# Imputation Variance Estimation Protocols for the NAS Poverty Measure: The New York City Poverty Measure Experience

Frank Potter[1], Eric Grau[2] and John Czajka[3], Dan Scheer[4] and Mark Levitan[5]
[1,2,3]Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543
[4,5]Center for Economic Opportunity of New York City, 180 Water St., New York, NY 10038

**Key Words:** Measurement of Poverty, Imputation Variance, ACS

## 1. Introduction

In a 1995 report entitled Measuring Poverty: A New Approach (Citro and Michael, 1995), the Panel on Poverty and Family Assistance appointed by the National Academy of Sciences (NAS) recommended changes to both the thresholds and the measurement of resources in the nation's official measure of poverty. The report also recommended accounting for geographic variation in the cost of living. Since then, poverty researchers have endorsed many of the panel's recommendations. Recently, the U.S. Census Bureau reported that a series of alternative estimates of poverty based on the NAS recommendations will be produced in the coming years. In the meantime, a number of individual jurisdictions, led by New York City (NYC), have begun to develop their own poverty measures that implement many of the recommendations in the report.

In addition to the cost of living for the local geographic area, the NAS panel recommended that the poverty measure account for taxes and tax credits to obtain an after-tax income value, the addition of cash-equivalent income (such as housing and nutritional assistance), and subtraction of the costs of daily living (such as housing and utility costs, work-related expenses, and medical out-of-pocket [MOOP] expenditures). With these additional factors accounted for, the new poverty measure would better reflect a family's actual resources and outlays.

The working papers of the Center for Economic Opportunity (CEO), as well as supplemental documents prepared by CEO staff, described the process for implementing the new poverty measure in NYC (CEO, 2008; Levitan et al., 2010). The CEO poverty measure is based on data collected in the American Community Survey (ACS; www.census.gov/acs/www). This ongoing survey collects extensive information about the family members' characteristics, activities (such as time to travel to work), and income. However, the ACS lacks information on key components needed for the new poverty measure. These include the federal, state, and city taxes paid and tax credits received by family members; the cash-equivalent income received from nutrition assistance programs (such as the Supplemental Nutrition Assistance Program [SNAP]); and commuting, child care, and MOOP expenditures. To incorporate these components, CEO imputes data from sources external to the ACS.

The CEO working papers describe the use of ACS data to compute the measure and show the variation in poverty across NYC based on the measure. CEO also wanted to assess the statistical significance of the variation within NYC and across years. Because of the degree to which external data are imputed to individuals and households in the ACS, CEO used measures of the random variation (1) associated with the ACS sampling design (the sampling variance) and (2) associated with assigning value from these external sources to individuals and family units in the ACS (the imputation variance). The Census Bureau assesses the sampling variance for estimates using a pseudo-replication procedure and provides replicate weights on the ACS public use files. The CEO contracted with Mathematica Policy Research to review its imputation procedures and recommend ways to assess the variation associated with these imputations.

A pseudo-replication method is available for computing the sampling variance for estimates based on the ACS data. However, there is not a clear process for estimating the variance due to the imputations, particularly since the data come from multiple sources. The purpose of this paper is to report on some of the methods that were used to estimate the variance associated with these imputations and to provide some insight into the effects of the methods for one component of the CEO poverty measure. In general, a random component was introduced into the imputation process and, by using the ACS pseudo-replication procedure, a precision estimate can be computed that includes both the sampling variance and the imputation variance.

### 1.1 Background

Poverty in the United States is measured by comparing a family's resources—defined as pre-tax cash income—to a threshold that varies by family size and is intended to reflect the level of resources required to meet basic needs. The current federal poverty thresholds, developed in the 1960s, were based on the latest survey data available at the time (the 1955 Household Food Consumption Survey from the U.S. Department of Agriculture [USDA]). From this survey, it was estimated that families of three or more spent about a third of their after-tax income on food. The cost of the USDA economy food plan ("thrifty food plan") was then multiplied by 3 to arrive at the minimum yearly income a family would need. Since then the thresholds have been adjusted for price changes using the Consumer Price Index, but not for changes in the general standard of living or the annual per capita cost of the thrifty food plan. These poverty thresholds and the aforementioned concept of family resources were designated by the Office of Management and Budget as components of the federal government's official statistical definition of poverty.[1]

In 1992, the National Research Council of the NAS received funding from Congress to appoint a Panel on Poverty and Family Assistance to conduct a study on measuring poverty. The study report was issued in 1995, and since then the NAS recommendations have gained wide acceptance among poverty researchers (Citro and Michael, 1995). The NAS poverty measure uses a broader set of needs than the official measure, taking into account the cost of clothing, shelter, utilities and "a little more" for other necessities, along with food. The new thresholds are also adjusted to account for differences in the cost of living across the nation.

---

[1] The thresholds were differentiated not only by family size, but also by farm/nonfarm status, by the number of family members who were children, by the gender of the head of household, and by aged/non-aged status. The result was a detailed matrix of 124 poverty thresholds. The matrix has been simplified over time, but otherwise the thresholds remain unchanged in constant dollars.

The NAS poverty measure also takes into account a wider variety of resources available to families. While family income is first adjusted for any federal, state, or local taxes and tax credits (to reflect an after-tax concept of income), it is supplemented by the cash-equivalent value of nutritional assistance and housing programs. Nutritional assistance can include the Supplemental Nutrition Assistance Program, or SNAP (the renamed food stamp program) and the National School Lunch Program (NSLP). Housing assistance is primarily from the Section 8 housing vouchers program of the U.S. Department of Housing and Urban Development (HUD). In addition, the cost of commuting to work, payments for child care, and MOOP expenses are subtracted from family income. For NYC, CEO adopted the NAS recommendations as a basis for establishing a new poverty threshold for the city. The new measure combines a more realistic poverty threshold with a measure of family resources that includes resources not included in the official measure. With these changes, the new measure will assess more accurately the capacity of NYC families to meet their basic needs.

## 1.2. New York City Poverty Measure

For the CEO poverty measure, CEO staff needed to first establish a new set of poverty thresholds to better reflect the cost of living in NYC. CEO then needed to use the ACS data to develop a net household income value following the recommendations of the NAS panel. Because the ACS definition of a household is different from that defined for the federal, state and city tax laws, CEO staff needed to define a household unit that matches current tax regulations. The federal, state, and city tax burden and tax credits were computed for each of these new household units using the ACS data on the household composition and characteristics to develop a after-tax income value. Finally the CEO staff developed algorithms to compute the dollar-equivalent value of other forms of income (such as nutritional assistance) and expenses (such as commuting costs, child care costs, housing costs, and health insurance and health care expenditures).

The first step in constructing its poverty measure required CEO staff to compute new poverty thresholds based on the NAS recommendations. CEO adjusted the thresholds to account for the difference in cost of living (specifically housing) in NYC, by using the Fair Market Rent value (an estimate related to the cost of renting a two-bedroom apartment) in the NYC metropolitan area as compared to the national average for a similar apartment (U.S. Department of Housing and Urban Development, 2007). This differential alone produced an increase of more than 40 percent in the portion of the threshold attributed to shelter and utility expenditures. Because such expenditures are estimated to represent about 44 percent of the threshold, the effective increase was nearly 30 percent relative to the official Census Bureau poverty thresholds.

As noted above, CEO staff then created a definition of a household unit (the minimal household unit, or MHU) based on the ACS data for simulating the federal, state, and city income tax. This definition extends the family definition used in the official poverty measure to include unmarried partners and their relatives living in the household. To construct this measure, CEO staff had to contend with deficiencies in the relationship data collected in the ACS.[2]

[2] The ACS does not identify relationships among household members except through their relationship to the householder. The CEO algorithm for defining the MHU infers relationships from the limited information collected in the survey, including age and gender.

### *1.2.1. Taxes*

CEO developed programs to simulate the taxes paid by each MHU according to the federal, state, and city tax regulations. Using the ACS income data and the characteristics of the members of the MHU, the CEO tax model computed the adjusted gross income (AGI) with the appropriate deductions (the standard and personal deductions). Based on the value of the AGI, the tax model simulation accounted for the tax credits available from each jurisdiction. These include the Earned Income Tax Credit, credits for elderly or disabled persons, and credits for children and other dependents.

### *1.2.2. Other Adjustments to Income*

Following the recommendations in the NAS report, CEO staff developed methods to include the monetary value of nutritional and housing assistance programs in after-tax income. The CEO process then deducted from this adjusted income the estimated costs for commuting, child care, and MOOP expenses, to reflect the characteristics of the MHU. The following is a brief description of some of these additions and deductions.

### *Nutritional Assistance Programs*

NYC families with low income can be eligible for one or more nutritional assistance programs sponsored by the USDA. The CEO poverty measure accounts for two: NSLP and SNAP. Under the NSLP, all schoolchildren whose family income is below 130 percent of federal poverty guidelines are eligible to receive free lunches, and those with family income between 130 and 185 percent can receive reduced-price lunches. The children must attend a public or nonprofit private school or a residential child care institution. The ACS database provides the information to determine eligibility, and cash-equivalent value of these meals can be obtained from the federal government, so the computations are fairly straightforward.

For SNAP, the process for developing a cash equivalent value is more difficult. First, the definition of a SNAP unit differs from that of the ACS household unit. The SNAP unit includes persons who reside in the same housing unit and purchase and prepare food together. An ACS household unit can include multiple SNAP units. In addition, the ACS asks only if anyone in the household received SNAP payments, which does not reveal either the number of recipients or the number of units to which they belong. Second, the ACS had previously collected data on both participation and the value of SNAP payments received by the household unit. Recently, the ACS dropped the question on the value of payments received to improve the response to the question on participation. The CEO process to develop an estimate of SNAP payments is now based on the response to the participation question combined with an imputation of the value of SNAP payments based on administrative data (NYC Human Resources Administration SNAP/Food Stamp database).

### *Shelter*

The cost of shelter is included in the poverty measure in two ways. First, the income threshold for the poverty measure is increased to account for the cost of living in NYC (as described previously). Second, the cost of rent and utilities is associated with each ACS household by matching the ACS households to households in the NYC Housing and

Vacancy Survey (HVS), conducted every three years by the Census Bureau and sponsored by the NYC Department of Housing Preservation and Development.[3]

### *Commuting Expenses*

The ACS does contain information on the primary mode of travel to work and the duration of travel. The CEO algorithm uses these data with cost estimates for travel within NYC. Travel costs include subway, bus, train, or taxi fare or the cost of using an automobile for commuting.

### *Child Care Expenditures*

The ACS does not contain information on the cost of child care. For estimating that cost, CEO uses national data from the Survey of Income and Program Participation (SIPP). Using a specific set of family characteristics available in both the SIPP database and the ACS database, CEO staff developed a series of regression models relating family characteristics to the likelihood of paying for child care and then the cost of that care. The models are then used with data from the ACS to associate a propensity to pay for child care and the cost of child care to each family.

### *Medical Out-of-Pocket Expenditures*

For MOOP expenditures, the ACS again does not contain the data needed for developing a value. The CEO measure of poverty uses data from the Medical Expenditure Panel Survey (MEPS), which collects data on health care costs over time for a national sample of households. The MEPS data are used to develop estimates for the annual out-of-pocket health insurance premium and for the percentile values of MOOP costs. The CEO procedure randomly assigns these estimates based on the MEPS database to families in the ACS database. The CEO staff are continually working on improving the imputation process and have recently researched refinements to the MOOP.

## 2. American Community Survey

### 2.1. Overview

The ACS is the basis for the income data used in the CEO poverty measure and for the development of the adjustments to those income data (U.S. Census Bureau, 2009).[4] The ACS consists of two separate samples: housing unit (HU) addresses and persons in group quarters (GQ) facilities. For the ACS, the sampling frames from which these samples are drawn are derived from the Census Bureau's Master Address File, the official inventory of known living quarters and selected nonresidential units in the United States. For the ACS, independent samples of HU addresses are selected for each of the 3,141 counties and county equivalents in the United States, including the District of Columbia.

### 2.2. Sampling Error

The complexity of the ACS sampling design and the adjustments performed on the weights result in the availability of no simple unbiased design-based variance estimators.

---

[3] Information on the HVS is available from NYC Department of Housing and Preservation and Development at www.nyc.gov/html/hpd/html/pr/vacancy.shtml and from the Census Bureau at www.census .gov/hhes/www/housing/nychvs/nychvs.html.

[4] The 2009 version of the ACS Design and Methodology report is available from the Census Bureau at www.census.gov/acs/www/SBasics/desgn_meth.htm.

To accommodate this, the Census Bureau employs the Successive Differences Replication (SDR) method (Wolter, 1984; Fay and Train, 1995; Judkins, 1990) for the ACS. The SDR method was designed to be used with systematic samples for which the sort order of the sample is informative, as in the case of the ACS's geographic sort. In the SDR method, the first step in creating a pseudo-replicate estimate is constructing the replicate factors, from which the pseudo-replicate weights are calculated by multiplying the base weight for each HU by the pseudo-replicate factor. The weighting process is then rerun to create a new set of weights. Given these pseudo-replicate weights, replicate estimates are created by using the same estimation method as the original estimate, but applying each set of replicate weights instead of the original weights. Finally, the replicate and original estimates are used to compute the variance estimate based on the variability between the replicate estimates and the full sample estimate measured across the replicates.

Given the replicate weights, the computation of variance for any ACS estimate is straightforward. Suppose that $X$ is an ACS estimate of any type of statistic, such as mean, total, or proportion. Let $\hat{X}_o$ denote the estimate computed based on the full sample weight, and $\hat{X}_1$, $\hat{X}_2$,…, $\hat{X}_{80}$, denote the estimates computed based on the replicate weights. The variance of $\hat{X}_o$, Var ($\hat{X}_o$) is estimated as a constant (4) times the sum of squared differences between each replicate estimate $\hat{X}_r$ ($r$ = 1, …, 80) and the full sample estimate $\hat{X}_o$. The formula is as follows:

(1) $$\mathrm{Var}\,(\hat{X}_o) \quad = \quad 4 * \sum_1^{80} (\hat{X}_r - \hat{X}_o)^2 / 80 \ .$$

The constant 4 is required because the SDR method is used to compute the sampling variance (Fay and Train 1995).

### 2.3. Imputations in the ACS

For the ACS, the Census Bureau uses a hot-deck imputation procedure that partitions the database of respondents into subgroups called imputation classes or cells. Although the ACS imputations are a potential source of error in the estimates, we have not incorporated any random factors to these data. As in all surveys, the income data are subject to more item nonresponse than most other variables and so have more imputed data. For purposes of the computation of the imputation variance for the CEO poverty measure, we have assumed that the ACS is fully reported.

## 3. Imputation Variance

Variance estimation that includes a component attributable to imputation is an important practical problem in survey sampling. Treating the imputed values as if observed and then applying the standard variance estimation formula often leads to the overestimation of the precision of survey-based estimates. The concept of computing a variance component attributable to imputation has been the subject of substantial research over the past 20 years (see Rubin, 1987; Rao and Shao, 1992; Fay, 1996; Rao, 1996; Kim and Fuller, 2004; Ibrahim et al., 2005; Kim and Rao, 2009). Common approaches have included the multiple imputation (MI) method of Rubin (1987), the adjusted jackknife method of Rao and Shao (1992), the population-model approach of Särndal (1992) and Deville and Särndal (1994), and the fractional imputation method of Kim and Fuller (2004).

Analysts have developed a number of procedures to handle variance estimation of imputed survey data. In particular, the MI procedure (Rubin [1987]) estimates the variance due to imputation by replicating the imputation process a number of times and estimating the between-replicate variation. The MI procedure, however, may not lead to consistent variance estimators for stratified multistage surveys in the common situation where the imputations involve multiple clusters in a multistage sample design (Fay 1991). More recently, Shao and Sitter (1996) proposed the implementation of imputation procedures independently on each bootstrap subsample to incorporate the imputation variability. Shao and Sitter (1996) proved that this method produces consistent bootstrap estimators for mean, ratio, or regression (deterministic or random) imputations under stratified multistage sampling. However, they believe that, in fact, the proposed bootstrap is applicable irrespective of the sampling design (single stage or multistage, simple random sampling or stratified sampling), the imputation method (random or nonrandom, proper or improper as defined for Rubin's MI method), or the type of estimator (smooth or nonsmooth).[5]

For the CEO poverty measure, the imputation process is somewhat different from the imputation process discussed in these sources, which treat imputation as a procedure that accounts for item nonresponse by an individual respondent. For the CEO poverty measure, imputation is based on using external data together with the ACS data to develop estimates that can be linked to the family. While similar procedures may be used (for example, nearest-neighbor matching or a regression imputation), the process is to link external data to the families in the ACS file. Because of this difference, we will be guided by the methods in these sources (in particular, Shao and Sitter, 1996), rather than using these methods explicitly. The ACS variance estimation procedure using pseudo-replicates can be viewed as somewhat comparable to the bootstrap method that Shao and Sitter (1996) proposed; this allows for both the sampling variance and the imputation variance to be computed at the same time. The addition of a stochastic error term to the process will introduce variability in the imputed values assigned to ACS family groups across the ACS pseudo-replicates and be the basis for the imputation variance. The sampling variance will be based on the variability associated with the different values of the pseudo-replicate weights across the 80 pseudo-replicates.

## 4. CEO Method for Estimating Income Adjustments

### 4.1. Introduction

The NAS report recommended including in family income the cash-equivalent of benefits received. Most non-cash benefits are related to housing assistance and nutritional assistance. For housing assistance, the primary source of non-cash benefits is the HUD Section 8 housing choice voucher program. Eligibility for a housing voucher is based on total annual gross income and family size. In general, the family's income cannot exceed 50 percent of the median income for the metropolitan area, but the public housing agency must provide 75 percent of its vouchers to applicants whose incomes do not exceed 30 percent of the median income. Costs for rent and utilities were imputed to each ACS household by matching the ACS households to households in the NYC Housing and Vacancy Survey.

[5] Smooth estimators are statistics that are generally functions of sample totals and means, whereas nonsmooth estimators are functions of order statistics, such as quantile estimates.

The second major adjustment to income is nutrition assistance, which includes NSLP and SNAP. The adjustment to income for the NSLP was based on the income and characteristics of the ACS household, the NSLP criteria for eligibility and the dollar-equivalent value for a school lunch established by the USDA. Because the dollar-equivalent value for the NSLP was based on the explicit characteristics of the household and the NSLP regulations, this component of the dollar-equivalent value of nutritional assistance was assumed to be known without error. For the SNAP payments, the imputations are based on a series of regression models and random selection and was assumed to be subject to imputation error. For the analysis of the imputation variance estimation procedures, we used the imputation of the SNAP payments.

### 4.2. CEO Methods Research

Income and program participation are often underreported in social surveys, and the ACS is no different. In the ACS, some respondents do not report their participation in SNAP. When participation is reported, respondents may understate the cash value of the benefits they have received in the prior 12 months.

Census Bureau testing of the ACS question on SNAP participation revealed that respondents were more likely to indicate receipt of the benefit if the follow-up question about the value of the benefit did not appear in the survey instrument.[6] Therefore, beginning with the 2008 survey, the ACS stopped asking for value of the benefit. Since SNAP payments are an important component of CEO's resource measure, CEO staff developed a methodology for estimating the value of those payments.

An additional problem affecting the accuracy of SNAP reporting in the ACS is that SNAP participation is measured at the household level, and the ACS household differs from a typical SNAP household. In the ACS, a "household" is comprised of all members living within the household unit, including, "the householder, occupants related to the householder, and lodgers, roomers, boarders and so forth." In contrast, SNAP family units (or cases) comprise co-resident individuals who purchase and prepare food together. The effect of these definitional differences is clearly shown in the data, where the NYC average SNAP case has 1.85 members while the average ACS household reporting SNAP payments has 2.81 members. This can result in a potential undercounting of SNAP cases, because some households may have more than one case.

To correct this undercount, CEO began by compiling administrative data on SNAP cases in NYC from the Human Resources Administration (HRA)'s internal database. The data included all cases in NYC that were active for any period between July 2006 and June 2007, a total of 769,303 cases. This process was repeated for the 2005 and 2006 surveys, using comparable June through July time periods. Consistent with the standard methodology used by CEO in its poverty measure, individuals in group quarters were removed from both the administrative data and the ACS sample.

This data set contained demographic information about the different SNAP case-heads and families, and relevant budget information such as household income, public assistance (PA) income, and monthly rent. For each case, SNAP payments for the previous year were summed. These data were used to develop a regression model using

---

[6] http://www.census.gov/acs/www/AdvMeth/content_test/H6_Food_Stamps.pdf

the demographic data—including household size, the number of children, income,[7] the presence of a household member 65 or older, and whether an elderly or disabled person headed the household—to predict the yearly value of SNAP payments of NYC families.

The regression model described above was then used to impute SNAP values through a predictive mean match (PMM) (see Little, 1988 and O'Donnell and Beard, 1999). First, the regression coefficients were used to estimate a predicted SNAP value for observations in the ACS and in the administrative data. The predicted value computed using ACS data and the predicted value computed using administrative values were matched using a nearest neighbor algorithm, whereby an ACS case would be matched with the administrative case with the closest estimated predicted value. The ACS case was then given the actual SNAP value from that administrative case. Once an administrative case donated its value to an ACS case, it was removed from the donor pool.

The advantage of using PMM rather than simply using the estimated values is that PMM preserves the actual distribution of SNAP values. Regression estimates accurately capture the mean and aggregate values of the distribution, but yield considerably less variation than seen in the actual data. To address the unit of analysis problem, CEO staff partitioned each ACS household into the maximum number of "SNAP units" that the program rules allowed. Using the SNAP unit rather than the ACS household increases the estimated number of SNAP cases in the 2007 ACS from 423,601 (55 percent of the administrative number) to 584,913 (76 percent of the administrative number).

## 4.3. Adjusting the Number of SNAP Cases in the ACS

Because of the gap between the number of SNAP cases in the administrative data and the number of reported cases in the ACS, CEO staff concluded that a number of ACS households that receive SNAP payments are not reporting them. Because it is known that SNAP participation is highly correlated with participation in other income support programs, such as PA and Supplemental Security Income (SSI), CEO staff assigned SNAP payments to individuals who were eligible for SNAP and reported PA or SSI receipt, but did not report SNAP participation.[8] This increased the number of SNAP units from 584,913 to 651,597.

## 4.4. Stochastic Error Component

Because the SNAP values were assigned using the nearest-neighbor imputation method with the predicted means, referred to as predicted mean matching (PMM) imputation, the stochastic component can be incorporated into the imputation procedures by defining a "neighborhood" for the predicted means. More specifically, the predictive mean using the HRA data can be used to define the "neighborhood" of predicted values in the data file in terms of a prespecified distance, using any distance function, from the donor's predicted mean or means, as opposed to directly using the values of the imputation covariates. This process, referred to as predictive mean neighborhoods (PMN), is discussed in Grau et al. (2004) and Singh et al. (2002). Assuming that the predicted mean for a randomly selected HRA SNAP family from this neighborhood (the donor) and the ACS family are about equal, the residual defined by the difference between this predicted mean and the

[7] Income is measured as the log of total income within the SNAP unit.

[8] Analysis of administrative data showed that roughly 80 percent of people on PA and SSI participate in the SNAP program.

observed values of the donor should approximate the residual that would have been obtained if it had been drawn from a known error distribution.[9] By using a different random start for each of the ACS replicates within a range for the predicted means (as opposed to the closest), each replicate would be expected to exhibit variation corresponding to selecting a random component from an error distribution based on the characteristics of the predictive models.

The CEO staff computed assigned imputed values for the base weight and assigned a separate set of imputations for each of the 80 ACS replicates. Thereby, each replicate weight was associated with a set of imputed values for the persons in the ACS file for NYC. The reported estimate of the poverty measure used the base weight and the base set of imputations. Following the ACS variance estimation procedure, 80 estimates (one for each replicate weight and imputation set) were also computed. The variance of the estimate was computed using the ACS variance estimation equation (see equation 1).

## 5. Imputation Variances for SNAP Imputations

The process for imputing the cash-equivalent of SNAP payments requires multiple steps and the computation of 81 sets of imputed values for the family units identified as SNAP recipients. To assess the effect of the imputation in a manner that is consistent, we chose to use an estimation model with the ACS data and the imputed values similar to the one used for the CEO poverty measure. Using the ACS data file for 2008 with the computed CEO net income and the imputed values for the SNAP payments, we estimated a pseudo poverty rate using the sum of the net income and the imputed SNAP value for each person and comparing this value to the 2008 official federal poverty thresholds. The estimates computed do not incorporate the other imputed values developed by CEO and therefore, should be considered as artificial estimates developed for this specific analysis.

The primary question about the imputations is whether the imputations change the variance estimate. Because of the effort involved in computing the 81 separate sets of imputed values, a secondary question is whether fewer separate sets of imputations could produce an equivalent measure of the imputation variation.

### 5.1. Methodology

For this analysis, we used the 2008 ACS sample data file as modified by CEO. As indicated previously, the CEO computed net income and 81 sets of imputed values. The data file also contained information on the household and demographic information on the individuals in the households in the ACS sample. These data included borough of residence, poverty unit family type, size, and number of adults and children under 18 for the family; and, for individual sample members, age, race/ethnicity, educational attainment, work experience, and citizenship status.

For the analysis, we chose to partition the sample by borough (the 5 counties comprising New York City) and age (11 categories in 5-year increments up to 24 years and 10-year increments from 25 to 74, with a 75 or older category). As shown in Table 1, the full sample size for NYC is 61,508, with the largest sample count in Brooklyn and the

[9] Bias in the estimate of the mean and the standard error can result if the predicted means are far apart. However, such bias would occur regardless of the imputation method used, since any method would be based on the same set of covariates.

smallest in Staten Island. For the city as a whole, each age category had at least 3,000 sample members. For Staten Island, the sample count was at least 200 for all age categories. By using borough and age, we had a substantial range of sample sizes.

**Table 1.** Unweighted and Weighted Sample Counts from American Community Survey

| | New York City | | Bronx | | Brooklyn | | Manhattan | | Queens | | Staten Island | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample | Weighted Sample | Sample | Weighted Sample | Sample | Weighted Sample | Sample | Weighted Sample | Sample | Weighted Sample | Sample | Weighted Sample |
| All Ages | 61,508 | 8,173,304 | 9,509 | 1,340,385 | 19,769 | 2,517,504 | 10,749 | 1,569,255 | 17,577 | 2,267,715 | 3,904 | 478,445 |
| Age 0-4 | 3,633 | 564,341 | 684 | 108,740 | 1,295 | 189,109 | 516 | 95,269 | 918 | 143,203 | 220 | 28,020 |
| Age 5-9 | 3,461 | 494,988 | 678 | 101,065 | 1,239 | 165,205 | 421 | 74,416 | 902 | 125,765 | 221 | 28,537 |
| Age 10-14 | 3,658 | 512,453 | 744 | 105,873 | 1,300 | 174,201 | 389 | 65,588 | 949 | 132,123 | 276 | 34,668 |
| Age 15-19 | 3,956 | 530,698 | 737 | 108,796 | 1,314 | 176,916 | 438 | 62,956 | 1,161 | 145,652 | 306 | 36,378 |
| Age 20-24 | 4,235 | 554,686 | 683 | 99,469 | 1,394 | 187,353 | 770 | 85,387 | 1,158 | 151,068 | 230 | 31,409 |
| Age 25-34 | 9,167 | 1,220,702 | 1,257 | 198,875 | 2,909 | 359,720 | 2,141 | 285,700 | 2,422 | 314,027 | 438 | 62,380 |
| Age 35-44 | 8,403 | 1,299,918 | 1,255 | 185,668 | 2,544 | 362,373 | 1,575 | 308,375 | 2,480 | 371,410 | 549 | 72,092 |
| Age 45-54 | 8,638 | 1,148,261 | 1,276 | 176,985 | 2,560 | 336,726 | 1,501 | 221,501 | 2,674 | 338,531 | 627 | 74,518 |
| Age 55-64 | 7,243 | 858,769 | 967 | 118,572 | 2,284 | 263,099 | 1,309 | 166,276 | 2,174 | 255,080 | 509 | 55,742 |
| Age 65-74 | 4,918 | 530,362 | 657 | 72,422 | 1,596 | 159,883 | 929 | 109,115 | 1,425 | 156,579 | 311 | 32,363 |
| ≥Age 75 | 4,196 | 458,126 | 571 | 63,920 | 1,334 | 142,919 | 760 | 94,672 | 1,314 | 134,277 | 217 | 22,338 |

## 5.2. Analysis

### *5.2.1. Changes in the estimates and the imputation variance*

The variance was estimated using the ACS procedure by computing 81 separate estimates (a base estimate and an estimate for each replicate weight). In the following tables, we show the estimate and the relative standard error (RSE). The RSE is the ratio of the standard error to the value of the estimate. The RSE represents a measure of the variation relative to the value being estimated and can be presented as a percentage. For example, for an estimated percentage of 11.9 percent for NYC (see Table 2) and an RSE of 2.1 percent, the standard error of the estimate is 0.25 percent (2.1 percent of 11.9 percent is 0.25 percent) Table 2 shows the percentage of persons with a CEO-adjusted family income below the 2008 federal poverty guidelines before and after including the SNAP imputed values for the full city and for each borough. Since the SNAP imputations represent an addition to the household income, the percentage is decreased for all boroughs and age categories.

**Table 2.** Estimated Percentages and Effect of Imputations in Estimates and Relative Standard Errors (RSE)

| | New York City | | Bronx | | Brooklyn | | Manhattan | | Queens | | Staten Island | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percentage | RSE | Percentage | RSE | Percentage | RSE | Percentage | RSE | Percentage | RSE | Percentage | RSE |
| Before SNAP Imputation | 11.9% | 2.1% | 16.6% | 5.1% | 14.6% | 4.0% | 8.7% | 5.3% | 9.8% | 5.1% | 5.7% | 11.2% |
| After SNAP Payments Added | 10.6% | 2.3% | 13.8% | 5.1% | 12.7% | 4.3% | 8.2% | 5.4% | 9.1% | 5.1% | 5.2% | 11.2% |
| With Multiple SNAP Imputations | 10.6% | 2.4% | 13.8% | 6.3% | 12.7% | 4.9% | 8.2% | 6.3% | 9.1% | 5.7% | 5.2% | 12.8% |
| Relative Change from Multiple Imputations | | 5.9% | | 22.9% | | 15.0% | | 15.6% | | 13.1% | | 13.8% |
| Relative Change Across Age Categories | | | | | | | | | | | | |
| Mean | | 13.8% | | 20.5% | | 15.7% | | 13.2% | | 14.7% | | 7.0% |
| Median | | 10.3% | | 16.1% | | 16.7% | | 4.9% | | 8.2% | | 4.1% |
| Minimum | | -3.1% | | 3.1% | | 1.1% | | 0.7% | | 1.1% | | -0.7% |
| Maximum | | 39.4% | | 48.7% | | 33.6% | | 53.3% | | 43.0% | | 34.1% |

When considering the effects of the single versus multiple imputation sets, the percentage below the poverty line is the same (lines 2 and 3 on Table 2), the RSE increased by 0.1 percent (a relative change of 5.9 percent) for the city as a whole, with an even greater percentage increase for individual boroughs (as much as 1.6 percent for Staten Island). The relative change in the RSE for each borough ranged from 13.1 percent for Queens to 22.9 for Bronx. By age group, relative change in the RSE averages 13.8 percent for the city as a whole and between 7 percent and 21 percent within the boroughs. For some age groups, the relative RSE was as much as 50 percent higher; for others, there was almost no change in the RSE. Based on Table 2, it appears that the addition of the stochastic error component in the SNAP imputations does increase the variation in the point estimates and that for some estimates the increase in the variation may be substantial.

### *5.2.2. Number of imputation sets*

The second question is whether the number of imputations can be reduced from 80 replicates, or whether fewer imputation sets can be computed and then randomly assigned to the 80 replicates. If fewer sets of SNAP imputations are computed, more time can be spent on other imputations. For this analysis, we reduced the number of imputations sets to 40, 20, 16, 8, and 4. The imputation sets were then randomly assigned across the replicates. In the case of 40 imputation sets, the 80 ACS replicates were paired (40 pairs) and one imputation set was selected and assigned to both ACS replicates. Similarly for the case of 8 imputation sets, the 80 ACS replicates were grouped into sets of 10 ACS replicates and one imputation set was selected among the 10 imputation sets associated with the 10 replicates and assigned to all 10 ACS replicates. To measure the effect of the reduction in the number of imputation sets, we computed 10 sets for each reduction to correspond to 10 different implementations of the same process.

For each implementation of the imputation process and variance estimation, a different value was computed for the relative standard error. We wanted to see how much variation exists in the RSE over repeated implementation of the same process, and we used the standard deviation of the RSEs over the 10 implementations as the measure.

In Figure 1 and Table 3, we show the standard deviation of the RSEs for the 10 implementations by borough. Staten Island has fewer people and the smallest sample size in the ACS, and also shows the greatest range in the standard deviations of the RSEs. The highest value of the standard deviation occurs when the fewest imputation sets are used. With an increasing number of imputation sets, the standard deviation declines. In Figure 2, the same analysis is done by individual age group. It is interesting to note that the greatest variation in the RSEs occurs for two age categories, ages 55 to 64 and ages 65 to 74, but the sample size for age category 55 to 64 is relatively large and the sample size for the age category 65 to 74 is larger than that for the 5-year age categories of persons under 24 years.

The primary finding is that the variance estimation is affected when few imputation sets are used, but when 20 or more imputation sets are used the variation in the RSEs may not be substantial. In other analyses (not shown), the sample size does affect the stability of the RSE and estimates based on a smaller sample size show greater variation in the RSEs. For estimates of small subpopulations, the number of imputation sets needs to be higher than that for estimates of larger subpopulations.

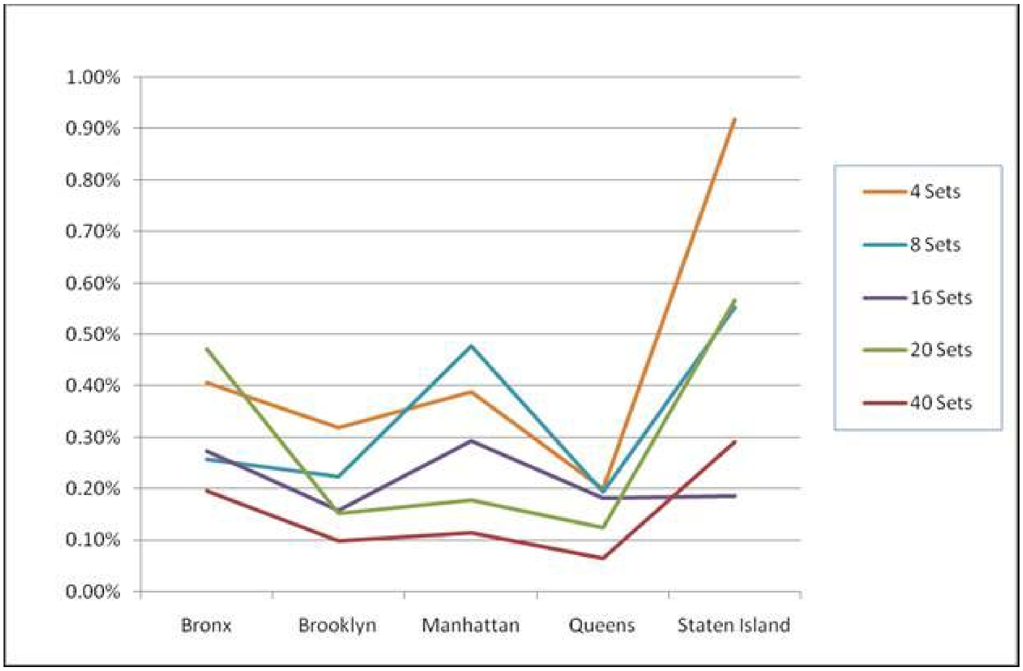


**Figure 1. Standard Deviation (STD) of Relative Standard Errors by Borough and Number of Imputation Sets (10 Replicates)**

**Table 3.** Mean and Standard Deviation of Relative Standard Errors across 10 Random Assignments

| | | 80 SETS | 40 SETS | | 20 SETS | | 16 SETS | | 8 SETS | | 4 SETS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| POPULATION | PERCENTAGE BELOW FPL[1] | MEAN | MEAN | STD DEV | MEAN | STD DEV | MEAN | STD DEV | MEAN | STD DEV | MEAN | STD DEV |
| CITY | 10.6% | 2.4% | 2.5% | 0.06% | 2.5% | 0.12% | 2.4% | 0.10% | 2.5% | 0.07% | 2.5% | 0.12% |
| BOROUGH | | | | | | | | | | | | |
| BRONX | 13.8% | 6.3% | 6.0% | 0.20% | 5.9% | 0.47% | 5.8% | 0.27% | 5.8% | 0.26% | 6.0% | 0.41% |
| BROOKLYN | 12.7% | 4.9% | 4.8% | 0.10% | 4.6% | 0.15% | 4.4% | 0.16% | 4.6% | 0.22% | 4.5% | 0.32% |
| MANHATTAN | 8.2% | 6.3% | 6.4% | 0.11% | 6.6% | 0.18% | 6.4% | 0.29% | 6.5% | 0.48% | 6.5% | 0.39% |
| QUEENS | 9.1% | 5.7% | 5.5% | 0.06% | 5.4% | 0.13% | 5.5% | 0.18% | 5.6% | 0.19% | 5.6% | 0.20% |
| STATEN ISLAND | 5.2% | 12.8% | 12.7% | 0.29% | 12.5% | 0.57% | 12.6% | 0.19% | 12.4% | 0.55% | 12.1% | 0.92% |
| AGE CATEGORY | | | | | | | | | | | | |
| AGE 0-4 | 12.8% | 6.4% | 6.5% | 0.26% | 6.3% | 0.24% | 6.0% | 0.33% | 6.2% | 0.46% | 6.2% | 0.48% |
| AGE 5-9 | 13.4% | 5.9% | 5.9% | 0.21% | 5.9% | 0.19% | 5.7% | 0.19% | 5.8% | 0.27% | 5.7% | 0.35% |
| AGE 10-14 | 12.6% | 6.4% | 6.8% | 0.25% | 6.9% | 0.15% | 7.0% | 0.30% | 7.0% | 0.39% | 7.1% | 0.41% |
| AGE 15-19 | 12.2% | 6.2% | 6.2% | 0.18% | 5.9% | 0.19% | 5.9% | 0.33% | 5.9% | 0.20% | 5.9% | 0.30% |
| AGE 20-24 | 12.3% | 6.4% | 6.4% | 0.13% | 6.5% | 0.25% | 6.4% | 0.39% | 6.7% | 0.41% | 6.6% | 0.30% |
| AGE 25-34 | 7.9% | 4.9% | 5.1% | 0.06% | 5.1% | 0.07% | 5.2% | 0.15% | 5.2% | 0.11% | 5.2% | 0.13% |
| AGE 35-44 | 7.9% | 4.8% | 5.0% | 0.14% | 5.0% | 0.17% | 5.0% | 0.30% | 5.0% | 0.16% | 4.9% | 0.20% |
| AGE 45-54 | 9.0% | 4.6% | 4.7% | 0.10% | 4.6% | 0.21% | 4.7% | 0.14% | 4.6% | 0.13% | 4.6% | 0.14% |
| AGE 55-64 | 9.2% | 6.7% | 6.6% | 0.17% | 6.6% | 0.24% | 6.8% | 0.19% | 6.4% | 0.44% | 6.4% | 0.59% |
| AGE 65-74 | 13.0% | 4.6% | 4.8% | 0.34% | 4.6% | 0.34% | 4.6% | 0.32% | 4.9% | 0.50% | 5.1% | 0.93% |
| AGE 75 OR OLDER | 16.8% | 5.7% | 5.8% | 0.10% | 5.7% | 0.14% | 5.9% | 0.14% | 5.7% | 0.27% | 5.9% | 0.37% |

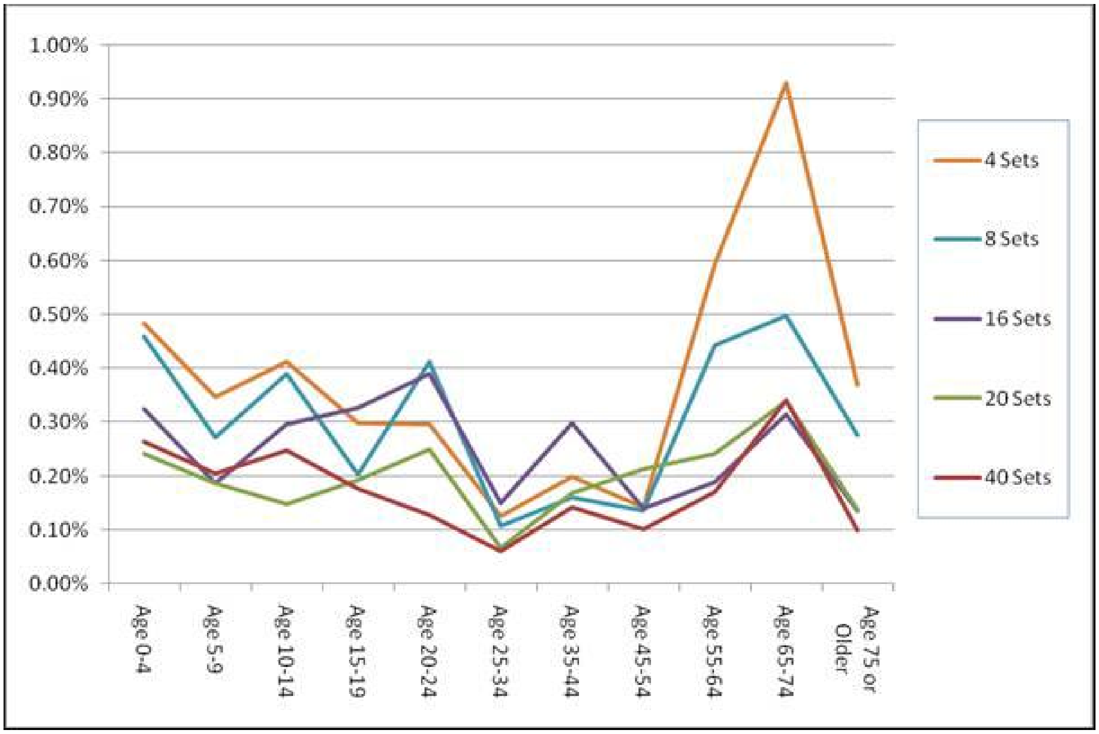


**Figure 2. Standard Deviation of Relative Standard Errors by Age Category and Number of Imputation Sets (10 Replicates)**

## 5.3. Conclusions

Mark Levitan (the director of the CEO) raised an interesting question when he and his team were working on the CEO poverty measure. While the ACS has a protocol to compute the sampling variance of the poverty measure, what about the variation that is being introduced by the imputations and adjustments used in the poverty measure computations? He came to Mathematica requesting help on possible methods to estimate the variance resulting from the uncertainty inherent with the imputations. The Mathematica team suggested procedures for most of the imputations and adjustments, and the CEO staff implemented most of these suggestions. Fortunately, attempts to systematically account for the imputations variance are becoming more frequent (see Sinclair et al., 2003).

The imputations and adjustments used in the CEO poverty measure are more extensive than often seen in sample surveys, and all are performed to compute a single estimate. In the CEO poverty measure, the imputation of SNAP payments to the income is only one of a series of imputations. Hopefully similar analyses can be performed on other imputations.

For these imputations of SNAP payments, the key findings are that by adding a stochastic error component in the imputation process, an imputation variance can be computed. The question of whether the amount of variation introduced is reasonable needs additional research by looking at the data for multiple years. It is apparent that the sample size for the estimate and the number of independent sets of imputations affect the stability of the standard error of the estimate. A possible reason for this lack of stability by using fewer imputation sets is that the imputation sets used may represent extreme values. If for one set of imputations, one or more imputation sets assign smaller values and others assign larger values, the standard error can be increased. Using more imputations sets is likely to produce a distribution of values that balance out the extreme values.

## REFERENCES

Center for Economic Opportunity. 2008. The CEO poverty measure: A working paper. New York City Center for Economic Opportunity. New York: Center for Economic Opportunity.

Citro, C.F., & Robert T. Michael (Eds) (1995). *Measuring poverty: A new approach.* Washington, DC: National Academy Press.

Deville, J.-C., & Särndal, C.-E. 1994. Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10.** 381-394.

Fay, R.E. 1991. A design-based perspective on missing data variance. *1991 Annual Research Conference Proceedings,* Washington, DC: U.S. Department of Commerce, Bureau of the Census. 429-440.

Fay, R.E. 1996. Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association,* 91. 490-498.

Fay, R., & Train, G. 1995. Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics*. Alexandria, VA: American Statistical Association. 154-159.

Grau, E., Frechtel, P., Odom, D.M. &Painter, D. 2004. A simple evaluation of the imputation procedures used in NSDUH. *2004 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association.

Judkins, D.R. 1990. Fay's method for variance estimation. *Journal of Official Statistics,* 6(3).

Kim, J.K., & Fuller, W. 2004. Fractional hot deck imputation. *Biometrika,* 91(3). 559-578.

Kim, J.K. & Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 9. 917-932.

Levitan, M., D'Onofrio, C., Krampner, J., Scheer, D. & Seidel, T. 2010. *The CEO poverty* measure: 2005-2008. New York City Center for Economic Opportunity.

Little, R.J.A. 1988. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6(3). 287-296.

O'Donnell, S, & Beard, R. 1999. Measuring MOOP (Medical out of pocket) expenditures using SIPP/MEPS. *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Rao, J.N.K. 1996. On variance estimation with imputed survey data. *Journal of the American Statistical Association,* 91. 499-506.

Rao, J.N.K., & Shao, J. 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79. 811-822.

Rubin, D.B. 1987. *Multiple Imputations for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Särndal, C.E. 1992. Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology,* 18( 2). 241-252.

Shao, J., & Sitter, R.R. 1996. Bootstrap for imputed survey data. *Journal of the American Statistical Association,* 91. 1278-1288.

Sinclair M., Diaz-Tena, N. & Wun, L. 2003. A simulation study to evaluate the robustness of recent methods for preparing variance estimates in the presence of hot deck imputation. *2002 Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association.

Singh, A.C., Grau, E.A., & Folsom Jr., R.E. 2002. Predictive mean neighborhood imputation with application to the person-pair data of the National Household Survey on Drug Abuse. *2001 Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association.

U.S. CensusBureau 2009. *Design and Methodology: American Community Survey* Washington.

U.S. Department of Housing and Urban Development 2007. *Fair Market Rents for the Section 8 Housing Assistance Payments Program.* Washington, DC: Office of Policy Development and Research.

Wolter, K. 1984. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association.*