Creating Improved Survey Data Products Using Linked Administrative-Survey Data

Michael Davern
Bruce D. Meyer
Nikolas Mittag

Davern: NORC at the University of Chicago, 55 East Monroe Street, 30th Floor, Chicago, IL 60603, Davern-Michael@norc.org
Meyer: Harris School of Public Policy Studies, University of Chicago, 1155 E. 60th Street, Chicago, IL 60637, bdmeyer@uchicago.edu
Mittag: CERGE-EI Politických vězňů 7, Prague 1, 110 00, Czech Republic. nikolas.mittag@cerge-ei.cz

Any opinions and conclusions expressed here are those of the authors and do not necessarily represent the views of the New York Office of Temporary and Disability Assistance (OTDA) or the U.S. Census Bureau. The CPS-OTDA data analysis was conducted at the Chicago Census Research Data Center by researchers with Special Sworn Status and the results were reviewed to prevent the disclosure of confidential information.

Introduction

Survey researchers catalogue the potential sources of survey errors that can influence the estimates derived from surveys (Federal Committee on Statistical Methodology (FCSM 2001). These survey errors can enter the survey at many different points and in different ways. The goal of a good survey design is to make conscious decisions about what types of error you are willing to reduce given the level of funding and the specific question you would like the data to answer. The five basic sources of error are: (1) Sampling error (2) Sample coverage error (3) Non-response error –including both unit and item non-response (4) Measurement error and (5) Processing error (FCSM 2001). In this paper we focus specifically on the improvements to survey estimate quality that are possible through data linkage by reducing the substantial amount of measurement error and bias that has been observed in critical policy relevant estimates derived from surveys in Medicaid enrollment and Supplemental Nutrition Assistance Program (SNAP) receipt. A similar approach could also be used to reduce bias from non-response.

Past research has demonstrated substantial measurement error and bias in the estimates for the policy relevant concepts of Medicaid enrollment and SNAP receipt derived from surveys. Work done on the Current Population Survey found that 43% of those linked to administrative data showing Medicaid coverage to not self-report having the coverage (false negatives). On the other hand one percent of respondents in the CPS reported having Medicaid coverage that could not be confirmed through the linkage. With the substantial portion of Medicaid enrollees not reporting Medicaid in the CPS there is a substantial overall undercount (Davern et al 2009a). Research on survey misreporting of SNAP has found that a substantial share of true recipients do not report receipt in the survey. For New York, Celhay et al. (2015) find false negative rates of 42 and 26 percent in the CPS and ACS. Meyer et al. (2014) find even higher rates in the same surveys for Illinois (48 and 32 percent) and Maryland (53 and 37 percent). On the other hand, the false positive rates (true non-recipients reporting SNAP receipt) are low at around one percent (e.g. 1.2 percent for the NY ACS), resulting in the substantial net underreporting of food assistance that is documented in Meyer et al. (2015a,b) and Meyer and Mittag (2015).

This amount of survey error and estimate bias for Medicaid and SNAP is a serious problem for the policy research community and the Federal Statistical system as these survey estimates are used for critical purposes. Medicaid and SNAP are two critical noncash benefits provided by states and funded through a federal-state partnership and they are critical for surveys to measure accurately for several reasons. First, those people who receive these benefits are better off than a similar family or individual who does not receive these benefits as they have more resources to acquire food and access to medical care. When measuring concepts like the Supplemental Poverty Measure, having accurate knowledge of who has and who has not received these benefits is critical to coming up with a complete picture of the resources that a person or family has access to provide for their needs (U.S. Census 2015). The impact of making adjustments for these noncash benefits on poverty measures can have large demographic and overall poverty rate implications for understanding who is in or not in poverty (U.S. Census 2015). The problem associated with making these kinds of adjustments for non-cash benefits is that they often rely on survey estimates known to have significant measurement error and that undercount the participation in these programs.

In addition to measuring poverty these data are critical for (1) providing general knowledge and statistics on the programs (2) evaluating these programs to see whether specific policy objectives are met over time (3) aiding official budgeting by the Congressional Budget Office as they "score" legislation and provide cost estimates for critical legislative initiatives such as the Affordable Care Act (Congressional Budget Office 2007) as well as simulation models used by federal agencies such as the Urban TRIM model (Urban Institute 2015). They are also used for official purposes by agencies to develop important health expenditure estimates for the country and states (Cuckler et al. 2013) Given these important uses of the survey data and the evidence that these data have considerable measurement error and bias it is critical that the survey research community take steps to improve the data products for the data that are used for these purposes. In this paper we use past research findings to estimate the magnitude of data quality gains that would be possible if agencies or policy research began to routinely use the partially corrected estimates that can be obtained using linked data methods.

Linking Data as a Way to Reduce Measurement Error in Estimates

One way to try to improve on the potential limitations of any data system is to combine it with other sources of data through linkage in an attempt to estimate error and minimize bias. For example, by combining survey reported data with program administrative data we can create improvements in the ultimate estimates and data products used for important policy related purposes. Linked data have been used to assess and potentially improve sample coverage (Celhay, Meyer and Mittag, 2015; Bee, Gathright and Meyer 2015), the linked data have been used to impute variables (Davern et al. 2009a), substituting administrative values for reported values (Nicholas and Wiseman,2010; Hokayem et al., forthcoming; Meyer and Mittag, 2015), supplement survey reported data (Abowd et al. 2006), and for making simple corrections (Davern et al 2009a, Mittag 2013, Schenker et al. 2010). In this paper we explore one of the potential benefits of combining administrative data with survey data by estimating bias and reducing the measurement error in survey responses.

We use methodologies that have been used by survey researchers to validate self-reported survey data against some other external or validated standards. The approach takes survey reported data and uses them in an equation to predict some external or validated standard for a sub-set of cases or all the cases that could be linked. Then the model developed using this approach is used to develop a partial correction for survey data that are not able to be linked to the external source or validated. An example of the method is Schenker et al. (2010) who start with a set of data from NHANES that has both the self-reported survey items and clinically measured items to diagnose hypertension, diabetes, and obesity. They then model the clinically diagnosed values using the self-reported values along with potential covariates of measurement error. Once the model is developed on the NHANES data they use the model to multiply impute clinical outcomes for data on which they do not have the actual clinical outcomes in the National Health Interview Survey using a model based multiple imputation methodology. Davern et al. (2009a) used a similar procedure where they linked Medicaid administrative data to earlier releases of the Current Population Survey (CPS) data and then used the model developed on these earlier years of linked data to impute an administrative data indication of Medicaid enrollment given self-reported Medicaid status and other measurement error related covariates. Mittag (2013) used a similar approach using food stamp (SNAP) administrative data linked to the American Community Survey (ACS) data to correct estimates of receipt of food stamps out of sample in the ACS. We discuss the advantages and disadvantages of these methods compared to other approaches such as direct substitution further below.

To illustrate the impact these models can have on improving the data and reducing measurement error we use a Mean Squared Error (MSE) metric of the gains in estimator quality that would be possible using linked data. Survey research often aggregates estimator bias and variance into MSE. The MSE is defined by:

MSE=Bias Squared + Variance.

The MSE of an estimator is the expected value of the square of its deviation from the true parameter of interest, so when evaluating the quality of different survey estimators, preference is given to the one with the smaller MSE. In our tables below we take the square root of the MSE or the Root Mean Squared Error (RMSE) in order to put the measure on the same scale as the original statistics. It is more accurate to say that we report an estimate of the RMSE, since our bias and variance are estimates, but, the result of such calculations is often just referred to as the RMSE.

We use findings from past administrative-survey linked research with policy relevant estimates of Medicaid and SNAP receipt to make the case that data producers should continue to invest in data linkage research. More importantly, they should start to take advantage of the large measureable improvements in survey estimate quality by creating enhancements to existing survey data products (including microdata, summary data and report tabulations) that partially correct for the known measurement errors.

Table 1 presents results from our application to Medicaid receipt. The first four columns of numbers in Table 1 are drawn from Davern et al. (2009a).[1] In that paper the authors used the 2001-2002 Current Population Survey (CPS) linked to Medicaid Statistical Information System (MSIS) data from 2000-2002 to create a person level logistic regression model of Medicaid receipt. Of the CPS respondents linked to MSIS and who show Medicaid enrollment at some point during the reference period, roughly 43% do not report having Medicaid, resulting in a Medicaid undercount (Davern et al. 2009b). However, because 47% of those linked do correctly self-report, Medicaid enrollment is a critical variable to use in the regression model. Stratifying on self-reported Medicaid status, Davern et al. (2009a) estimated two models to partially correct for survey measurement error (See Appendix A for the estimated model parameters). The first used logistic regression to predict whether a person received Medicaid in MSIS given that they did not report having Medicaid in the survey (i.e., a false-negative model). The second predicted whether a person received Medicaid given that they had reported Medicaid coverage in the survey (a true positive model). The coefficients from these two logistic regression models were used to predict each person's probability of being enrolled in Medicaid in the 2007 and 2008 CPS given their self-reported coverage and other key co-variates such as age, sex, income and state of residence (representing coverage for calendar years 2006 and 2007). This process generated a predicted probability for each person in the 2007 and 2008 CPS and these person level predicted probabilities were used to develop estimates by state of having Medicaid (details of the model are presented in Davern et al. 2009a).

The point of this reanalysis of those data is to add the last four columns below. The estimated bias is measured as the difference between the state estimate of enrollment in 2006-2007 and the Medicaid enrollment numbers found on Kaiser State Health facts. This number is likely biased as well and that bias can vary from state to state given how Kaiser compiles the estimates. Nevertheless, the numbers are an independent estimate of enrollment in those years for comparison purposes. The first column RMSEs are for the unadjusted CPS (i.e., what you would get if you simply tabulated the CPS public use file for those two years and created a two-year average). Bias is estimated as the difference between the Kaiser rate and the CPS rate. The second column of RMSEs represents the RMSEs from comparing the Kaiser rate to the CPS imputation rate based on the individual level predicted probabilities. The final column is the percent reduction (negative numbers are the percent increase) between the two RMSEs for any given state.

[1] The standard errors for the imputed Medicaid Enrollment estimates in Davern et al. (2009a) were incorrect and did not appropriately adjust for the design effect of the CPS complex sample design. The standard errors in Table 1 of this paper for imputed Medicaid by state have been adjusted using the design effect of the CPS direct survey estimates.

Table 1: Comparison of Medicaid Enrollment Estimates from our Partially Corrected Imputation Model to the Regular CPS Estimates, by Selected Characteristics and State: Average of Calendar Year 2006 and 2007

	Medicaid Enrollment Estimate - CPS		Medicaid Enrollment Estimate - Imputed		Kaiser Medicaid Enrollment	Root Mean Squared Errors (RMSE)		Percent Reduction from RMSE-CPS to
State	Percent	SE	Percent	SE	Percent	RMSE-CPS	RMSE-Imputed	RMSE-Imputed
Alabama	11.2%	0.85%	13.9%	0.93%	14.7%	3.58%	1.24%	65.39%
Alaska	7.9%	0.68%	10.3%	0.77%	11.9%	4.06%	1.77%	56.40%
Arizona	15.0%	0.98%	17.5%	1.05%	15.8%	1.26%	2.00%	-58.20%
Arkansas	15.3%	0.93%	17.4%	0.98%	17.9%	2.75%	1.10%	59.91%
California	13.8%	0.35%	16.5%	0.38%	17.7%	3.92%	1.27%	67.69%
Colorado	7.6%	0.50%	8.7%	0.54%	8.0%	0.64%	0.94%	-48.03%
Connecticut	7.9%	0.56%	9.0%	0.59%	11.4%	3.59%	2.48%	30.98%
Delaware	10.0%	0.74%	13.7%	0.85%	16.9%	6.98%	3.37%	51.71%
District of Columbia	18.5%	1.09%	20.5%	1.14%	21.8%	3.53%	1.79%	49.38%
Florida	8.3%	0.40%	11.7%	0.47%	11.6%	3.35%	0.47%	85.91%
Georgia	9.8%	0.59%	12.9%	0.67%	13.4%	3.70%	0.86%	76.67%
Hawaii	9.6%	0.65%	12.6%	0.73%	14.5%	4.96%	2.09%	57.82%
Idaho	9.9%	0.78%	10.9%	0.82%	11.3%	1.56%	0.90%	42.19%
Illinois	10.3%	0.54%	13.4%	0.60%	15.3%	5.00%	1.99%	60.20%
Indiana	10.3%	0.71%	12.5%	0.78%	12.5%	2.33%	0.78%	66.57%
Iowa	11.0%	0.67%	12.2%	0.70%	10.7%	0.71%	1.59%	-125.22%
Kansas	8.5%	0.69%	10.8%	0.77%	9.1%	0.94%	1.87%	-98.70%
Kentucky	13.6%	0.84%	14.7%	0.86%	16.7%	3.17%	2.21%	30.35%
Louisiana	12.8%	1.00%	15.6%	1.08%	20.4%	7.63%	4.85%	36.38%
Maine	18.2%	0.84%	21.6%	0.89%	19.7%	1.75%	2.10%	-20.05%
Maryland	7.0%	0.51%	8.3%	0.55%	9.5%	2.53%	1.35%	46.57%
Massachusetts	14.7%	0.86%	13.9%	0.84%	16.1%	1.60%	2.32%	-44.66%
Michigan	11.9%	0.64%	12.7%	0.65%	15.1%	3.24%	2.47%	23.95%
Minnesota	10.3%	0.60%	12.2%	0.65%	11.3%	1.20%	1.06%	12.09%
Mississippi	16.7%	1.13%	16.5%	1.12%	18.0%	1.68%	1.84%	-9.40%
Missouri	11.5%	0.72%	15.8%	0.83%	12.5%	1.23%	3.43%	-178.24%
Montana	10.7%	0.90%	6.6%	0.72%	9.2%	1.78%	2.69%	-51.01%
Nebraska	7.8%	0.66%	11.6%	0.79%	10.0%	2.23%	1.81%	18.94%
Nevada	5.2%	0.54%	7.1%	0.62%	6.8%	1.71%	0.68%	60.39%
New Hampshire	5.6%	0.42%	7.3%	0.48%	8.3%	2.77%	1.16%	58.11%
New Jersey	7.4%	0.52%	8.6%	0.56%	8.8%	1.52%	0.59%	61.33%
New Mexico	14.7%	1.01%	18.1%	1.09%	20.3%	5.66%	2.46%	56.52%
New York	15.6%	0.53%	16.2%	0.54%	21.6%	6.02%	5.36%	10.92%
North Carolina	11.9%	0.66%	16.8%	0.76%	13.3%	1.47%	3.61%	-144.80%
North Dakota	8.0%	0.68%	10.3%	0.76%	8.4%	0.76%	2.11%	-178.57%
Ohio	12.0%	0.63%	13.5%	0.66%	14.1%	2.20%	0.93%	57.91%
Oklahoma	12.3%	0.84%	15.5%	0.92%	14.7%	2.59%	1.20%	53.56%
Oregon	10.0%	0.74%	11.8%	0.80%	9.1%	1.13%	2.79%	-146.14%
Pennsylvania	9.3%	0.51%	13.3%	0.60%	15.3%	6.03%	2.10%	65.12%
Rhode Island	17.1%	0.90%	16.7%	0.89%	15.7%	1.73%	1.35%	22.11%
South Carolina	13.2%	0.87%	16.8%	0.96%	14.6%	1.60%	2.40%	-49.70%
South Dakota	8.8%	0.70%	9.8%	0.73%	11.5%	2.84%	1.88%	33.65%
Tennesse	14.1%	0.95%	22.0%	1.13%	20.6%	6.54%	1.79%	72.60%
Texas	10.9%	0.38%	13.2%	0.42%	12.0%	1.23%	1.20%	3.09%
Utah	8.0%	0.77%	9.8%	0.84%	7.2%	1.08%	2.71%	-151.51%
Vermont	17.2%	0.93%	20.9%	1.00%	19.2%	2.21%	1.95%	11.74%
Virginia	7.1%	0.52%	8.1%	0.55%	8.5%	1.46%	0.66%	54.60%
Washington	11.1%	0.69%	15.1%	0.79%	13.4%	2.36%	1.89%	20.06%
West Virginia	14.0%	0.91%	16.3%	0.97%	16.7%	2.88%	1.06%	63.38%
Wisconsin	11.5%	0.79%	12.0%	0.80%	12.2%	1.04%	0.83%	20.68%
Wyoming	7.5%	0.70%	9.0%	0.76%	10.9%	3.40%	1.99%	41.61%
Total - United States	11.4%	0.11%	13.8%	0.12%	14.3%	2.89%	0.54%	81.29%

Source: 2007 and 2008 CPS ASEC data files

* Independent Medicaid Enrollment Estimate Downloaded September 2015 from Kaiser State Health Facts Downloaded Notes: See http://kff.org/medicaid/state-indicator/monthly-medicaid-enrollment-in-thousands/ for notes and sources.

For the U.S. as a whole the RMSE for the model based imputed direct estimate is 81% lower than the RMSE for the direct CPS estimate. This is a substantial reduction in RMSE which results mainly from the bias being reduced. The direct CPS estimate is 11.4% and the imputed estimate is 13.8% which is much closer to the 14.3% in the Kaiser State Health Facts. In most states the MSE decreased between the CPS direct survey estimate and imputed estimate. There are, however, 12 states that saw an increase in bias with the imputed estimate. The largest were in Utah, Arizona, North Dakota, North Carolina and Missouri. There is not a uniform reason explaining why these states' estimates do not improve with the current model but future research can look for potential reasons and attempt to improve on the fit of the model for these states. For the state of Montana, the increase in the bias in the modeled results derives from the fact that over half of those on Medicaid were missing the linking information. Thus in Montana's case, too few people are imputed to have Medicaid as over half the enrollees were not linkable to the CPS (US Census Bureau 2008a). One way to fix this problem would be to not add a state specific fixed effect for Montana. Other reasons for the decrease seem to be states with significantly lower than the national average Medicaid enrollment rate (North Dakota and Utah) and also could be due to how the states set up their Children's Health Insurance Program--future work should examine this.

Our second illustration of how simulations based on models from validation data can improve survey estimates examines SNAP receipt for small geographic areas[2] in New York State. The results in Table 2 are similar to those for Medicaid in Table 1. They are based on the model and results in Mittag (2013), which uses administrative SNAP records linked to the ACS to develop a method of correcting survey estimates for measurement error. The validation data were created by linking administrative records on monthly SNAP payments for all recipients in New York State from the New York State Office of Temporary and Disability Assistance (OTDA) to the 2008-2012 ACS survey data. The administrative records are based on actual payments that have been validated and the two data sources are linked at the household level with a high match rate. Thus, even though they are not free of error, the linked data appear accurate enough that we consider them to be the unbiased estimate of receipt. For further descriptions of the data and its accuracy, see e.g. Celhay et al. (2015), Harris (2014), Mittag (2013) and Scherpf et al. (2014). As Celhay et al. (2015) show, the linked data reveal substantial error in reported SNAP receipt and amounts at the household level. For example, 26 percent of true recipient households do not report SNAP receipt in the ACS (false negatives). On the other hand, the false positive rate (true non-recipients reporting SNAP receipt) is low at 1.2 percent, resulting in the substantial net underreporting of government transfers that is documented in Meyer et al. (2015a,b) and Meyer and Mittag (2015).

The fifth column of Table 2 provides estimates of receipt rates and the number of recipients using the linked data that we consider to be the unbiased estimate for the 39 county groups that can be identified in the ACS public use data. Comparing receipt rates to the survey based estimates in the first two columns underlines that there is net underreporting in all but one area, and that reporting rates vary between these areas. Harris (2014) examines reporting rates at the county level in detail.

The main objective of this paper is to assess how the survey estimates compare to the results in columns three and four, which contain estimates of the receipt rate and number of recipients using an imputation model to partial correct the survey reports. The imputations are based on the method in Mittag (2013), who uses the linked ACS data to estimate the conditional distribution of administrative SNAP receipt and amounts received given reported receipt and a large set of covariates. The conditional distribution of SNAP amounts can be seen as a continuous distribution with a mass point at 0. However, we are only concerned with receipt and not with amounts received here, so we only use the estimate of the binary part of the distribution. We discuss extensions to continuous or mixed distributions below. Using the estimated parameters of this conditional distribution, we predict a probability of SNAP receipt for each household as with Medicaid above. We then generate a receipt variable by taking 20 random draws from a Bernoulli distribution with the predicted probability for every household in the New York ACS sample. Since we are interested in subgroup means here, which are consistent under classical measurement error, the estimates are

[2] We use the counties that can be identified in the public use ACS data and pool counties that cannot be separated in the public use data.

consistent with one imputation and standard multiple imputation yields the same results as the procedure we use here.[3] However, taking multiple draws makes simulation error negligible and thus avoids having to correct the SEs for it.

Table 2: Comparison of Estimated SNAP Receipt Rates in New York State and Counties From 2010 ACS Survey Reports and Imputation

	SNAP Receipt Rate Estimate - ACS		SNAP Receipt Rate Estimate - Imputed		Linked Data	Root Mean Squared Errors (RMSE)		Percent Reduction from RMSE-ACS to
Counties	Percent	SE	Percent	SE	Percent	RMSE-ACS	RMSE-Imputed	RMSE-Imputed
Albany	11.6%	1.49%	15.6%	1.66%	14.6%	3.3%	1.9%	40.85%
Allegany, Cattaraugus	14.6%	1.83%	18.4%	1.97%	19.3%	5.0%	2.2%	56.95%
Bronx	41.1%	1.02%	47.7%	1.04%	52.1%	11.0%	4.5%	58.94%
Broome, Tioga	15.3%	1.59%	19.3%	1.80%	18.8%	3.9%	1.9%	51.69%
Cayuga, Madison, Onondaga	12.6%	0.92%	16.6%	1.03%	16.4%	4.0%	1.0%	73.98%
Chautauqua	15.9%	1.94%	19.2%	2.18%	21.5%	5.9%	3.2%	46.48%
Chemung, Schuyler	17.1%	2.35%	19.4%	2.47%	21.7%	5.2%	3.4%	35.55%
Chenango, Cortland	18.2%	2.28%	20.6%	2.47%	19.6%	2.7%	2.7%	1.77%
Clinton, Essex, Franklin, Hamil	16.9%	2.05%	19.7%	2.17%	19.2%	3.1%	2.2%	27.05%
Columbia, Greene	9.4%	2.26%	14.5%	2.72%	10.2%	2.4%	5.0%	-107.30%
Delaware, Otswego, Schoharie	10.8%	1.61%	16.0%	2.06%	15.1%	4.6%	2.3%	50.68%
Dutchess	8.5%	1.34%	12.4%	1.62%	11.3%	3.1%	2.0%	36.72%
Erie	17.4%	0.95%	20.5%	1.00%	20.9%	3.6%	1.1%	70.42%
Fulton, Montgomery	14.6%	2.07%	19.5%	2.63%	28.3%	13.8%	9.2%	33.37%
Genesee, Orleans	11.3%	2.26%	15.6%	2.51%	16.4%	5.6%	2.7%	52.74%
Herkimer, Oneida	16.6%	1.55%	19.6%	1.65%	21.8%	5.4%	2.8%	48.86%
Jefferson, Lewis	19.1%	2.39%	22.0%	2.45%	19.5%	2.4%	3.4%	-41.32%
Kings (Brooklyn)	26.1%	0.68%	32.8%	0.73%	35.9%	9.9%	3.2%	67.74%
Livingston, Wyoming	12.2%	2.42%	16.1%	2.64%	13.8%	2.9%	3.5%	-19.51%
Monroe, Wayne	14.2%	0.82%	18.4%	0.92%	18.1%	4.0%	1.0%	75.95%
Nassau	4.2%	0.42%	9.5%	0.67%	7.0%	2.8%	2.6%	9.78%
New York (Manhattan)	16.5%	0.69%	20.3%	0.73%	21.2%	4.7%	1.2%	75.07%
Niagara	15.9%	1.79%	18.7%	1.90%	19.4%	3.9%	2.0%	48.05%
Ontario	8.6%	2.78%	13.0%	2.98%	10.6%	3.4%	3.8%	-12.09%
Orange	11.1%	1.44%	17.2%	1.85%	11.0%	1.4%	6.4%	-342.91%
Oswego	18.1%	2.63%	20.9%	2.71%	22.1%	4.8%	3.0%	37.36%
Putnam, Westchester	6.0%	0.62%	10.6%	0.81%	9.4%	3.4%	1.5%	56.78%
Queens	17.2%	0.65%	24.4%	0.74%	23.9%	6.7%	0.9%	86.91%
Rensselaer	14.7%	2.32%	17.5%	2.42%	17.1%	3.3%	2.4%	26.25%
Richmond (Staten Island)	11.0%	1.21%	15.5%	1.41%	16.9%	6.0%	1.9%	67.51%
Rockland	13.8%	1.59%	18.1%	1.78%	15.1%	2.0%	3.5%	-68.77%
Saratoga	8.0%	1.65%	10.6%	1.76%	9.3%	2.1%	2.1%	-2.12%
Schenectady	10.2%	1.93%	13.9%	2.16%	17.3%	7.3%	4.0%	45.77%
Seneca, Tompkins	11.0%	2.70%	15.5%	2.91%	13.4%	3.6%	3.6%	-0.68%
St. Lawrence	16.4%	3.41%	20.3%	3.44%	21.5%	6.2%	3.6%	41.12%
Steuben, Yates	10.7%	1.72%	14.1%	1.90%	18.6%	8.1%	4.9%	39.79%
Suffolk	5.6%	0.53%	10.7%	0.73%	9.1%	3.6%	1.7%	51.93%
Sullivan, Ulster	13.1%	1.70%	17.4%	1.93%	18.1%	5.2%	2.0%	60.84%
Warren, Washington	11.1%	1.78%	13.8%	1.95%	15.9%	5.1%	2.8%	44.51%
Total - New York State	16.1%	0.20%	21.1%	0.22%	21.4%	5.3%	0.4%	92.86%

Note: Source is the 2010 American Community Survey. The measure of truth in the first two columns and the parameters of the imputation model are from NY OTDA administrative data linked to the 2010 ACS. RMSE is $((\text{estimate-truth})^2+\text{Var(estimate)})^{0.5}$

The last three columns of Table 2 contain RMSE defined the same way as for Medicaid above. We compute the bias in the survey and imputation based estimates as the difference in the numbers from the linked data in the fifth column. Thus, contrary to the Medicaid application, the imputation model has been estimated using the same sample. Mittag (2013) further discusses extrapolation across time and geography. Our main statistic of interest is the percent reduction in RMSE when replacing the survey reports by the imputations in the last column, i.e. by how

[3] As discussed in Mittag (2013), correlations and model parameters as in Schenker et al. (2010) are inconsistent under single and standard multiple imputation, but the methods discussed here yields consistent estimates.

much the imputations reduce error compared to uncorrected survey based estimates. The numbers for the entire state of New York in the last row show that the imputation procedure reduces RMSE by an impressive 93 percent. This is similar in magnitude to the reduction in RMSE for Medicaid and again mainly driven by the reduction in bias. The survey understates receipt by 25 percent, while the imputations fall short of the actual number of recipients by 1 percent only. Standard errors are of a similar magnitude, but slightly higher for the imputation.

This pattern also drives the results at the local level. The survey numbers underestimate receipt rates in all but one county, while the imputation based numbers do not seem to be systematically biased. They are larger than the true numbers in 21 out of 39 areas and smaller in 18 areas. While the standard errors are slightly larger than in the survey, the reduction in bias more than makes up for this. Consequently, the imputation based rates are more accurate than the survey in terms of estimated RMSE in 31 out of 39 areas. The reductions in RMSE are substantial: In 29 of these 31 areas, RMSE is reduced by 25 percent or more, and in 15 areas the imputation based measure cuts the error by more than half. However, RMSE of the imputed receipt rate is larger than the survey RMSE in 8 of the 39 areas. Note that this result is primarily due to the fact that the survey closely replicates the numbers from the linked data for these 8 areas, i.e. it is mainly driven by the good performance of the survey.

Limitations of the Analysis

The Davern et al. (2009a) model does not account for variance added as a result of imputation so that the MSEs for the imputed model are too small, though the amount of variance due to imputation modeling will be minimal relative to the reduction in bias. Future enhancements to the method should account for imputation model variance through using, for example, multiple imputation (Shenker et al 2007; 2010; Rubin 1996) and corrections for the fact that the parameters of the imputation model are estimates (Murphy and Topel, 2002).

The Kaiser State health Facts estimates of Medicaid enrollment are not measured without bias and each state has different ways they compile the data for Kaiser.

There is not perfect concept alignment between the CPS measure and the Kaiser measure (the Kaiser measure is an average monthly enrollment and the CPS is a measure of Medicaid enrollment at any point in the last year). In general this would mean the administrative data counts should be even higher than the Kaiser counts.

Universes between CPS and Kaiser are not the same. Kaiser includes people in group quarters and who may have died during the year who would not be counted in CPS. The impacts of these adjustments are important although will not significantly impact the findings of the paper (see US Census Bureau, 2008 to better understand the magnitude).

The model was created using 2000-2001 MSIS data linked to 2000-2002 CPS data and was applied to microdata from the 2007-2008 COPS. Several states experienced changes in their Medicaid program over this time span leading to some (but not all) of the anomalous findings. In addition in many states the State Children's Health Insurance Program (CHIP) was also changing and can often be confused or misreported as Medicaid coverage (Plotzke et al 2011).

One final limitation is that these types of techniques are only useful when there is an administrative data source with high quality linking variables available to link to the survey data.

Discussion

In this discussion we address the immediate advantages and disadvantages of the model based imputation approach using linked survey and administrative data. We then look at this approach to reduce measurement error and compare it to the survey costs and error reduction achieved through other commonly used approaches to reduce survey MSEs and address survey error.

Policy researchers and survey researchers have advocated approaches to improve survey data using linked data. One approach that has been considered is the direct substitution of administrative data for survey data. In this case, instead of asking the survey respondent whether they have received food stamps, the receipt indicator would come directly from the administrative data. While this approach has advantages such as accuracy and the potentially better

maintenance of correlations between variables it also has disadvantages that may include timeliness (the speed at which the linkage can be done can sometimes delay the overall release of the data and estimates), and confidentiality of the survey data. Someone with access to the administrative data could more easily identify individuals on the survey data with such linkage and therefore access to the linked data is usually restricted. Our approach is not a direct substitution but a model, so that it allows for some uncertainty in imputed estimate ensuring a higher level of confidentiality. The model based imputation allows for models to be developed and improved on older vintages of the data and then implemented quickly on new data and could be incorporated toward the end of the processing/editing system assuming the program itself and the mechanisms that result in measurement error do not change significantly over time (an assumption that needs to be continually evaluated). A final advantage of the modelling approach is that if the data production agency (e.g., the Census Bureau) did not want to produce these imputed estimates due to increased cost and complexity in producing and processing the data they could simply produce model coefficients based on linked data (similar to those used by Davern 2009b; and Mittag 2013 that are included in Appendix A) that policy researchers could use to create their own imputations and edits (and could potentially be distributed through data systems such as IPUMS). [4]

Beyond the binary variables explored in this paper on SNAP receipt and Medicaid enrollment these types of models could be used to partially correct measurement error in amounts received or other continuous variables as well. Mittag (2013) imputes both receipt and amounts received by estimating an otherwise continuous distribution with a mass point at zero. This estimation can be done by combining a Probit-type model for being a recipient (i.e. being at the mass point) with a continuous model such as the truncated normal model in Mittag (2013). One could also use the simpler approach of estimating a regression model for amounts in addition to the take-up model of receipt in the validation data. See Scholz, Moffitt and Cowan (2009) for further discussion of this approach and an application to receipt of transfer programs, but without access to validation data. The work we referenced earlier by Schenker et al. (2007 and 2010) explored these types of models for continuous variables such as height, weight and body mass index as well.

Comparisons to Other MSE Reduction Approaches

Most survey researchers use a standard set of tools to ensure a quality survey data collection. In this vein, to reduce non-response bias we need to increase response rates, and we need to post-stratify the data to known census control totals; we need to reduce item non-response and impute missing data using high quality imputation procedures; to reduce coverage bias we need to increase coverage; to reduce measurement error we need to conduct record check studies, compare estimates to alternative sources, and conduct cognitive interviews and pre-test; to reduce processing error we need to check for errors being introduced during processing – making sure input data match output data and errors are not introduced during editing, weighting, imputation, and disclosure editing processes. Sampling error is measured using survey sample design variance estimates and is reduced by increasing the sample size and decreasing the design effect.

Table 3 below highlights these approaches along with a rough assessment of the cost associated with reducing the MSE using these methods. It is not meant to be an exhaustive list nor provide excessive detail on how these corrections are implemented but it is meant to provide the context for understanding where data linkage used to reduce measurement error can fit in with other commonly used approaches.

[4] As a final note on modelling, to have high quality models both Mittag (2013) and Davern et al. (2009a) found the most important predictor variable for these imputation models to be the self-reported indicator of receipt or enrollment from the survey data. For these models to work well it would be important for agencies collecting the data to retain a minimal set of indicator items on the survey despite the high level of known measurement error.

Table 3: Relative Cost of Selected Commonly Used Corrections for Survey Errors

Type of Survey Error	Correction	Rough Relative Cost ($ to $$$$$)
Sampling	increase sample size	$$$$$
	reduce design effect	$
Non-response (item and unit)	increase Response Rate	$$$$$
	impute item missing data	$
	Post-stratification weights	$
Coverage	increase coverage	$$$$$
	listing	$$$$
Measurement	cognitive interviews	$
	validate against other source	$
	link to other data	$
Processing	transformation data checks	$
	data disclosure editing	$
	metadata checks	$
	variable output checks	$
	imputation checks	$
	editing checks	$

Reducing sampling error is possible by increasing sample size but this option is expensive and grows less effective at reducing variance with each additional case that is added. On the other hand, decreasing the design effect of a survey can be a very cost effective way to reduce variance and MSEs. A very commonly employed approach to reducing the MSE is to attempt to reduce unit level non-response in surveys. Attempting to reduce unit level non-response has come under considerable scrutiny as of late as it is costly and there is little evidence it improves accuracy. We have learned that spending a considerable amount of project funds on strategies aimed at increasing response rates (through working sample hard and incentives etc.) can increase response rates. However, survey research is concerned with response bias and not response rates. These expensive efforts have demonstrated little impact on final estimates and non-response bias (Groves 2006; Groves et al. 2008). Survey researchers were optimizing the intermediate measure of response rate but it had little demonstrated impact on the ultimate measure of response bias (Groves 2006; Davern et al. 2010; Davern 2013). In addition to addressing the problem of misreporting, evidence from linkages to administrative data can also reassure us that unit nonresponse bias is small for key policy relevant variables such as income (Bee, Gathright and Sullivan, 2015; Celhay, Meyer and Mittag, 2015).

In light of the recent response rate research findings, additional work should also be done on other expensive attempts to reduce bias in estimates (such as coverage error) to make sure the costs of reducing coverage error is justified by a reduction in bias. We know that survey listing operations, for example, lead to improved coverage. But the question is how significantly does improved coverage reduce non-response bias?[5] From our two analysis of Medicaid and Food Stamps we argue that in the realm of survey errors that (a) we can do something about and (b)

[5] Data linkage to administrative data can facilitate other survey improvements besides reducing measurement error. For example, there is strong evidence that linking the sample frame to other sources of data can help surveys more efficiently allocate resources used in household listing (Montaquila 2011).

have a measureable impact on increasing the data quality (as measured by Mean Squared Error), reducing measurement error through linkage of administrative data to survey data seems to be an attractive area for achieving substantial MSE reductions. And the cost of such measures is low compared to other approaches, so that it would seem funds to pay for the linkage programs and modeling could be pulled from efforts that cost substantially more but do not have the same MSE impact such as overly aggressive measures to increase a survey's response rate.

Current Infrastructure to Support this Work Exists but Needs Enhancement

Federal statistical agencies routinely get administrative data from agencies that run administrative programs for linkage purposes. This arrangement works well for some programs which are operated by the federal government like social security and Medicare. However, many programs like TANF, Food Stamps, Medicaid and unemployment insurance are state run and acquiring the data from all states will require intense efforts over many years (with a good example being the LEHD program). While some of these state programs have useful national level data systems such as Medicaid, many of them do not (such as TANF). Thus, the bottom line is that the infrastructure exists, and some data linking and sharing is occurring. Also the modelling is advanced as well. What has been lacking is the incorporation of the results from this research into the most widely used and circulated data products produced by the Federal Statistical Agencies. In our opinion the additional funds needed to make this happen should be invested and will pay off not only in terms of higher quality estimates but also will allow critical policy research organizations such as the Congressional Budget Office, the Congressional Research Service and the Office of the Actuary at CMS to have access to better estimates and microdata as they score legislation and forecast costs of programs into the future. The CBO and the CMS Office of the Actuary already use the results of Davern et al. (2009a) in their modeling but they would appreciate the consistent production of these kinds of estimates and data sets year after year, rather than having them as one-off research projects.

Conclusion

The federal statistical community should do more to correct for known survey measurement error. It is convenient to create official estimates of uninsurance, Medicaid enrollment and food stamp participation based on data from a single survey (e.g., the Current Population Survey or the American Community Survey). We know these data products have pronounced measurement error for policy relevant variables and we have also developed approaches that allow analysts to partially correct the measurement errors. The examples of Medicaid and SNAP receipt underline that the improvements can be substantial as they reduce RMSE by 81 and 93 percent compared to estimates based on the survey data. The corrections we propose do not compromise confidentiality of the data, privacy of the respondents or violate the terms of the data sharing agreements among the agencies. We know that all data (including survey and administrative data) have errors. However, it is critical we move beyond acknowledging the data's limitations and begin to create new data products which blend the strengths of each data system in innovative ways to correct for known errors in one or the other set of data. We need to use innovative methods to mitigate the flaws in any one data system to make better public policy related decisions.

The reasons why it is now imperative for the Federal Statistical system to use linked data in the creation of official statistics, reports and data products are (1) the foundational research for use of linked administrative data and survey data has been conducted for several potential sources (2) there is clear evidence from these research projects studying linked survey and administrative data that the amount of bias due to measurement error in the survey data responses could be significantly reduced (3) the unit level and item level non-response to household surveys is growing over time putting more pressure on our models (e.g., post-stratification adjustments) that adjust for unit level non-response and impute missing item data (4) a substantial sum of funds is being spent on surveys to reduce unit level non-response through expensive/aggressive follow-up that have demonstrated little improvement in reducing bias (5) the necessary infrastructure for sharing data among federal agencies and directives have been supplied by the Office of Management and Budget (Burwell 2014). Now is the time to start building the data products that use administrative data in production as it will improve official statistics, reports and data products. While not all linked administrative data and survey data are ready for production we believe that there are substantive areas of policy research (Medicaid enrollment, Medicare enrollment, SNAP, Social Security, Public Assistance, and uninsurance) that have needed agreements in place and ongoing linkage projects that could be leveraged for improving our ability to make policy relevant estimates to evaluate and cost out policy proposals.

References

Abowd, J. M., Stinson, M., and Benedetto, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project. *U.S. Census Bureau Working Paper*.

Alexander, J. Trent, Michael Davern and Betsey Stevenson. *(2010)* "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly.* 74 (3): 551-569.

Bee, C. Adam, Graton Gathright, and Bruce D. Meyer (2015), "Bias from Unit Non-Response in the Measurement of Income in Household Surveys." University of Chicago working paper.

Burwell, Sylvia. 2014. "M-14-06: MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: Guidance for Providing and Using Administrative Data for Statistical Purposes." Office of Management and Budget. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

Congressional Budget Office. 2007. "Background Paper: CBO's Health Insurance Simulation Model A Technical Description." Congressional Budget Office. October 2007. Washington DC. https://www.cbo.gov/publication/19224?index=8712

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2015. "Measurement Error in Program Participation." Working Paper.

Cuckler, Gigi, and Andrea Sisko. 2013. "Modeling Per Capita State Health Expenditure Variation: State-Level Characteristics Matter." *Medicare and Medicaid Research and Review.* 3(4):E1-E24. https://www.cms.gov/mmrr/Downloads/MMRR2013_003_04_a03.pdf

Davern, Michael, Holly Rodin, Timothy J. Beebe, and Kathleen Thiede Call. *(2005)* "The Effect of Income Question Design in Health Surveys on Family Income, Poverty and Eligibility Estimates." *Health Services Research.* 40(5):1534-1552.

Davern, Michael, Holly Rodin, Kathleen Thiede Call, and Lynn A. Blewett. *(2007)* "Are the CPS Uninsurance Estimates Too High? An Examination of Imputation." *Health Services Research*. 42(5): 2038-2055.

Davern, Michael, Jacob Alex Klerman, David Baugh, Kathleen Call, and George Greenberg. *(2009b)* "An Examination of the Medicaid Undercount in the Current Population Survey (CPS): Preliminary Results from Record Linking." *Health Services Research.* 44(23) 965-87.

Davern. Michael, Jacob Klerman, Jeanette Ziegenfuss, Victoria Lynch, and George Greenberg. *(2009a)* "A Partially Corrected Estimate of Medicaid Enrollment and Uninsurance: Results from an Imputational Model Developed off Linked Survey and Administrative Data." *Journal of Economic and Social Measurement.* 34(4):219-240.

Davern, Michael, Donna McAlpine, Timothy J. Beebe, Jeanette Ziegenfuss, Todd Rockwood and Kathleen Thiede Call. *(2010)* "Are Lower Response Rates Hazardous to Your Health Survey? An Analysis of Three State Health Surveys." *Health Services Research.* 45 (5): 1324–1344.

Davern, Michael. 2013. "Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research." *Health Services Research*: 48(3):905-912.

Federal Committee on Statistical Methodology (FCSM). 2001. "Measuring and Reporting Sources of Error in Surveys." Washington DC: Statistical Policy Office, Office of the Management and Budget. http://www.fcsm.gov/01papers/SPWP31_final.pdf

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (4): 646-75.

Groves, R.M., E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly*. 72: 167-189.

Harris, B. C. (2014). Within and Across County Variation in SNAP Misreporting: Evidence from Linked ACS and Administrative Records. *CARRA Working Paper #2014-05.* U.S. Census Bureau.

Kaiser State Health Facts. "2006-2007 total monthly Medicaid enrollment December avg." Downloaded from Kaiser state health facts 9/20/2015. http://kff.org/medicaid/state-indicator/monthly-medicaid-enrollment-in-thousands/ for notes and sources.

Kish, Leslie. 1965. Survey Sampling. Wiley and Sons. New York; New York.

Montaquila, Jill, Hsu, Valerie, and Brick, J. Michael. (2011). Using a match rate model to predict areas where USPS-Based address lists may be used in place of traditional listing. Public Opinion Quarterly, 75, 317-335.

Meyer, Bruce D. and Nikolas Mittag. 2015. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness and Holes in the Safety Net," NBER Working Paper 21676, October.

Meyer, B.D., Mok, W.K.C. and Sullivan, J.X. 2015a. The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences. *Harris School of Public Policy Studies, University of Chicago Working Paper.*

Meyer, B.D., Mok, W.K.C. and Sullivan, J.X. 2015b. Household Surveys in Crisis. Journal of Economic Perspectives, forthcoming, Fall 2015.

Mittag, Nikolas. 2013. "A Method of Correcting for Misreporting Applied to the Food Stamp Program." Harris School of Public Policy, University of Chicago. Chicago IL.

Nicholas, J. and Wiseman, M. 2010 Elderly Poverty and Supplemental Security Income, 2002-2005. Social Security Bulletin, Vol. 70(2).

Plotzke, Michael, Jacob Alex Klerman and Michael Davern. *2011*. "How Similar Are Different Sources of CHIP Enrollment Data?" *Journal of Economic and Social Measurement*, 36(3): 213 – 25.

Rubin, Donald B. 1996. "Multiple Imputation after 18+ years." *Journal of the American Statistical Association*. 9(434):473-89.

Schenker, N. and Raghunathan, T. E. (2007), Combining information from multiple surveys to enhance estimation of measures of health. Statist. Med., 26: 1802–1811. doi: 10.1002/sim.2801

Schenker, N., Raghunathan, T. E. and Bondarenko, I. (2010), Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. Statist. Med., 29: 533–545. doi: 10.1002/sim.3809

Scherpf, E., Newman, C., and Prell, M. (2014), "Targeting of Supplemental Nutrition Assistance Program Benefits: Evidence from the ACS and NY SNAP Administrative Records". *Working Paper.*

Scholz, J.K., Moffitt, R. and Cowan, B. 2009. Trends in income support. In: *Changing poverty, changing policies*, M. Cancian and S. Danziger,eds. Washington, D.C.: Russell Sage Foundation.

Urban Institute. 2015. TRIM3 project website, trim3.urban.org, downloaded on November 13 2015.

U.S. Census Bureau. 2015. The Supplemental Poverty Measure: 2014. P60-254, September 2015.

US Census Bureau. 2008a. "Phase II Research Results: Examining Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic

Supplement (ASEC)." US Census Bureau: Washington DC.
https://www.census.gov/did/www/snacc/docs/SNACC_Phase_II_Full_Report.pdf

US Census Bureau. 2008b. "Phase III Research Results: Refinement in the Analysis of Examining Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)." US Census Bureau: Washington DC.
https://www.census.gov/did/www/snacc/docs/SNACC_Phase_III_Executive_Summary.pdf

Actuary paper that uses the imputation model

https://www.cms.gov/mmrr/Downloads/MMRR2011_001_04_a03-.pdf

This approach has already been used by several health policy researchers to inform their work and the early working paper version was cited as a key source in the Congressional Budget Office's health reform model in 2009 and is currently being used to by the office of the actuary at CMS the CBO work can be seen here (http://www.cbo.gov/ftpdocs/87xx/doc8712/10-31-HealthInsurModel.pdf) and the CMS work can be seen here: (http://www.cms.gov/MMRR/Downloads/MMRR2011_001_04_A03.pdf). The Office of the Actuary at CMS offered the following comments:

> After studying the impact of using the unadjusted and adjusted Current Population Survey estimates of the uninsured population [using the method proposed by Davern et al 2010] in our per capita health spending model, we believe that your Medicaid enrollment adjustments represent an improvement to our analysis of health spending behavior by state. By adjusting Medicaid enrollment, the uninsured population is reduced somewhat, resulting in a smaller magnitude for the coefficient that represents the impact of the uninsured on health spending. Although both the adjusted and unadjusted estimates of the uninsured have a negative impact on health expenditures, the Medicaid adjustments from your models help us refine the magnitude of that effect.

Appendix: Model Parameter Estimates for the models used to impute Table 1 are in Table A1 and estimates used to impute Table 2 are in Table A2.

Table A1: Logistic Regression Coefficients for Those CPS Cases Without Medicaid Recorded on the CPS (Model 1) and Those Cases With Medicaid Recorded (Model 2) Predicting the Probability of Being Linked to the Medicaid Statistical Information System (MSIS) (Model 1) or Not Being Linked to the MSIS (Model 2)

Variable	Model 1	SE		Model 2	SE	
Intercept	-0.6089	0.0424	***	0.7521	0.0635	***
Age						
Age 00 - 05	1.3364	0.0391	***	0.396	0.0465	***
Age 06 - 14	0.8797	0.0359	***	0.4068	0.0554	***
Age 15 - 17	0.6517	0.0411	***	0.1538	0.0629	**
Age 18 - 44	-0.0311	0.0253		0.1553	0.0411	***
Age 45 - 64	-1.0515	0.0434	***	-0.2539	0.0578	***
Age 65 +	-1.7853	0.0669	***	-0.8579	0.0717	***
Health Insurance Allocation Status						
Medicaid Status Edited	.	.		-0.3439	0.0471	***
Health Insurance Status Imputed	0.3617	0.0174	***	-0.7819	0.0402	***
Health Insurance Status Reported	-0.3617	0.0174	***	1.1258	0.0318	***
CPS Health Insurance Codes						
Only Another Public Insurance Program Reported on CPS	1.1714	0.0383	***	.	.	
Only Private Insurance Reported on CPS	-1.0714	0.0344	***	.	.	
Other Public and Private Insurance Reported on CPS	0.0936	0.0549	***	.	.	
Uninsured Reported on CPS	-0.1936	0.0352	***	.	.	
Only Medicaid Reported on CPS	.	.		0.1033	0.0474	*
Race and Ethnicity						
Hispanic	0.1155	0.046	**	-0.0447	0.0615	
Black	0.5177	0.0364	***	0.1324	0.0606	*
American Indian	0.1917	0.0932	*	0.0797	0.1306	
Asian or Pacific Islander	-0.2467	0.0619	***	-0.00385	0.1035	
White	-0.5782	0.0341	***	-0.1635	0.0501	***
Sex						

Variable	Model 1	SE		Model 2	SE	
Male	-0.5109	0.0205	***	-0.3084	0.0358	***
Relationship to Reference Person						
Parent	0.888	0.0859	***	0.8454	0.1277	***
Spouse	-0.5062	0.0424	***	-0.6605	0.0603	***
Child	-0.2866	0.0392	***	-0.1461	0.0561	**
Other	0.1965	0.0344	***	0.1694	0.0675	**
Self	-0.2917	0.0286	***	-0.2083	0.0501	***
Income						
Zero Family Income Reported	0.2475	0.0803	***	-0.2862	0.1356	*
Ratio to Poverty Lvl 0-49%	0.3891	0.0506	***	0.4922	0.0596	***
Ratio to Poverty Lvl 050-75%	0.6237	0.0459	***	0.5247	0.0595	***
Ratio to Poverty Lvl 075-99%	0.45	0.0423	***	0.5368	0.0657	***
Ratio to Poverty Lvl 100-124%	0.1944	0.0479	***	0.1999	0.06	***
Ratio to Poverty Lvl 125-149%	0.0504	0.0437		-0.1616	0.0658	**
Ratio to Poverty Lvl 150-174%	-0.1552	0.0453	***	-0.21	0.0755	**
Ratio to Poverty Lvl 175-199%	-0.2717	0.0449	***	-0.5104	0.0832	***
Ratio to Poverty Lvl >200%	-1.2808	0.0299	***	-0.8718	0.0431	***
State						
Alabama	-0.1379	0.0838		-0.0488	0.1543	
Alaska	-0.1272	0.1283		-0.0857	0.193	
Arizona	0.0813	0.0924		0.1248	0.1968	
Arkansas	0.1515	0.1091		-0.2814	0.1458	
California	-0.124	0.0571	*	0.3479	0.0803	***
Colorado	-0.3486	0.1268	**	-0.3851	0.1663	*
Connecticut	-0.1982	0.1463		-0.7219	0.161	***
Delaware	0.2252	0.1268		0.2802	0.1787	
District of Columbia	0.0206	0.1474		-0.0589	0.1606	
Florida	-0.1452	0.0674	*	-0.0341	0.1078	
Georgia	-0.3799	0.1081	***	-0.2252	0.1415	

Variable	Model 1	SE		Model 2	SE	
Hawaii	0.2828	0.1152	**	0.0564	0.1883	
Idaho	-0.2137	0.1245		-0.1441	0.1475	
Illinois	0.1144	0.0772		-0.1066	0.1135	
Indiana	0.1683	0.0907		-0.0716	0.1313	
Iowa	0.0545	0.1058		0.348	0.2038	
Kansas	-0.3241	0.1097	**	0.2111	0.1952	
Kentucky	0.0305	0.1504		-0.2099	0.1682	
Louisiana	-0.1636	0.0813	*	-0.571	0.1666	***
Maine	1.18	0.0842	***	0.8533	0.1544	***
Maryland	-0.4281	0.1426	**	-0.8764	0.2201	***
Massachusetts	0.2211	0.1252		-0.1872	0.1296	
Michigan	-0.1803	0.086	*	0.0434	0.1341	
Minnesota	0.223	0.1305		0.2205	0.2103	
Mississippi	-0.3619	0.1207	**	-0.9372	0.1653	***
Missouri	0.4235	0.0936	***	0.3584	0.1665	*
Montana	-1.0005	0.1522	***	-1.6887	0.2051	***
Nebraska	0.159	0.0927		0.6703	0.146	***
Nevada	-0.6962	0.1272	***	-0.6033	0.1331	***
New Hampshire	-0.1836	0.1159		0.7746	0.144	***
New Jersey	-0.3858	0.0947	***	-0.6282	0.1425	***
New Mexico	0.1199	0.0765		0.0559	0.1575	
New York	-0.1396	0.0643	*	0.0361	0.0714	
North Carolina	0.2104	0.0876	**	0.4162	0.1437	**
North Dakota	-0.0914	0.1087		0.3506	0.1943	
Ohio	-0.0658	0.1049		0.2443	0.121	*
Oklahoma	0.08	0.1111		-0.0671	0.1658	
Oregon	-0.0195	0.0911		-0.0192	0.1641	
Pennsylvania	0.3005	0.077	***	0.5203	0.1267	***
Rhode Island	0.3507	0.1055	***	0.2558	0.1281	*

Variable	Model 1	SE		Model 2	SE	
South Carolina	0.174	0.126		0.1124	0.2257	
South Dakota	-0.1485	0.1266		-0.2581	0.1482	
Tennessee	0.9171	0.1029	***	0.9406	0.1872	***
Texas	-0.6106	0.0661	***	-0.1475	0.105	
Utah	-0.3107	0.1199	**	0.0172	0.1419	
Vermont	1.1751	0.1149	***	0.853	0.1672	***
Virginia	-0.5826	0.1316	***	-0.4431	0.1817	**
Washington	0.6428	0.0758	***	0.4109	0.163	**
West Virginia	0.3519	0.0882	***	0.2588	0.1363	
Wisconsin	-0.0958	0.1147		0.1001	0.1189	
Wyoming	-0.1949	0.1135		-0.0608	0.1771	

Source: 2001 and 2002 Expanded Sample CPS ASEC data files Linked to the 2000 and 2001 MSIS
Note: Effect coding (as opposed to dummy coding) was used for all categorical variables except for "Sex" (reference category for sex is female), "Only Medicaid Reported on the CPS" recorded on the CPS (the reference category was Medicaid and at least one other type of coverage reported on the CPS) and the Variable "Zero Family Income Reported" (the reference category was having at least some income --or loss of income reported).
***P<.001, ** P<=.01, * P<.05