Effects of Imputation on CPS Poverty Series: 1987 – 2007[1]

Joan Turek, Department of Health and Human Services
Fritz Scheuren, NORC University of Chicago
Charles Nelson, Bureau of the Census
Edward Welniak Jr., Bureau of the Census
Brian Sinclair-James, Department of Health and Human Services
Bula Ghose, Department of Health and Human Services

1.0 Introduction

The Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS) is the source of official Federal statistics on poverty[2]. U.S. poverty estimates based on the poverty thresholds originally developed by Molly Orshansky of the Social Security Administration in 1963-1964[3] are published annually.

Since the mid-1960's, many changes have occurred in our society and in the degree to which eligible CPS respondents are willing to report the cash income needed to construct these poverty figures. Many improvements have occurred in the CPS as well. Still, the series of annual cross-sectional poverty estimates are believed to provide a meaningful measure of change in U.S. poverty over time. A systematic review of this conjecture, though, seemed due and this paper continues this process by building on research findings presented at the 2009 Joint Statistical Meetings[4].

Why have we undertaken such a systematic review and believe it is in order? The principal reason is the large increase in CPS income "missingness"[5] since 1987, and the consequent increase in CPS imputation as seen in Chart 1. Between income years 1987 and 2007, the percent of dollar income received by persons that is imputed rose from 20.5% to 32.3%. This constitutes a 58% increase in the dollar amount of income that is imputed.

How has this added imputation impacted poverty rates during the period? Given the central role that mitigating poverty has played in Federal income transfer and health programs, it is important to try an answer this question. While the Census Bureau hot deck imputation procedures have known strengths, it is still possible that the growing nonresponse during this period may have altered historical trends in the characteristics of the poor.

Chart 1: Trend in the Percent of Total Dollar Income Imputed, All Persons with Positive Income
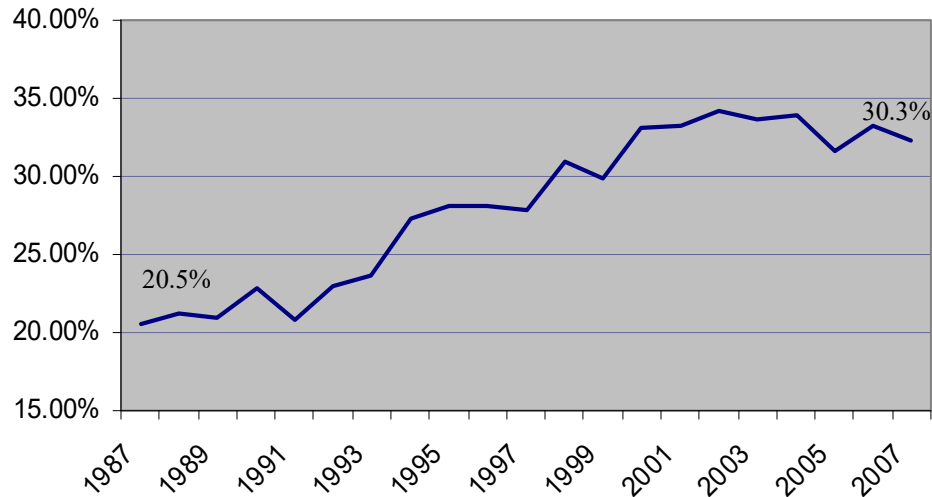
Although the current focus is from 1987 to the present, data have been compiled from income years 1976 to 2007[6]. Assembling available documentation and improving its accessibility, readability and interpretability are a main project goal. Recent advances in the creation of metadata systems at the Census Bureau and elsewhere motivate us. We want, when done, to have a better understanding about how imputation affects the CPS across a number of dimensions, including being able to say in what ways this historical series is a meaningful yardstick of the state of America's poor.

Now, frankly we expect to be at this big task for some time to come. Still we think that you may find some of our efforts and conjectures worth listening to. Perhaps the natural place to begin is with how poverty is defined in the CPS. We then go on to discuss the treatment of CPS income nonresponse. Finally, trends in poverty rates since 1987 classified by imputation type and by selected demographic groups are presented. Demographic groups were selected that are either commonly included in research studies and/or that are expected to include persons likely to be eligible for Federal transfer programs. The impact of eliminating item imputation on overall poverty rates is also discussed. Some technical details are covered in the extensive footnotes.

2.0 Defining Poverty and the Official Measure

During the early 1960's, public policy began to reflect a growing believe that poverty resulted partly from societal failure rather than primarily from the unwillingness of the poor to be gainfully employed. Michael Harrington's book, *The Other America: Poverty in the United States,* influenced both the Kennedy Administration and the Johnson Administration's "War on Poverty"[7].

Efforts to create public programs helping the poor required some method of estimating who is poor and needing assistance. Molly Orshansky's poverty thresholds, which she characterized in her 1965 **Social Security Bulletin** article as income inadequacy thresholds, provided this measure. They were based on the Department of Agriculture's economy food plan and were adjusted for family size, farm/nonfarm status, sex of the family head, number of children and by age for families of size one and two. Food costs were multiplied by three to reflect other needs.

These thresholds, adjusted by the Consumer Price Index since 1969, were, with minor modification, used until 1981, to develop the annual Official Poverty Thresholds applied to the CPS income supplement each year to estimate US poverty. In 1981, the farm/nonfarm and female/male headed distinctions were

eliminated and family size extended to nine or more persons. Several other efforts to revise the poverty thresholds have been proposed since then but have not yet been implemented[8].

3.0 Perspective on CPS Income "Missingness"

Problems with missing data are always present to some degree on surveys. When missing data are accounted for through imputation or by some other means, usually there is an implicit assumption that data are missing at random after controlling for other variables. However, evidence indicates that CPS income missingness may not be completely random. Thus, even after imputation, bias can result. [9]

Now there are many statistical goals in imputation. The one we find particularly important is the extent that imputation reduces the missingness bias in CPS survey estimates. This goal is met to the degree that patterns of nonresponse are correctly identified and corrected for.[10] In the CPS there are two basic patterns of missingness – where responses to some but not all of the CPS income questions are missing (Item Imputes) and where the entire CPS supplement is missing (Whole Imputes).

- Item Imputes. Sample persons or other household members who fail to respond to a specific question and "item" imputation is performed.

- Whole Imputes. Sample persons who only responded to the basic labor force questions. In this case, the "whole" or entire supplement has to be imputed.

This distinction is important because in the case of item imputes, many more variables are available to find a good match to impute any missing data. For whole impute cases, where the whole supplement is missing, there are fewer variables to match on and thus we believe the chances of finding a way to reduce the missingness bias are likely to be less. In either case, after imputation a complete data set is created[11].

The Census Bureau started regularly imputing for missing CPS income in 1962. Since then, the same basis strategy, "hot deck" imputation, has been employed[12]. With this procedure, non-respondents are assigned income amounts reported by respondents with similar characteristics. The process is conducted at the person level for each income source identified. A complex set of demographic, economic and social characteristics is used in identifying similar person-level respondents.[13] Different types of missingness are treated differently. Item (partial) imputes are based on responses both to the entire monthly survey and the ASEC supplement, while whole imputes are only based on the monthly survey[14].

A new processing system was introduced in the 1988-1989 period. This revision expanded the editing and imputation process to reflect a substantial expansion in the number of income sources collected on the March questionnaire starting in 1979[15]. The hot deck procedure was also modified in the 1988-1989 period to impute all missing supplemental items from one source, and to retain all reported data during the nonrespondent/respondent match process[16]. The new questionnaire allowed for over 50 sources of income and the recording of 27 different amounts. Prior to 1979, 11 income categories were collected. With minor change, the processing system implemented in 1989 is still being used today.[17]


4.0 Trends in Poverty Rates

Chart 2, *Poverty Trends by Types of Missingness: 1987-2007*, shows trends in poverty rates separately for persons with item, whole or no imputes (reporters) for the time period 1987-2007. Only persons with positive income are included.[18] The pattern of item, whole and no imputes are consistent over time. Persons with no imputes have the highest poverty rate, those with whole imputes are next and those with item imputes have the lowest poverty rate.[19]

Whole imputes do not reduce poverty much because there is no information about the previous year's status of these persons and they are drawn from everyone in the imputation pool. If full-year workers are

more likely to need imputation, there is no way to reflect this in the whole imputation system. It can be reflected in the item imputation system, as more characteristics are available for use in doing the imputation. As seen later, zeroing out item imputes for persons with positive income in 2007 dramatically increased poverty rates.

Chart 2: Poverty Trends by Types of Missingness 1987-2007
Persons with Positive Income
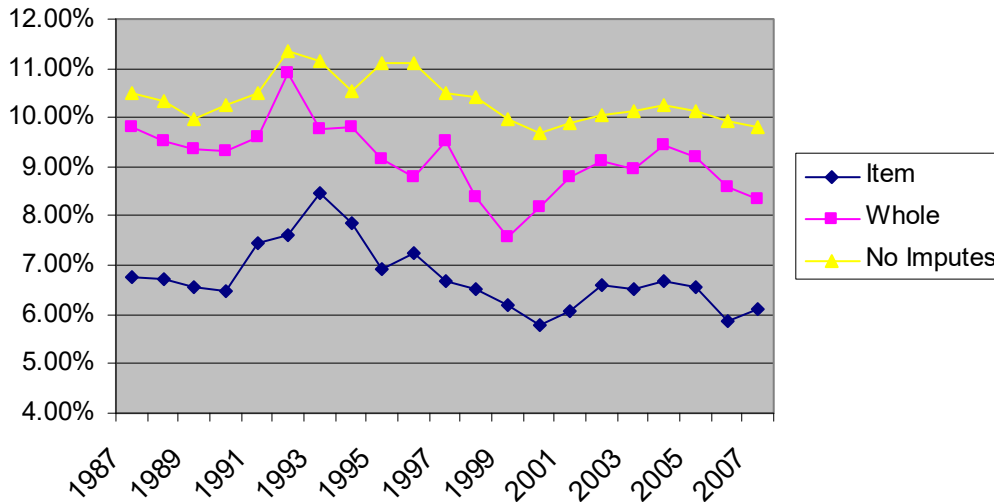
### 5.0 Poverty Rate Comparisons: Selected Demographic Groups

How imputation affects demographic groups important for policy is of major importance. For this reason, tabulations were developed for selected demographic characteristics at five-year intervals from 1987 through 2007. Charts were developed for persons with positive income that provide average estimates of item, whole and no imputation during this period.

Averages are constructed by summing each imputation category across all years and then dividing by the number of years (five). Appendix Table 1 provides separate poverty rate estimates for each of the five years for which tabulations were developed and for the five-year averages.

Five demographic groups are examined. The groups are: Family Type, Age, Race, Gender and Education. Each of these categories is further subdivided into smaller groupings. In selecting subgroups attention was paid to separately categorizing those likely to contain persons eligible for Federal programs.

Family Type: Chart 3, *Poverty Rates of Single and Married Parents* looks at single parent and married-parent families by type of imputation. . This split was selected because single parent families with children are important recipients of Federal aid, such as welfare programs. Single parents are about four times as likely to be poor as married-parent families. Single parent families have the highest poverty rates of any of the demographic groups studied. While they have similar patterns by type of imputation, that is, persons with no imputes generally having the highest poverty rates and those with item imputes generally having significantly lower poverty rates, the higher poverty rates for single parents dominate.[20] Overall, a significantly larger proportion of single parents are in poverty and more likely to be candidates for Federal aid.

Chart 3: Poverty Rates of Single and Married Parents
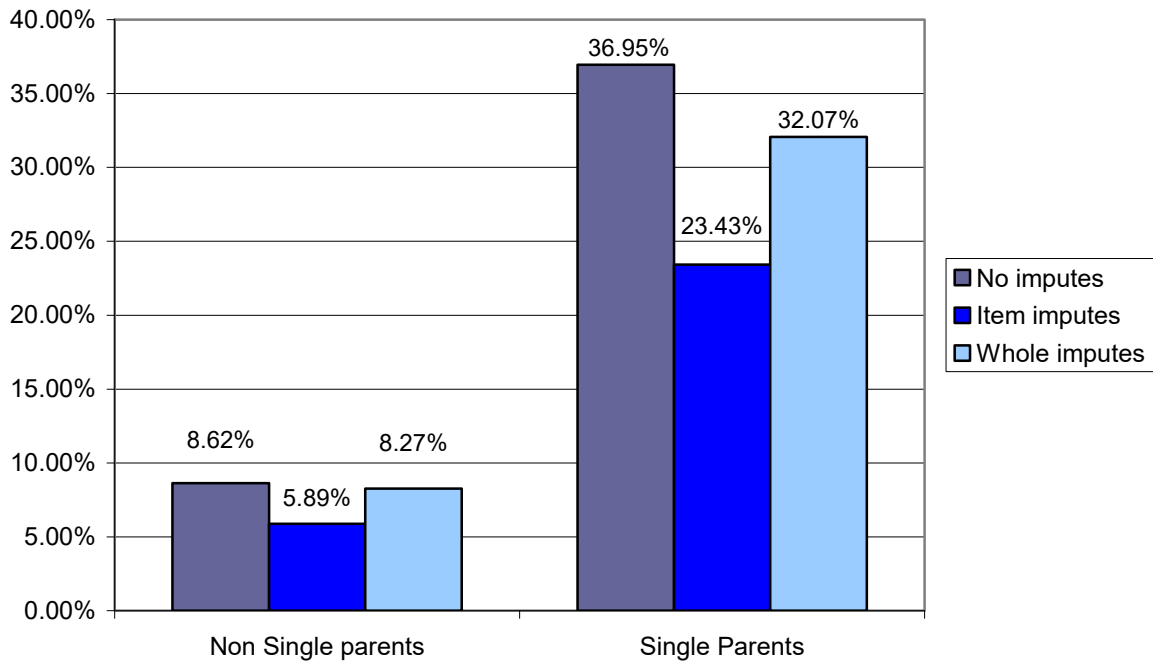
Age: Three age categories were chosen for presentation: persons aged 18 to 44, aged 45 to 64, and 65 and older in Chart 4, *Poverty Status by Age of Person.* These are commonly used categories and can be viewed as separating persons in the earlier stages of their careers and more likely to have children under 18 from those who are farther advanced and more likely to have college aged children. Persons 18 to 44 are more likely to have children under 18 and to be eligible for welfare programs. Many persons 65 and older are likely to be retired, although this may be changing as health improves and economic circumstances worsen.

Chart 4: Poverty Rates by Age
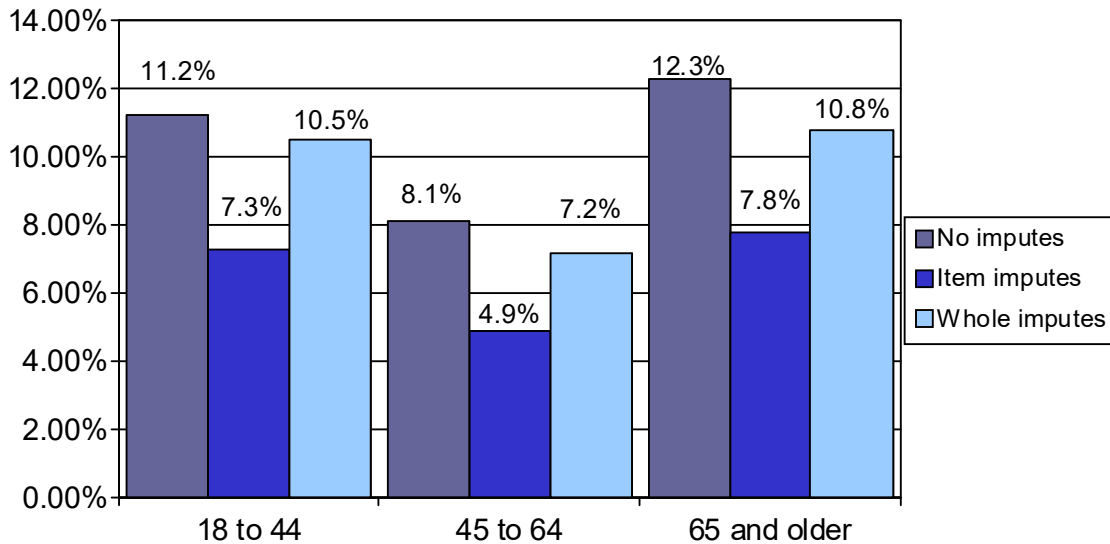
People aged 18 to 44 with no imputes or item imputes have slightly lower poverty rates than people 65 and older. The difference between the whole imputes was not statistically significant. People 45 to 64 have the lowest poverty rates. The higher poverty rate for people 65 and older may reflect underreporting of asset income on the CPS and their higher percentage of income from unearned sources.

Race: Two race categories are presented, Whites and Blacks. Finer distinctions identifying additional race categories (e.g., Asians) and separately identifying Hispanics could not be constructed covering the entire time period 1976 to 2007. Large disparities exist by race, The poverty rate of Blacks with no imputes is more than double that for Whites as seen in Chart 5, *Black and White Poverty Rates.* That is they are 23.7% vs. 8.6%. However, the pattern of poverty rates for Blacks and Whites is similar to those found earlier with persons with no imputes having the highest poverty rates and those with item imputes having the lowest poverty rates.

Chart 5: Black and White Poverty Rate
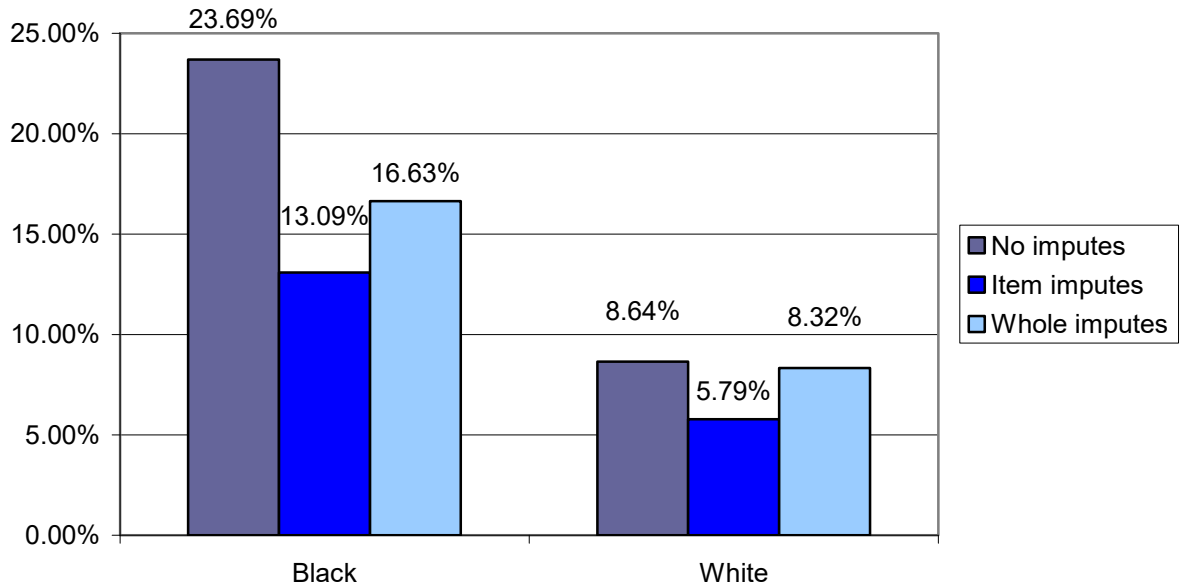
Gender: Women also have considerably higher poverty rates than men, as seen in Chart 6, *Poverty Rate by Gender*. Females are about 50 percent more likely to be poor than males Again, the biggest differences are between persons with no or whole imputes. The pattern continues with persons with no imputes having the highest poverty rates and those with item imputes having the lowest rates.

**Chart 6: Poverty Rates by Gender**
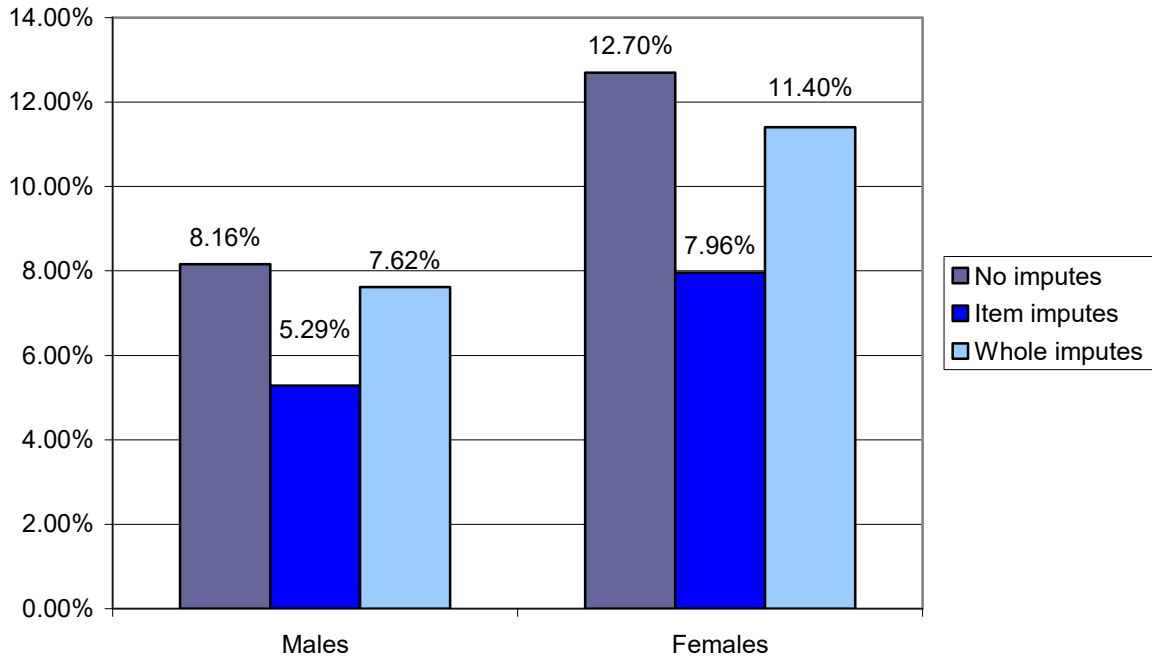
Education: As seen in Chart 7, *Poverty Rates by Education Level*, poverty rates vary inversely with education level. Overall, poverty rates for persons without a high school education are substantially higher. In fact, persons with a less than high school education have poverty rates about 8 times higher than for college graduates. Again the pattern identified earlier as between the item, whole and no imputes generally continues.[21]

Chart 7: Poverty Rates by Education
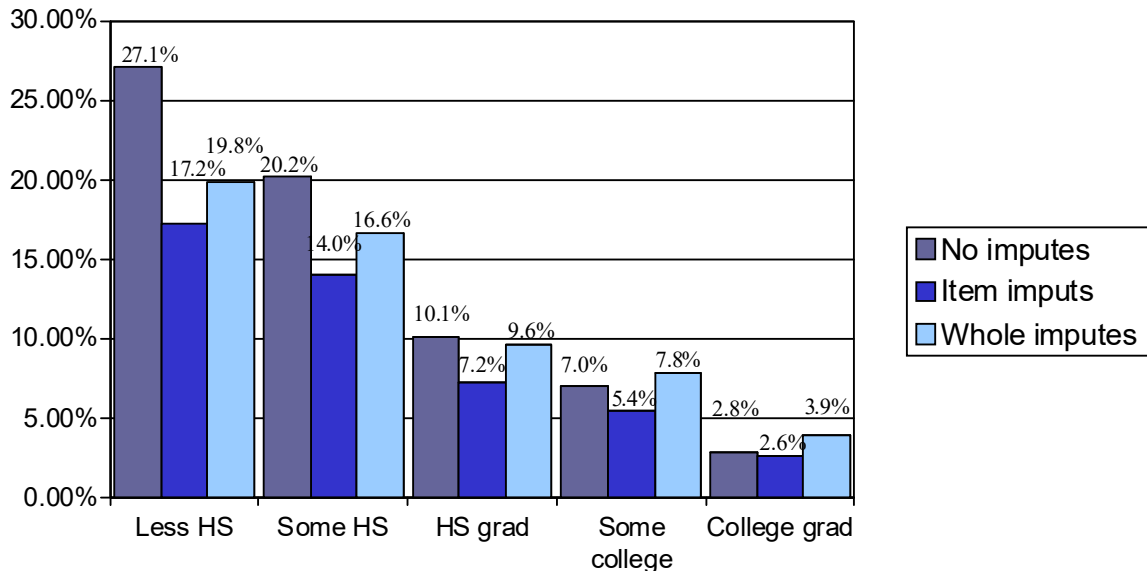
6.0 Interpretation of Results

The three series we have been looking at tell similar stories but this may be a cause of concern and not comfort. In particular, why were the two series with missing data similar across demographic groupings? Was this because only some of the known bias in the missingness[22] (e.g., Scheuren et al. 1980) was "corrected" by the imputation? We frankly do not know.

There have been recent CPS efforts to compare survey and administrative data[23] (e.g., O'Hare 2004) but these do not address poverty directly and so are only partially of value in firming up or rebutting the working hypotheses we began this exercise with. The fact that the whole impute series (8.35%) and that for the combined (item plus no impute) series (8.25%) are not statistically different, as seen below in Table 1, supports our concerns. For the time period 1987 to 2007, none of the ratios of whole imputes to item plus no imputes are statistically different from each other. But this makes sense, given how weak the variables are to match whole imputes to the rest of the sample. Maybe another approach to handling the whole imputes should be tried to see what the two distributions might be?

When looking at demographics, the higher poverty rates for persons in categories likely to be eligible for Federal programs is not surprising. At the same time, the significantly higher contribution of item imputation to lowering poverty rates stands out. Getting item imputation "right" appears critical to efforts toward "correctly" identifying the poor.

The greater effect that item imputes have on reducing poverty is also shown in Table 1 below where poverty rates are presented for persons with positive income in 2007 for a number of different measures. In all cases, the denominator is all persons in the category being measured in the numerator. For the estimates, all persons with item imputes set equal to zero and item imputes only equal zero, income

estimates for persons identified with positive income have been modified by setting all item imputes equal to zero[24] and their poverty status re-estimated As can be seen, poverty rates increase dramatically when item imputes are set equal to zero. Poverty rates increase from 8.3% to 35.1% for all persons with positive income and from 6.1% to 51.7% for persons whose positive item imputes have been set equal to zero. When whole imputes are excluded, the poverty rate remains at 8.3%. By comparison, the poverty rate for persons with whole imputes was 8.4%. Further examination indicates that a major portion of the imputations that have been zeroed out are for persons with jobs—a not surprising finding, since earnings account for around 80% of total income.

Table 1: Role of Imputes on Poverty Rates in 2007

No Imputes Only	9.82%	Pre1962 Measure
Item Imputes Only	6.11%	
Whole Imputes Only	8.35%	
All of Above	8.26%	
Whole Imputes Plus No Imputes	9.56%	
Item Plus No Imputes	8.25%	
All Persons with Item Imputes Equal Zero	35.07%	
Item Imputes Only Equal Zero	51.66%	

7.0 Next Steps and Future Directions

This is the second paper to undertake analysis of the complex data set developed for analyzing trends in imputation, A variety of future efforts and studies on the effects of imputation on poverty rates are under consideration. Some examples include record check studies to check for bias caused by the imputation and investigating refining the hot deck cells by perhaps considering the use of control card income. Additional analysis of imputation trends since 1976 will also be undertaken with initial focus on the effects of imputation on the income distribution i.e. a quintile-by quintile analysis as well as examination of other income distribution measures. More on these plans will be forthcoming.[25]

Appendix: Poverty Rates by Selected Demographic Characteristics Five Year Intervals and Five Year Average, All Persons with Positive Income

Family Composition

Non Single parents	1987	1992	1997	2002	2007	5 Yr. Average
No	8.60%	9.29%	8.46%	8.44%	8.20%	8.62%
Item	5.98%	6.69%	5.85%	5.96%	5.45%	5.89%
whole	8.64%	9.29%	8.18%	8.18%	7.07%	8.27%
Total pop	8.09%	8.76%	7.55%	7.38%	6.91%	7.70%
Single Parents						
No	42.05%	41.83%	37.17%	31.28%	31.78%	36.95%
Item	29.41%	31.85%	25.06%	20.10%	20.57%	23.43%
whole	33.38%	39.32%	35.44%	25.12%	29.16%	32.07%
Total pop	39.78%	40.26%	34.09%	27.06%	28.30%	33.40%

Age

18 to 45	1987	1992	1997	2002	2007	5 Yr. Average
No	10.46%	11.89%	11.44%	10.94%	11.06%	11.17%
Item	7.07%	8.13%	7.49%	7.25%	6.97%	7.33%
whole	10.24%	11.62%	10.50%	10.47%	9.61%	10.51%
Total pop	9.87%	11.21%	10.17%	9.50%	9.52%	10.06%
45 to 64						
No	8.34%	8.36%	7.89%	7.79%	8.09%	8.10%
Item	4.97%	5.37%	4.79%	5.07%	4.69%	4.92%
whole	7.78%	7.55%	7.56%	6.49%	6.80%	7.15%
Total pop	7.51%	7.57%	6.72%	6.44%	6.51%	6.87%
65 and older						
No	13.70%	15.39%	11.19%	11.36%	9.46%	12.31%
Item	7.98%	9.24%	7.42%	7.78%	7.18%	7.75%
whole	11.56%	13.67%	10.08%	10.36%	7.61%	10.83%
Total pop	12.19%	12.45%	9.60%	9.58%	8.28%	10.31%

Race

White

No	8.17%	9.14%	8.87%	8.61%	8.38%	8.64%
Item	5.91%	6.77%	5.85%	5.79%	5.21%	5.79%
whole	8.64%	9.34%	8.10%	8.23%	7.25%	8.32%
Item + no	7.68%	8.60%	7.75%	7.28%	7.03%	7.65%
Total pop	7.77%	8.68%	7.78%	7.38%	7.05%	7.71%

Black

No	27.27%	28.21%	22.20%	20.21%	19.53%	23.69%
Item	14.21%	15.34%	12.69%	12.47%	12.70%	13.09%
whole	16.96%	18.85%	17.30%	15.62%	14.31%	16.63%
Item + no	24.96%	25.57%	18.93%	16.90%	16.80%	20.31%
Total pop	24.00%	24.45%	18.73%	16.74%	16.48%	19.82%

Gender

	1987	1992	1997	2002	2007	5 Yr. Average
Males						
No	7.86%	8.75%	7.94%	8.12%	8.09%	8.16%
Item	5.73%	6.12%	5.21%	5.29%	4.77%	5.29%
whole	8.52%	8.17%	7.71%	7.11%	6.82%	7.62%
item plus no	7.40%	8.15%	6.93%	6.81%	6.69%	7.16%
Total pop	7.51%	8.15%	7.00%	6.84%	6.70%	7.21%
Females						
No	13.03%	13.80%	12.91%	11.87%	11.54%	12.70%
Item	7.74%	9.11%	8.15%	7.88%	7.44%	7.96%
whole	11.06%	13.56%	11.23%	11.09%	9.89%	11.40%
item plus no	11.92%	12.76%	11.19%	10.04%	9.81%	11.08%
Total pop	11.84%	12.85%	11.19%	10.15%	9.82%	11.11%

Education

Less HS

	1987	1992	1997	2002	2007	5 Yr. Average
No	27.89%	28.89%	26.81%	25.66%	24.78%	27.14%
Item	15.26%	18.40%	18.20%	16.90%	17.25%	17.22%
whole	21.15%	21.89%	19.52%	16.36%	18.25%	19.75%
Total pop	25.21%	26.34%	23.95%	21.94%	22.05%	24.15%

Some HS

	1987	1992	1997	2002	2007	5 Yr. Average
No	18.23%	22.12%	20.63%	19.23%	21.01%	20.15%
Item	12.57%	14.31%	14.88%	13.39%	14.51%	13.96%
whole	15.10%	19.46%	16.42%	14.98%	16.54%	16.58%
Total pop	17.03%	20.48%	18.71%	16.54%	18.64%	18.24%

HSGrad

	1987	1992	1997	2002	2007	5 Yr. Average
No	8.88%	10.55%	10.07%	10.22%	10.89%	10.07%
Item	6.33%	8.04%	7.06%	7.40%	7.18%	7.22%
whole	8.65%	10.99%	9.79%	9.41%	8.63%	9.55%
Total pop	8.37%	10.10%	9.04%	8.99%	9.28%	9.16%

Some Col

	1987	1992	1997	2002	2007	5 Yr. Average
No	6.10%	7.31%	7.09%	7.13%	7.21%	6.99%
Item	4.75%	5.74%	4.95%	5.80%	5.48%	5.42%
whole	6.70%	7.75%	7.51%	8.54%	7.81%	7.77%
Total pop	5.88%	7.04%	6.38%	6.73%	6.61%	6.56%

Col Grd

	1987	1992	1997	2002	2007	5 Yr. Average
No	2.56%	2.81%	2.77%	3.03%	2.90%	2.82%
Item	2.84%	2.74%	2.50%	2.78%	2.46%	2.62%
whole	4.28%	3.36%	4.15%	4.31%	3.46%	3.88%
Total pop	2.76%	2.85%	2.78%	3.03%	2.75%	2.84%

The educational categories are "less than a high school education" (less HS); "some high school, but no degree" (Some HS); "high school graduates" (HS Grad), some college, (Some Col), and college graduate with a BA or advanced degree (Col Grd).

References

[1] The views are those expressed by the authors and are not the official position of their organizations.

[2] Technical Paper 66: *Current Population Survey Design and Methodology*, October 2006, U.S. Department of Commerce, Bureau of the Census, Washington D.C

[3] Social Security Bulletin, January 1965

[4] Joan Turek, Fritz Scheuren, Brian Sinclair-James, Bula Ghose, Sameer Desale, *Effects of Imputation on CPS Income and Poverty Series: 1981-2007*, Joint Statistical Meetings, Washington D.C., August 4, 2009.

[5] "Missingness" is used as a term of art in this paper to mean a failure by a survey respondent to fully report their income.

[6] From 1947 until 1961, income estimates were derived only from fully reported cases. The Census Bureau began systematically assigning values to missing income items in 1962. Significant changes in income questions and imputation procedures occurred during this period, hence, analysis in limited to data from income year 1976 to the present.

[7] Michael Harrington, *The Other America: Poverty in the United States* (New York: Macmillan, 1962.

[8] Gordon M. Fisher, *The Development and History of the U.S. Poverty Thresholds --- A Brief Overview,* U.S. Department of Health and Human Services, http://aspe.hhs.gov/poverty/papers/hptgssiv.htm)

[9] H. Lock Oh, Social Security Administration. Fritz Scheuren, Internal Revenue Service, Hal Nisselson, WESTAT, *Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data,* Papers an d Proceedings, Joint Statistical Meetings, 1980.

[10] U.S. Census Bureau, Survey of Program Dynamics, *Data Editing and Imputation Goals,* http://www.census.gov/spd/goals.html).

[11] Technical Paper 54: *Income Nonresponses: March 1983 CPS*, September 1985, U.S. Department of Commerce, Bureau of the Census, Washington D.C and Edward J Welniak, *Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty,* Bureau of the Census, Washington D.C. 20233.

[12] Ford, Barry, *Hot Deck Imputation* in Theory of Incomplete Data*,* Volume II, National Academy of Science, 1983.

[13] Technical Paper 54: *Income Nonresponses: March 1983 CPS*, September 1985, U.S. Department of Commerce, Bureau of the Census, Washington D.C and Edward J Welniak, *Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty,* Bureau of the Census, Washington D.C. 20233.

[14] Noninterview nonresponses, where both the income supplement and monthly labor force surveys are missing, are not covered in this paper, since they are handled by weight adjustments and not imputation.

[15] The new questionnaire was fully adopted in 1980 with minor modification.

[16] The processing system in use before this time was implemented in 1976. Prior to 1979, 11 income categories were collected. Due to resource constraints, the expanded income categories were collapsed back to these 11 categories for processing until 1988. While our estimates of total imputation compared to those in Technical paper 54 for 1983 were highly similar, there were significant differences in the split between item and whole imputes.

[17] op.cit., Welniak, This discussion relies heavily on that source.

[18] All persons in families where someone has negative income should probably been excluded. When persons with negative income are included, there is a virtually insignificant impact on poverty rates. Poverty rate changes in 1983 and 1997 were substantially less than one percent and do not materially affect the results presented in this paper.

[19] The differences between poverty rates for "no imputes" and "whole imputes" are not statistically significant for the years 1987-1992, 1994, 1997, 2001-2002, and 2004-2005.

[20] The differences between "no imputes" and "whole imputes" were not statistically significant for married-couples, high school graduates and whites and not statistically significant for "item imputes" for college graduates.

[21] The poverty rates for whole imputes for people with some college and for college graduates are higher than for people with no imputes. The apparent difference between the rates for no imputes and whole

imputes for high school graduates was not statistically significant, nor was the apparent difference between no imputes and item imputes for college graduates.

[22] ibid, Welniak.

[23] Amy O'Hare, *Allocated Values in Linked Files*, US Bureau of the Census, presented at the Federal Committee on Statistical Methodology.

[24] The population being considered is all persons who have positive income including item imputes.

[25] Dr. Scheuren is taking the lead in developing proposals for further analysis. Documentation describing development of this comprehensive database. Joan Turek and Bula Ghose, *Documentation for CPS Income Imputation Data Set,* is available from Joan Turek, joan.turek@hhs.gov.