Statistical Journal of the IAOS 33 (2017) 495–501
DOI 10.3233/SJI-160334
IOS Press

Imputation and money income distribution measures[1]

Joan L. Turek
Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, 200 Independence Avenue S.W., Washington, DC 20201, USA
E-mail: JLTurek@aol.com

Abstract. This paper examines the growing inequality in American money income at the person level. It tests whether the growing trend in the imputation of money income data on the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS) conducted by the U.S. Bureau of the Census may be a contributing factor. The growing disparity in the distribution of money income over time does not appear to be an artifact of the imputation methods employed in developing aggregate money income estimates.

Keywords: Income inequality, income distribution, imputation, GINI index, income quintiles, total money income

1. Introduction

Numerous studies have reported on the growing income disparity in the United States since the 1970s.[2] By 2008, the trend towards greater inequality in the distribution of money income had reached levels not seen since the Great Depression of the 1930's.[3] Growing inequality has gained increased attention recently.

This is evident in the popularity of related books such as Thomas Piketty's "Capital in the 21st Century"[4] and in contemporary political discussions.[5] This paper's research builds on several earlier studies that examined the impact of imputation on total positive money income and on those in poverty.[6]

It examines the growing trend in income inequality and tests whether the growing trend in the imputation of money income on the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS) may be a contributing factor. Trends in the imputation of income at the person level are examined for selected calendar years from 1977 to 2007 in order to see the degree to which these trends might have influenced trends in inequality. The paper does not attempt to look at the many other factors affecting inequality such as changes in family composition, the growth of two income families, changes in the occupational mix over time, etc.

2. Imputation

Missing data are always a factor, to some degree, in surveys. When missing data are accounted for through

[1] The views are those expressed by the author and are not the official position of her organization.

[2] See: Arthur E. Jones Jr. and Daniel H. Weinberg, *The Changing Shape of the Nation's Income Distribution: 1947–1998*, Current Population Reports, P60–204, U.S. Census Bureau, U.S. Department of Commerce, June 2000, p.11.

[3] See: Arthur E. Jones Jr. and Daniel H. Weinberg, *The Changing Shape of the Nation's Income Distribution: 1947–1998*, Current Population Reports, P60–204, U.S. Census Bureau, U.S. Department of Commerce, June 2000, p.11.

[4] Piketty, Thomas. "Capital in the 21st Century." *Cambridge: Harvard* University (2014). His book has created considerable debate. Also see: Bill Gates notes: http://www.gatesnotes.com/Books/Why-Inequality-Matters-Capital-in-21st-Century Review. Wikipedia, the free encyclopedia, also provided a detailed bibliography of articles responding to Thomas Piketty's book: Https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century.

[5] Peter Whoriskey, *Income gap widens as executives prosper*, The Washington Post, Sunday, June 19, 2011, page A1.

[6] Joan Turek, Fritz Scheuren, Brian Sinclair-James, Bula Ghose and Sameer Desale, *Effects of Imputation on CPS Income and Poverty Series?* Joint Statistical Meetings, Washington D.C.; August 4, 2009 and Joan Turek, Fritz Scheuren, Charles Nelson, Edward Welniak Jr., Brian Sinclair-James, and Bula Ghose; *Effects of Imputation on CPS Poverty Series: 1987–2007*; Federal Committee on Statistical Methodology; Washington D.C., November 3, 2009.

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 missing ASEC money income data may not be completely random. Hence, if the non-random nature of missing data is not properly accounted for, bias can result.[7]

The Census Bureau started imputing for missing ASEC income data in 1962. Since then, the same basic strategy, "hot deck" imputation, has been employed, although the exact process has been revised several times.[8] With this procedure, non-respondents are assigned income amounts reported by respondents with similar characteristics. This process is conducted at the person level for each income source identified. A complex set of demographic, economic and social characteristics is used to identify similar person-level respondents.[9] Hot deck imputation is preferable to mean imputation or just dropping cases with missing values because it maintains the distribution, whereas mean imputation would reduce the variance. In addition, hot deck imputation recognizes that the population that does not answer the survey might be different from the population that did answer the question. Applying a mean imputation might not represent the group that did not answer. Of course, this also can happen with hot deck imputation. But the bias is thought to be less.

In the ASEC there are two basic types of imputation for missing data:

- Item Imputes. Sample persons or other household members fail to respond to a specific question on the ASEC and data for that particular "item" is imputed. Responses to more than one item may be imputed.
- Whole Imputes. Sample persons respond to the basic monthly survey questions but refuse to respond to the ASEC supplement. In this case, the "whole" or entire supplement is imputed.

Item imputes use information reported in both the monthly survey and the ASEC supplement, while whole imputes only use information reported on the monthly survey.[10] 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.[11] For whole impute cases, where the whole supplement is missing, there are fewer variables to match on and thus the chances are greater that bias is increased from missing data. In either case, after imputation a complete data set is created.

3. Trends in the imputation of total positive money income

Imputation has increased so much over the years that it is important to know whether the increase in inequality is real or if it is a statistical illusion influenced by declining data quality resulting from increased imputation over time. Total positive money income estimates were developed by type of imputation at the person level, since this is the level at which income imputation is performed. Only positive money income is included in the analysis. Instead of excluding persons whose total income was negative, each item of negative income was eliminated and each person's positive income was summed. This was done to permit a separate analysis of individual income components during previous research studies. The differences were minuscule. When poverty estimates are compared including and excluding negative income, restricting to positive values had only a marginal effect on them.

Some persons' total positive money income may consist of both reported and imputed income. Estimates of the number of persons receiving income with no imputes (e.g. reported) or a combination of no and item imputes are constructed such that they are mutually exclusive.[12] Top codes also have been replaced by the mean above the top code to permit comparisons

[7] See for example H. Lock Oh, Social Security Administration, Fritz Scheuren, Internal Revenue Service, Hal Nisselson, WESTAT, *Differential Bias Impacts of Census Bureau Procedures for Imputing Missing CPS Income Data*, Papers and Proceedings, Joint Statistical Meetings, 1980.

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

[9] *Technical Paper 54: Income Nonresponse: March 1983 CPS*, September 1995, Department of Commerce, Bureau of the Census, Washington D.C. 20203 and Edward J. Welniak Jr., *Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty*, 1990, U.S. Department of Commerce, Bureau of the Census, Washington D.C. 20233.

[10] Non-interview non-responses, 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.

[11] The ASEC collects detailed information on income by source, on employment, earnings hours of work; and on a variety of demographic characteristics such as age, race, marital status, educational attainment health insurance coverage and participation in transfer programs.

[12] Moreover, although estimates are shown separately for item and whole imputes, the estimate for 1977 are an exception. Public use files for that year do not separately identify the two types of

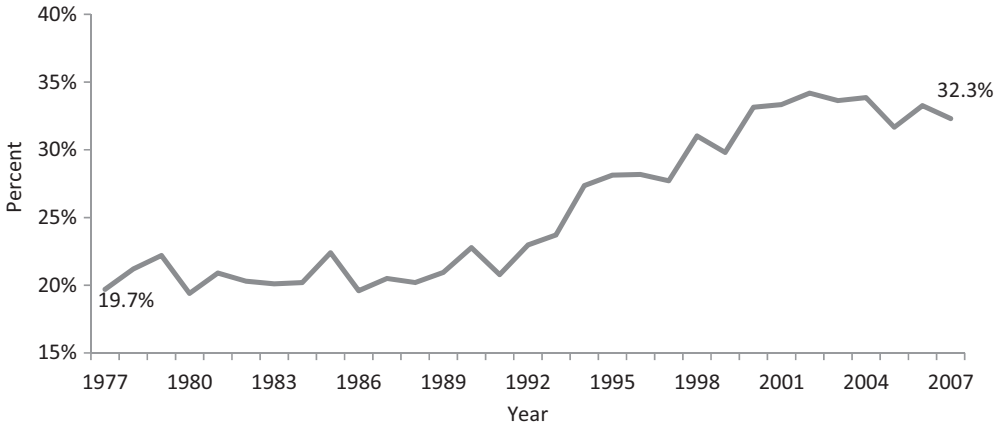

Fig. 1. Trend in the percent of total money income imputed, all persons with positive income. Source: Current population survey, annual social and economic supplement.

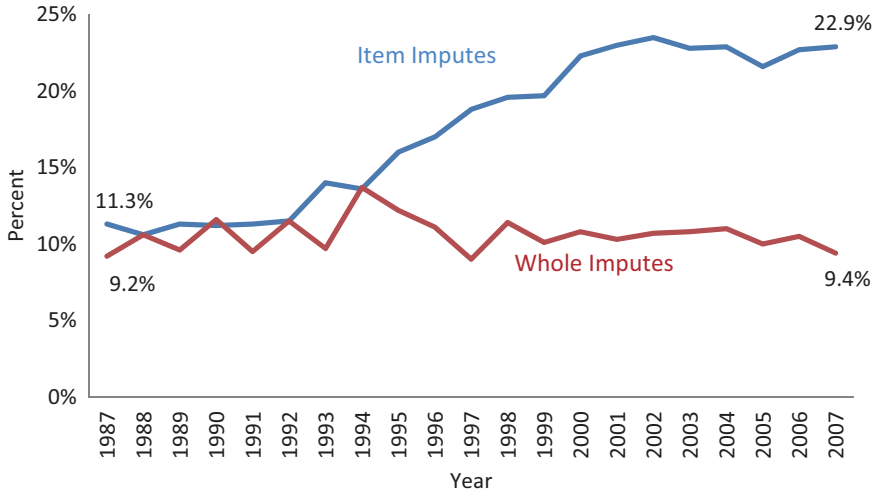

Fig. 2. Trend in percent of total money income that is item and whole imputed, all persons with positive income. Source: Current population survey, annual social and economic supplement.

over time on the same basis for persons with very high income.

Between 1977 and 2007, the percent of imputed positive total money income rose from 19.7% to 32.2% as seen in Fig. 1. This constitutes a 64% increase in the total amount of money income that is imputed. This growth slowed in the early 2000's perhaps due to enhanced efforts by the Census Bureau to reduce nonresponse.

Figure 2 shows trends in item and whole imputation from 1987 to 2007.[13] As can be seen, whole imputes have remained fairly stable over this period, ranging from 9.2% in income year 1987 to 9.4% in 2007. At their peak in 1994, whole imputes accounted for 13.7% of total dollar income. Between the 1987 and 2007 time period, item imputes doubled from 11.3% to 22.9%. Thus, while persons were willing to indicate they received income they were more reluctant to say "how much".

imputation and the authors estimated them. That is, they checked to see if some or all of the individual income items were identified as imputed. Where all were shown as imputed, it was assumed that a whole impute had been performed. Otherwise, it was assumed they were item imputes. See: Joan Lee Turek Ph.D. and Bula Ghose, Documentation for CPS Income Imputation Dataset, 2007, Joan.turek@hhs.gov.

[13]Figure 2 does not cover 1977 to 1986 since Census did not separately identify item and whole imputes during this time period.

4. Developing the GINI coefficient

The GINI coefficient provides a summary measure of money income inequality.[14] It incorporates detailed share data into a single statistic summarizing the dispersion of income across the entire income distribution. The GINI coefficient ranges from 0, indicating perfect income equality, to 1 indicating perfect income inequality. With perfect equality everyone has the same money income share while with perfect inequality only one person or group of persons receives all of the income. GINI coefficients are absolute numbers that can be easily compared at different points in time. These can be derived from the Lorenz curve.[15]

GINI coefficients are frequently produced for households or families. However, this report, as stated earlier, constructs GINI coefficients at the person level to better understand the connection between income inequality and imputation which is performed at the person level. It looks at persons who have positive total money income and attempts to examine whether elevated imputation rates have influenced income inequality. Initially, inequality measures for all persons with positive money income are examined. Later, analysis focuses on policy relevant demographic groups.

5. GINI coefficient trends

Several inequality estimates are constructed for comparison. Initially, GINI coefficients are used to examine trends in inequality of total positive money income at five year intervals from 1977 to 2007. Later years are excluded due to the effects of the "Great Recession." A separate measure of inequality, the aggregate shares of total positive money income received by each fifth of the income distribution at five year intervals between 1992 and 2007, is also presented. Finally, GINI coefficients are developed at five year intervals from 1992 to 2007 for selected demographic groups.

As seen in Table 1, the GINI Coefficients for all persons with positive money income have not varied greatly over the 1977–2007 time period. In recent years, "item and whole imputes" tend to show slightly

Table 1
GINI coefficients, all persons with total positive money income by type of imputation: 1977 to 2007

Year	No imputes	Item imputes	Whole imputes	Total positive money income
2007	0.50036	0.50139	0.50636	0.50112
2002	0.50951	0.51128	0.51993	0.51140
1997	0.50880	0.51519	0.53283	0.51191
1992	0.50178	0.50819	0.51157	0.49880
1987	0.50764	0.50576	0.50868	0.50285
1982	0.50657	0.50285	0.56842*	0.50680
1977	0.49903	0.53122	0.51432	0.49771
Ratio of each GINI coefficient to no imputes (parent)				
2007	100.0	100.2	101.2	100.2
2002	100.0	100.4	102.1	100.4
1997	100.0	101.3	104.7	100.6
1992	100.0	101.3	102.0	99.4
1987	100.0	99.6	100.2	99.1
1982	100.0	99.3	112.2*	100.1
1977	100.0	106.5	103.1	99.7

*Differences tend to be small and not significant at the 95% confidence level, except when asterisked. Source: Current population survey, annual social and economic supplement.

Table 2
Percent of total positive money income in each quintile: 1992 to 2007

Quintile	Percent of total positive money income			Total positive money income
	No imputes	Item imputes	Whole imputes	
2007				
Lowest	2.6	2.8	2.7	2.6
2	7.5	8.5	8.1	7.8
3	13.8	13.9	14.8	13.9
4	22.9	21.7	22.8	22.6
Highest	53.2	53.1	51.6*	53.1
1997				
Lowest	2.3	2.6	2.6	2.4
2	7.4	7.9	7.6	7.5
3	13.9	13.5	13.4	13.8
4	23.4	20.0	21.7	22.6
Highest	53.0	56.1*	54.7	53.7
1992				
Lowest	2.2	2.3	2.4	2.3
2	7.7	7.9	7.9	7.7
3	14.6	14.4	14.7	14.6
4	24.9	23.0	23.8	24.6
Highest	50.6	52.5	51.3	50.9

*Differences tend to be small and not significant at the 95% confidence level, except when asterisked. Source: current population survey, annual social and demographic supplement.

[14] Op.cit. Arthur E. Jones Jr. and Daniel H. Weinberg, *The Changing Shape of the Nation's Income Distribution: 1947–1998*, p.11.

[15] As stated in Damgaard, Christian. "Lorenz Curve." From *MathWorld* – A Wolfram Web Resource, created by Eric W. Weisstein; http://mathworld.wolfram.com/LorenzCurve.html, the Lorenz curve is used in economics and ecology to describe inequality in wealth or size. The Lorenz curve is a function of the cumulative proportion of ordered individuals mapped onto the corresponding cumulative proportion of their size. Given a sample of n ordered individuals with x'_i the size of individual i and $x'_i < x'_2 < \ldots < x'_n$, then the sample Lorenz curve is the polygon joining the points $(h/n, L_h/L_n)$, where $h = 0, 1, 2, \ldots, n$, and $L_h = \sum_{i=1}^{h} x'_i$. Alternatively, the Lorenz curve can be expressed as $L(y) = \frac{\int_0^y xdF(x)}{\mu}$, where is the cumulative distribution function of ordered individuals and is the average size.

Table 3
GINI coefficient trends, by demographic group: 1977 to 2007

Quintile	No imputes	Item imputes	Whole imputes	All with oositive income
		Age		
18–44				
2007	0.48329	0.49263	0.47841	0.48398
1992	0.47903	0.49150	0.48692	0.47577
1977	0.45730	0.51531	0.49137	0.45881
45–64				
2007	0.47748	0.49648	0.48991	0.47769
1992	0.48720	0.49003	0.49700	0.48521
1977	0.46619	0.49146	0.47150	0.46305
65 and older				
2007	0.47147	0.49001	0.49765	0.48470
1992	0.45310	0.46398	0.47470	0.45603
1977	0.45265	0.47169	0.44989*	0.45399
		Gender		
Male				
2007	0.48515	0.48009	0.50546	0.48569
1992	0.46867	0.47496	0.48768	0.46562
1977	0.44658	0.47893	0.46198	0.44388
Female				
2007	0.48992	0.49556	0.48184	0.49145
1992	0.49813	0.49753	0.49861	0.48014
1977	0.47042	0.48617	0.49472	0.47401
		Race		
White (Hispanic and nonHispanic)				
2007	0.49953	0.50169	0.50721	0.50077
1992	0.49909	0.51060	0.50830	0.49633
1977	0.49839	0.52608	0.51962	0.49723
Black (Hispanic and nonHispanic)				
2007	0.46504	0.47053	0.48092	0.47368
1992	0.48525	0.46600	0.49808	0.48265
1977	0.47259	0.47718	0.47293	0.46540
Other**				
2007	0.51345	0.51423	0.51835	0.51636
1992	0.50693	0.49578	0.54908*	0.50829
1977	0.48677	0.53054*	0.48076	0.48532
		Family composition		
Single parent				
2007	0.46466	0.47146	0.45567	0.46466
1992	0.47046	0.48396	0.43412	0.47030
1977	0.43483	0.38558	0.43073	0.40832
Both parents				
2007	0.46942	0.47555	0.47365	0.47547
1992	0.39904	0.40807	0.43412	0.39563
1977	0.32865	0.38710	0.34137	0.32638
		Education		
Less than high school (HS)				
2007	0.41115	0.44092	0.47985	0.42566
1992	0.42982	0.45564	0.45162	0.43429
1977	0.46167	0.46510	0.46520	0.46398
Some high school				
2007	0.51664	0.53728	0.52150	0.52123
1992	0.52729	0.52345	0.55534	0.52966
1977	0.56057	0.52250	0.56275	0.55271
HS graduate/GED				
2007	0.42982	0.43832	0.43997	0.43103
1992	0.44143	0.44116	0.44287	0.43447
1977	0.42725	0.45232	0.45251	0.42440

Table 3, continued

Quintile	No imputes	Item imputes	Whole imputes	All with oositive income
Some college				
2007	0.44219	0.45294	0.45880	0.44483
1992	0.45638	0.47139	0.47340	0.45444
1977	0.46841	0.50135	0.48245	0.46613
College graduate				
2007	0.45157	0.45312	0.47689	0.45397
1992	0.43179	0.45145	0.45794	0.42968
1977	0.44461	0.46322	0.48338	0.44612

*Differences tend to be small and not significant at the 95% confidence level, except when asterisked. **All other racial categories. Source: Current population survey, annual social and economic supplement.

higher levels of income inequality as compared to "no imputes", although they all show declining income inequality over time. The differences tend to be small and not significant. Some of the values shown are outside the 95% confidence interval. When this occurs, they are asterisked (*). Comparisons are also made between GINI coefficients for no imputes and the other GINI coefficients by imputation type. Even with the greater variability for "item and whole imputes," the differences in the GINI coefficients are in most instances small. They are all less than one percent. Data from the "no imputes" still account for the largest portion of income as seen in Figs 1 and 2.

6. Percent of total money income in each quintile

In order to get an additional comparison of changes in the distribution of income over time, the aggregate shares of total positive money income received by each fifth of the income distribution in 1992, 1997 and 2007 were compared (Table 2). This provides a different and commonly used way of examining the growth in income inequality among persons over this time period. As can be seen in Table 2, the percent of total positive money income received by persons in the lowest quintile increased by 18% between 1992 and 2007 from 2.2% to 2.6%, while the percent in the highest quintile rose by 5% from 50.6% to 53.2%.

Both item and whole imputes tend to increase inequality; however, when the percent of total positive money income in each quintile is compared to the percent for no imputes the differences are marginal. In other words, if the income distributions within each subgroup are not significantly different from each other, a change in the prevalence of one type of imputation or another should not alter the distribution in the aggregate. Overall, these results mirror those for GINI coefficients.

7. GINI coefficients – selected demographic categories

Table 3 presents GINI coefficients for selected demographic groups. Again, for most groups, the fluctuations over time by imputation status are small and suggest that imputation has not greatly affected income distribution estimates for most groups. Asterisks indicate where results are significant. This only occurs for imputes in selected years in the Other Race category. As seen earlier for all persons, there is more variability in the GINI coefficients reported for item and whole imputes, but even here most differences are small and not significant at the 95% confidence level. The only significant coefficients are for persons of races other than black or white, who are either Hispanic or non-Hispanic.

8. Conclusion

The growing disparity in the distribution of money income over time does not appear to be an artifact of the imputation methods employed by the U.S. Bureau of the Census. Further, this analysis indicates that increases in imputation since 1977 do not appear to greatly alter measures of inequality. However, failure to adjust for missing total positive money income would have resulted in biased measures of this income.[16] The adjustments due to imputation do not appear to have created significant biases in these estimates. While there is more variation in the GINI coefficients for item and whole imputes, the GINI coefficients for "no imputes" tend to be similar to the GINI coefficients for total money income reflecting the larger share of this source of income in the total – ap-

[16]While now dated, the Social Security Administration Series by Scheuren et al. (1972–1980) *Studies from CPS-IRS-SSA Income Data Linkages* (in eleven volumes) is still a reliable general resource.

proximately 70% of total positive money income as reported in 2007, for example.

Similar findings occur when looking at quintiles of total positive money income in selected years from 1992 to 2007. Here, there is no evidence of increasing inequality when comparing the increase in the total positive money income share of the lowest fifth of the income distribution to the highest fifth of the income distribution. Imputation does not appear to have distorted these changing shares. There is also relatively little variation when examining GINI coefficients by demographic groups over time, although for some subgroups there are slightly larger differences. Only the differences for Other race are significant at the 95% confidence level.

Acknowledgments

The author greatly appreciates Bula Ghose, Office of the Assistant Secretary for Planning and Evaluation, for preparing all of the tabulations used in this report. A special thank you are also due to Fritz Scheuren, NORC; Edward Welniak, Bureau of the Census and Kendall Swenson, Office of the Assistant Secretary for Planning and Evaluation for reviewing earlier drafts of the report and for their constructive criticism. Any remaining errors are attributable to the author.

References

[1] A. Bustos, Estimates of the Distribution of Income from Survey Data, Adjusting for Compatibility with Other Sources, *Statistical Journal of the International Association of Official Statistics* **31** (2015).

[2] B. Gates, http://www.gatesnotes.com/Books/Why-Inequality-Matters-Capital-in-21st-Century Review.

[3] C. Faulker, More on Income Inequality, a Bustos Sequel, *Statistical Journal of the International Association of Official Statistics* **32** (2016).

[4] B. Ford, Hot Deck Imputation, Theory of Incomplete Data, Volume II, National Academy of Sciences, 1983.

[5] A.E. Jones Jr. and D.H. Weinberg, *The Changing Shape of the Nation's Income Distribution: 1947–1998*, Current Population Reports, P60–204, U.S. Census Bureau, U.S. Department of Commerce, June 2000, 11.

[6] GINI Coefficient, Wikipedia, the free encyclopedia, https://wiki/GINICoefficient.

[7] H.O. Lock, F. Scheuren and H. Nisselson, in: Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing Income Data, Papers and Proceedings, Joint Statistical Meetings, 1980.

[8] J.C. Moore, L.L. Stinson and E.J. Welniak Jr, Income Measurement Error in Surveys: A Review, 1997, U.S. Bureau of the Census, Department of Commerce, Washing D.C., 20233.

[9] T. Piketty and E. Saez, Inequality in the Long Run, *Science* **334**(6186) (2014 May 23).

[10] T. Piketty, Capital in the 21st Century, translated by Arthur Goldhammer, Cambridge: Harvard University, 2014.

[11] C.G. Renfro, Journal of Economic and Social Measurements, *JEMEEZ* **40**(1-4) (2015), 1–474.

[12] Scheuren et al., Studies from CPS-IRS-SSA Income Data Linkages (in eleven volumes), Social Security Administration, 1972–1980.

[13] Technical Paper 54: *Income Nonresponses*: March 1983 CPS, September 1885, U.S. Department of Commerce, Bureau of the Census, Washington D.C., 20233.

[14] Technical Paper 66: *Current Population Survey Design and Methodology*, October 2006, U.S. Department of Commerce, Bureau of the Census, Washington D.C., 20233.

[15] J. Turek and B. Ghose, Documentation for CPS Income Imputation Dataset, Joan.turek@hhs.gov.

[16] J. Turek, F. Scheuren, B. Sinclair-James, B. Ghose and S. Desale, Effects of Imputation on CPS Income and Poverty Series, Joint Statistical Meetings, Washington D.C., August 4, 2009.

[17] J. Turek, F. Scheuren, C. Nelson, E. Welniak Jr., B. Sinclair-James and B. Ghose, *Effects of Imputation on CPS Poverty Series: 1987–2007*, Federal Committee on Statistical Methodology; Washington D.C., November 3, 2009.

[18] J. Turek, F. Scheuren, K. Swenson and B. Ghose, Effects of Imputation on Trends and Demographic Characteristics, 33rd Annual Fall Research Conference, Association of Public Policy Analysis and Management, Washington D.C., November 4th, 2011.

[19] E. Welniak, Effects of the March Current Population survey's New Processing System on Estimates of Income and Poverty, U.S. department of Commerce, Bureau of the Census, Washington D.C., 20233

[20] P. Whoriskey, Income Gap Widens as Executives Prosper, The Washington Post, Sunday, June 19, 2011, page A1 .