data as well. One potential problem with the proximity preference is the difference in sample size between the two surveys. The CPS sample size for this period was roughly 65,000 households per year versus 890,698 households in the C2SS. (U.S. Census Bureau 2000; U.S. Census Bureau 2002). The geographic proximity preference within the hot deck cell may break down quickly with significantly smaller sample sizes within each of the states.

Finally, the third possible solution involves changing the hot deck methodology to some other form. The most logical alternative would be to use multiple imputation (Rubin 1996). Multiple imputation is preferable from a statistical standpoint but does have some practical drawbacks. As the name implies, multiple imputation does not just impute one value for each missing item, it imputes multiple values (e.g., from two to ten) using a model-based maximum likelihood approach. These imputations are then used to estimate, say, ten different models and the coefficients from these models are averaged together. This type of technique can use many more data inputs into the imputation model (for example all 51 states can be predictors). The major drawback is that it changes the way analysts interact with the data. Instead of running each analysis once, they will need to run them ten times and average the coefficients together. This would be a difficult adjustment for most data analysts. However, the payoff may be worth it for those few items on the CPS that have high (e.g., over 10 percent) imputation rates.

7. Conclusion

Even small amounts of bias in the CPS state level estimates of income and health insurance coverage are problematic because these estimates are used to distribute roughly $11 billion per year. Our analysis shows that bias is likely to exist and that the U.S. Census Bureau could take steps to eliminate it. The changes required to use multiple imputation would be great for both the U.S. Census Bureau and analysts of Census data. With this in mind, short-run changes to the current hot deck imputation procedure would be the most pragmatic and immediate solution. The U.S. Census Bureau could consider developing a grouping of states for its hot deck imputation procedure based on their relationship to the variable being imputed (high income states in one group, low income states in another group) and adding the geographic proximity preference to its CPS imputation procedure. It is more advantageous to use empirically driven categories for grouping states for both income and coverage than using no grouping or Census region. The Census should consider performing an analysis of its key imputation procedures to find those variables that explain most of the variance of a variable, but use the smallest number of unique

combinations of categories. This will make the imputation specifications of the U.S. Census Bureau more data driven. Furthermore, the geographic preference in the selection of the donor should also result in the reduction of bias. After experimenting with these changes the U.S. Census Bureau should be able to evaluate whether the agency is able to reduce the bias in the CPS state estimates. In addition, the U.S. Census Bureau should begin to develop a framework by which a multiple imputation procedure could be developed for key items.

8. References

Alexander, C.H. (1998). Recent Developments in the American Communities Survey. Presented to the Annual Meeting of the American Statistical Association. Dallas, Texas: American Statistical Association, August.

David, M., Little, R., Samuhel, M., and Triest, R. (1986). Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association, 81, 29–41.

Dalaker, J. (2001). Poverty in the United States: 2000. Washington, D.C.: U.S. Bureau of the Census.

DeNavas-Walt, C., Cleveland, R., and Roemer, M. (2001). Money Income in the United States: 2000. Washington, D.C.: U.S. Bureau of the Census.

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association, 78, 605–610.

Federal Register (2000). State Children's Health Insurance Program: Final Allotments to States, the District of Columbia, and U.S. Territories and Commonwealths for Fiscal Year 2000. Washington, D.C.: Health Care Financing Administration. 65 (101) 33638–33644.

Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds) (2001). Survey Nonresponse. New York: Wiley.

Heeringa, S., Little, R., and Raghunathan, T. (2001). Multivariate Imputation of Coarsened Survey Data on Household Wealth. In Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley.

Kalton, G. (1983). Compensating for Missing Survey Data. Ann Arbor: University of Michigan.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Little, R. and Rubin, D. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Mason, R., Lesser, V., and Traugott, M. (2001). Effect of Item Nonresponse on Nonresponse Error and Inference. In Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley.

Marker, D., Judkins, D., and Winglee, M. (2001). Large-Scale Imputation for Complex Surveys. In Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley.

Mills, R. (2001). Health Insurance Coverage: 2000. Washington, D.C.: U.S. Bureau of the Census.

National Research Council (2000). Small Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Washington, D.C.: Committee on National Statistics, National Academy of Sciences.

Rubin, D. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91, 473–489.

Ruggles, S., Sobek, M., et al. (1997). Integrated Public Use Microdata Series: Version 2.0. Minneapolis, Minnesota: Historical Census Projects, University of Minnesota.

U.S. Bureau of the Census (2002). Census 2000 Supplementary Survey: Source and Accuracy Statement 2001. Washington, D.C.: U.S. Bureau of the Census.

U.S. Bureau of the Census (2001). American Community Survey Joint Economic Edit Specification. Unpublished Document. Washington, D.C.: Housing and Household Economic Statistics, Income Branch.

U.S. Bureau of the Census (2000). Current Population Survey: Design and Methodology, Technical Paper #63. Washington, D.C.: U.S. Bureau of the Census.

U.S. Bureau of the Census (1998–2000). Annual Demographic Current Population Survey Source and Accuracy Statement. Washington, D.C.: U.S. Bureau of the Census.

U. S. Bureau of the Census (1998). Current Population Survey Health Insurance Edit and Imputation Specification. Unpublished Document. Washington, D.C.: Housing and Household Economic Statistics, Poverty and Health Branch.

U.S. Bureau of the Census (1993). 1990 Census of Population and Housing: Public Use Microdata Samples Technical Documentation. Washington, D.C.: U.S. Bureau of the Census.

U.S. Bureau of the Census (1987). Post Edit Completeness and Consistency Checks for the Work Experience, Earnings, and Longest Job Portion of the March 1988 CPS Income Rewrite. Unpublished Document. Washington, D.C.: Housing and Household Economic Statistics, Income Branch.

Received March 2002
Revised March 2004