

**THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION**

**THE QUALITY OF CENSUS BUREAU
SURVEY DATA AMONG RESPONDENTS
WITH HIGH INCOME**

No. 178

**C. T. Nelson
Bureau of the Census**

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author wishes to thank John Coder, also of the Census Bureau, for providing the computer file upon which most of the analysis in this paper is based, and Elaine Anderson for providing word processing assistance.

INTRODUCTION

Researchers who make use of income statistics gathered from surveys are generally aware of the fact that these data are subject to more than their share of nonsampling error.

Underreporting and high nonresponse rates appear to be consistent attributes of income surveys. This issue is a particularly important one for researchers who require an accurate picture of the upper tail of the income distribution, since (as will be discussed later in this paper) there appears to be a positive correlation between income, nonresponse rates, and levels of net income underreporting.

Researchers who use Census Bureau survey data to examine the upper tail of the income distribution are also somewhat hampered by the topcoding of high income dollar amounts on public use versions of survey microdata files. Topcoding refers to a system which preserves the confidentiality of survey respondents by suppressing very high income amounts on publicly available microdata files. For example, annual incomes from the March income supplement to the Current Population Survey (CPS) are topcoded to \$100,000 per income type for each person. Monthly income amounts from the Survey of Income and Program Participation (SIPP) are topcoded to \$8,333 (or \$100,000/12). Public use files from the 1990 decennial census employ a sliding topcode based on income type (See table 1).

Table 1. Decennial Census Public Use Microdata Sample (PUMS)
File Income Topcodes, by Type: 1990

Wages and salaries	\$140,000
Nonfarm self-employment	90,000
Farm self-employment	54,000
Interest, dividends, rents or royalties, estates or trusts	40,000
Social Security/Railroad Retirement	17,000
SSI, AFDC, other public assistance	10,000
Pensions	30,000
Other income	20,000

One response on the part of users who require accurate income data is to use other data sources to enhance the survey data released by the Census Bureau. One method of enhancement is the use of statistical matches between survey data and administrative records (such as tax returns). Another method of enhancement is to adjust the survey data (through reweighting and/or substitution of dollar amounts) to account for observed differences between survey income distributions and income distributions based on administrative records.

The former method of income survey enhancement is reflected in a Brookings Institution study (Pechman, 1985), that used statistical matches between the CPS and Federal income tax returns to examine various issues associated with tax policy.¹ An example of the latter method of enhancement is an annual set of CPS-based income estimates prepared by the Congressional Budget Office (Williams, 1993).

Neither of these enhancement methods is based on any concrete knowledge of how survey responses differ from the corresponding administrative records of those respondents. This paper, through an examination of exact matches between wages as reported on SIPP and Federal income tax returns, will hopefully provide some of that information. This study differs from previous Census Bureau exact match studies in that it focuses on those at the upper tail of the distribution, while previous studies have tended to use

exact match data to examine data quality and imputation issues as they relate to the entire income distribution.

The types of survey enhancements employed by the Brookings Institution and CBO are based, at least in part, on a desire to use survey data to address issues related to taxes and/or tax policy. Table 2 shows why it is probably unwise to use unadjusted survey data to address tax policy issues. The table compares adjusted gross incomes and taxes paid in 1990 from a simulation based directly on income values as reported on the CPS² with information based on tax returns, as published by the Internal Revenue Service (1992). As the distributions show, the CPS data track the IRS distribution fairly well until the upper tail; CPS simulations over-estimate income and taxes in the \$75,000-\$199,999 range, and miss a significant amount (over one-half) of income and taxes in the \$200,000 or more category. Given these differences, it would be hard to justify using unadjusted CPS data to examine such issues as the effect of the new Administration's proposal to raise the highest marginal tax rate to 36 percent, since the group most affected by this change is significantly underrepresented on the CPS, both in terms of the income they receive and the taxes they pay. This table illustrates why there is a particular concern over the ability of income surveys to adequately represent the upper tail of the income distribution.

This paper is divided into six major sections. The next two sections discuss factors that are leading causes of concern among data users examining the upper tail of the distribution: underreporting and nonresponse (including imputation bias). Section four provides a tentative look at the effect of using administrative data to correct for these types of nonsampling errors. Section five discusses topcoding. While topcoding is certainly more an issue of disclosure than one of data quality, I include it in this study because the effect of topcoding is similar to the other two factors as it represents a limitation on the part of Census Bureau income data to accurately reflect incomes at the upper tail. The final section consists of a summarization of findings and conclusions.

Estimates in this paper are based solely on wages and salaries, by far the largest single income type. Future Census Bureau studies will be based on other income types, including interest. Census Bureau research (Nelson, 1983) using CPS/IRS exact match estimates has shown considerable reporting and imputation bias for this income type.

UNDERREPORTING

The term "underreporting" is really a bit of a misnomer; exact match studies between tax returns and survey data reveal that many survey respondents either over- or under-estimate their

Table 2. Comparison of IRS and CPS-Simulated Estimates of Tax Returns, Aggregate Adjusted Gross Income, and Aggregate Taxes Paid, by Adjusted Gross Income Levels: 1990

(Numbers of returns in thousands. Aggregates in millions of dollars.)

	Returns			Adjusted Gross Incomes			Taxes Paid		
	CPS	IRS	CPS/ IRS Ratio	CPS	IRS	CPS/ IRS Ratio	CPS	IRS	CPS/ IRS Ratio
Total	111,370	113,799	.98	\$3,326,624	\$3,431,906	.97	\$412,642	\$450,808	.92
Under \$25,000	62,955	67,372	.93	700,880	710,520	.99	41,400	45,130	.92
\$25,000-\$49,999	29,233	28,967	1.01	1,034,693	1,036,644	1.00	106,505	110,715	.96
\$50,000-\$74,999	11,805	11,000	1.07	707,554	660,633	1.07	94,765	87,750	1.08
\$75,000-\$99,999	4,028	3,282	1.23	342,615	279,932	1.22	57,040	44,467	1.28
\$100,000-\$199,999	2,864	2,329	1.23	374,659	305,813	1.23	74,576	58,009	1.29
\$200,000 or more	483	850	.56	166,220	438,365	.38	38,356	104,736	.37

Table 3. Comparison of CPS Aggregate Incomes with Independent Estimates, by Type of Income: 1987

(Aggregates in billions of dollars)

	CPS aggregate	Independent aggregate	CPS/Independent ratio
Total income*	\$2,941.4	\$3,297.1	.89
Wages	2,202.4	2,215.9	.99
Nonfarm			
self-employment	172.5	219.8	.79
Farm			
self-employment	15.7	46.9	.34
Social Security	178.7	193.6	.92
SSI	9.5	11.5	.83
AFDC	11.9	16.4	.73
Interest	134.9	244.4	.55
Dividends	38.8	73.7	.53
Rents & royalties	29.2	40.3	.73
Veterans' payments	9.7	14.2	.69
Unemployment			
compensation	10.4	14.0	.75
Workers'			
compensation	9.2	14.2	.65
Private pensions	57.8	125.8	.46
Federal pensions	40.3	40.7	.99
State or local			
pensions	20.3	25.7	.79

*Those income types for which there are independent estimates available.

Source: U.S. Bureau of the Census (1992).

incomes. The term reflects the fact that, as a result of missed recipients and a tendency to under-estimate rather than over-estimate incomes, surveys invariably result in under-estimates of aggregate incomes. As table 3 (based on the March 1988 CPS) shows, the level of net underreporting is strongly associated with income type. Wages, Social Security, and Federal retirement benefits are within 8 percent of the independent benchmarks. (None of the survey aggregates were higher than the comparable independent figures.) Aggregates from some of the other income types are significantly lower than their benchmarks.³ Vaughn (1989) found SIPP to be more accurate than the CPS for some income types (transfer incomes, in particular). Unfortunately, these comparisons do little to help us understand the nature of these differences between surveys and administrative estimates (i.e., whether the differences result from missed recipients or from misreported or misallocated amounts).

To examine this issue, match studies, in which survey respondents are matched to their corresponding administrative records, are quite useful. These studies typically involve the use of identifiers common to the survey and the administrative record systems (such as Social Security Number) in order to examine the relationship between the two on a case-by-case basis. Exact match studies (for example, Scheuren et al., 1981) have proven to be invaluable as indicators of the extent and nature of survey nonsampling error.

Table 4. Percentage of CPS Married Couples Who Under- or Overreported Wages, by IRS Quintile: March 1986 CPS

CPS amount:				
	Over-reported	Mean absolute difference	Under-reported	Mean absolute difference
Lowest quintile	40.8	\$2,288	49.3	\$ 804
Second quintile	40.8	1,703	52.8	1,447
Middle quintile	39.5	1,356	55.0	1,587
Fourth quintile	39.3	1,418	55.8	2,135
Highest quintile	41.0	2,250	55.3	6,365
Top 5 percent	37.0	2,855	60.0	16,377

Source: Coder (1990)

More recently, Coder (1990), in examining the CPS wage and salary income of married couples matched to their tax returns (excluding those who were imputed earnings on the CPS), found that the mean CPS wages of respondents in 1985 was \$32,205, only about 2 percent lower than the wages on these respondents' tax returns (\$32,872). This seemingly small difference, though, is really the result of offsetting reporting errors rather than accuracy. The mean absolute difference between IRS and CPS wages for married couples was \$4,273, and only about one-half of the CPS reported amounts were within 5 percent of the corresponding IRS amounts. Table 4 shows the relationship between income level and income misreporting. The table shows the percentage of married couples with reported wages that differ from their IRS amounts, by IRS quintile (and those in the top 5 percent). Here we see that the tendency to underreport does not seem to be particularly correlated with income level (although those in the top 5 percent are clearly more likely to underreport). What is important to note, however, is the relationship between income level and the mean differences (particularly the underreported means) between CPS and IRS amounts.

SIPP represents another source of data through which we can examine underreporting on income surveys and the relationship between underreporting and income level. As such, it has two advantages over the CPS. First, for CPS match studies (such as the one cited above), Social Security Numbers (SSNs) are not

normally validated and no attempt is made to match survey respondents who fail to provide a SSN. For SIPP, SSNs are validated and attempts are made (through the Social Security Administration) to search for the correct SSN for those who fail to provide one during the survey (except for those who specifically refuse to provide a SSN during the survey). Thus, SIPP is able to provide a more complete universe of survey respondents. Second, in the CPS it is impossible to report longest job earnings of more than \$299,999. Thus, at least some of the difference noted above is the result of the inability to record high income amounts on the CPS. This is not a limitation on SIPP, as it is possible to report monthly wages up to a million dollars on this survey. As result, SIPP is particularly appropriate as a vehicle for examining accuracy among high-income survey respondents.

The SIPP/IRS comparisons in this paper are based on: 1) SIPP wage and salary incomes summed across all 12 months of calendar-year 1990; and 2) wages and salaries as reported on these respondents' 1990 Federal individual income tax returns. The two data sets were linked using SSNs; SIPP SSNs were validated by the Social Security Administration prior to the linkage to ensure accuracy and a more complete matching universe.⁴

The universes for this study consist of: 1) married couples with validated SSNs for both spouses who matched to an IRS married-

joint tax return; and 2) unmarried persons with validated SSNs who matched to an IRS single return. The study was restricted to cases in which wages were present on both the survey and the tax return.

As Coder (1992) points out, there are some conceptual differences between wages as reported on SIPP and Federal tax forms. SIPP collects gross wages, which includes wages deferred from current taxation. Contributions to a 401(k) plan would be a prime example of deferred wages. Tax returns, however, exclude these amounts. Another difference arises from the fact that taxable wages, in some cases, include pay received in-kind, while SIPP collects wages on a strictly "cash" basis.

In all, there were 6,858 matched married-joint filer returns and 6,374 matched single filer returns. Tables 5 and 7 provide some basic statistics for cases with reported SIPP wage amounts (the next section will discuss cases in which the survey data were imputed). There were 5,373 married filer returns with reported SIPP wage amounts (about 78 percent of all matched married filer returns) and 4,947 single filer returns with reported wage amounts (about 78 percent of all matched single filer returns).

For matched married couples, the median reported SIPP wage amount was \$35,361, about 5 percent less than their median IRS wages (\$37,413). The SIPP mean wage (\$39,634) was 8 percent less than

Table 5. IRS and SIPP Wage and Salary Summary Measures: 1990 Married-Joint Filer Returns
 (Restricted to those with reported SIPP earnings)

SIPP				IRS			
	Lower limit	Mean	Share	Lower limit	Mean	Share	
Median		\$35,361		Median		\$37,413	
Mean		\$39,634		Mean		\$43,078	
Gini index		.348		Gini index		.371	
Lowest quintile	X	\$ 10,589	5.4	X	\$ 10,900	5.0	
Second quintile	\$18,843	24,524	12.4	\$19,272	25,793	11.9	
Third quintile	29,774	35,375	17.9	31,791	37,408	17.3	
Fourth quintile	41,227	48,145	24.3	43,483	50,465	23.4	
Highest quintile	56,351	79,488	40.2	58,980	90,740	42.2	
Top 5 percent	84,472	115,344	14.6	89,216	151,771	17.6	

X Not applicable.

the IRS figure, while the SIPP Gini index (.348) was 10 percent less than the IRS Gini index.

Data shown by quintile clearly demonstrate that underreporting patterns are correlated with income level, at least at the upper end of the distribution. SIPP mean wages in the first through fourth quintiles (\$10,589, \$24,524, \$35,375, and \$48,145) were all within 5 percent of these respondents' IRS means, while in the highest quintile the SIPP mean was 12 percent less than the IRS figure. Among the top 5 percent of married-joint filer returns, the difference was 24 percent. An examination of the shares of aggregate incomes received by each quintile shows the highest quintile as the only one in which the SIPP share (40 percent) was less than the IRS share (42 percent). Actually, the SIPP share of aggregate income received by the 80th through 95th percentile (26 percent) appears to be slightly higher than the IRS share (25 percent), but this difference was more than offset by the difference (3 percentage points) among those in the top 5 percent.

Table 6 examines the issue of net underreporting a little more closely by showing for SIPP married couples with IRS earnings in the middle quintile, highest quintile, and top 5 percent, the percentage whose wages were more than 5 percent less than their IRS wages, within 5 percent (more or less) of their IRS wages, and more than 5 percent greater than their IRS wages. The table

Table 6. Differences between SIPP and IRS Wages Among Those With IRS Wages in the Middle Quintile, Highest Quintile, and Top 5 Percent: 1990 Married-Joint Filer Returns

	SIPP/IRS RATIO:					
	Less than .95		.95 - 1.05		Greater than 1.05	
	Percentage	Mean absolute difference	Percentage	Mean absolute difference	Percentage	Mean absolute difference
All quintiles	45.3	\$11,463	30.1	\$127	24.6	\$ 7,249
Middle quintile	48.1	6,692	31.4	187	20.6	6,249
Highest quintile	47.5	32,931	32.6	175	19.9	10,670
Top 5 percent	60.4	76,056	24.1	41	15.6	12,530

shows a pronounced tendency to underreport wages; overall, 45 percent of married couples had a SIPP/IRS ratio of less than .95, 30 percent had a ratio between .95 and 1.05, and 25 percent had a ratio of more than 1.05. Given the fact that, as table 5 shows, SIPP/IRS differences tend to become more pronounced in the top quintile, it is somewhat surprising to note that the middle and top quintiles display quite similar tendencies toward underreporting wages (48 percent of those in both the middle and top quintiles had SIPP/IRS ratios of less than .95). The two quintiles differed, however, in the relationship between mean absolute differences. Overall, the ratio of the mean absolute difference of those with a SIPP/IRS ratio of less than .95 to those with a ratio greater than 1.05 was 1.58 (\$11,463 versus \$7,249). For those in the middle quintile, the ratio was 1.07, while in the top quintile the ratio was 3.09. Those in the top 5 percent showed both a greater tendency to underreport (60 percent had a ratio of less than .95, while only 16 percent were greater than 1.05) and an extremely large mean absolute difference ratio of 6.07 (\$76,056 versus \$12,530).

Table 7 examines the results of single filer SIPP and IRS matches. Overall, SIPP/IRS differences are much smaller for single filer returns than were observed among married-couple filers. The SIPP median and mean incomes were both within 4 percent of the comparable IRS figures, and the SIPP Gini index (.490) was virtually identical to the IRS Gini index (.492).

Unlike married couples, for single filers there was no widening of SIPP/IRS differences at the top end of the distribution. Both sets of numbers indicate that 19 percent of aggregate wages were received by the top 5 percent of filers. These data indicate that married couples, with their higher overall wage levels, are significantly more vulnerable to net underreporting problems in the highest quintile and top 5 percent than single filers.

NONRESPONSE

Survey literature has long recognized the positive correlation between income level and nonresponse. (See, for example, Lillard et al., 1986.) In the March 1986 CPS, 40 percent of persons with incomes of \$50,000 or more had at least some of their income imputed; the comparable figure for those with incomes of less than \$50,000 was 27 percent (see table 8). Here we see a good illustration of perhaps the two main reasons why the Census Bureau imputes for missing income responses: income nonresponse rates are relatively high and the characteristics of income nonrespondents differ significantly from those of respondents. The table also illustrates how the quality of income data, on average, is tied more closely to the quality of the survey imputation system among high-income persons than among those with lower incomes. Of course, for many of these persons imputed income accounted for a relatively small proportion of total

Table 8: Percentage of Persons With Any Income Items Allocated,
by Total Money Income: March 1986 CPS

All income levels	27.6
Under \$50,000	27.3
Loss	41.7
\$1-\$999	7.1
\$1,000-\$1,999	25.5
\$2,000-\$2,999	26.0
\$3,000-\$4,999	24.8
\$5,000-\$6,999	25.5
\$7,000-\$9,999	26.5
\$10,000-\$11,999	27.7
\$12,000-\$14,999	26.8
\$15,000-\$19,999	27.2
\$20,000-\$24,999	27.9
\$25,000-\$49,999	29.8
\$50,000 or more	39.8

Table 9. Earnings (From Longest Job) Imputation Rates in 1991:
 All Workers; Those in Health Diagnosing Occupations;
 and Lawyers and Judges: March 1992 CPS

(Numbers in thousands)

	Number		Imputation rate
	Total	Imputed	
All workers	134,036	24,443	18.2
Health diagnosing occupations	880	283	32.2
Physicians	577	208	35.7
Dentists	174	50	28.7
Other health diagnosing occupations	129	27	20.9
Lawyers and judges	748	182	24.3
Lawyers	713	171	24.0

income. Indeed, part of the differences in overall imputation levels as one moves up the income distribution is the result of the greater likelihood of high-income persons to receive asset income (which has a very high imputation rate). Since 80 percent of money income comes from earnings, it is important to focus on the earnings nonresponse rates of workers.

Are workers in the highest-paying occupations more likely to have imputed earnings than other workers? Table 9 examines this issue by showing the imputation rates (from the March 1992 CPS) of workers in the two highest-paid occupation groups: health diagnosing occupations, and lawyers and judges. The table clearly shows that workers in these occupations are more likely to have imputed earnings than other workers. Overall, the imputation rate for these workers was 29 percent, compared to the overall imputation rate of 18 percent. Physicians had an imputation rate of 36 percent, and lawyers had a rate of 24 percent. These two occupation groups represent only 1 percent of all workers, though they comprise almost one-third of all workers with annual longest job earnings over \$100,000. Clearly, if one wishes to examine the quality of high-income survey data, it is important to examine imputed cases closely.

Table 10 examines the relative accuracy of matched SIPP/IRS married-joint returns in which SIPP earnings were imputed. A comparison between table 5 (married couples with reported SIPP

Table 10. IRS and SIPP Wage and Salary Summary Measures: 1990 Married-Joint Filer Returns

(Restricted to those with imputed SIPP earnings)

	SIPP				IRS			
	Lower limit	Mean	Share		Lower limit	Mean	Share	
Median		\$35,417			Median	\$38,616		
Mean		\$42,043			Mean	\$50,361		
Gini index		.376			Gini index	.439		
Lowest quintile	X	\$ 11,680	5.5		X	\$ 11,434	4.5	
Second quintile	\$19,608	24,936	11.9		\$ 20,761	26,703	10.6	
Third quintile	30,003	35,284	16.8		32,563	38,564	15.3	
Fourth quintile	40,148	46,847	22.3		44,011	51,758	20.6	
Highest quintile	55,016	91,122	43.6		61,466	122,748	49.0	
Top 5 percent	89,499	160,652	19.3		109,279	262,054	26.3	

X Not applicable.

earnings) and table 10 shows that: 1) imputed cases exhibit much larger levels of differences between IRS and SIPP wages (not a particularly surprising finding); and 2) as was true of cases with reported SIPP earnings, SIPP/IRS differences widen significantly at the top of the income distribution for imputed cases.

A comparison of SIPP/IRS summary measures indicates that SIPP wage imputations are downwardly biased. The SIPP median (\$35,417) was 8 percent less than the IRS median for these respondents. The mean (\$42,043) was 17 percent lower, and the SIPP Gini index (.376) was 14 percent less than the comparable IRS figure. Among reported cases, the SIPP/IRS median, mean, and Gini index differences were 5 percent, 8 percent, and 10 percent, respectively.

Comparisons of mean income by quintile show that, in the lowest four quintiles, SIPP/IRS ratios ranged from 1.02 (lowest quintile) to .91 (fourth quintile). In the highest quintile the ratio was .74; among the top 5 percent it was .61, as the IRS mean was \$101,000 higher than the SIPP mean for this group. The higher concentration of IRS wages in the top quintile is reflected in the shares of aggregate wages received by this group based on the two data sources. IRS wage data show 49 percent of aggregate wages going to the highest quintile (and 26 percent going to the top 5 percent). The comparable SIPP figures were

44 percent and 19 percent, respectively.

One of the reasons for imputing income data is to lessen the bias resulting from the fact that the characteristics of income nonrespondents differ from those of respondents. Comparisons between the IRS and SIPP figures in tables 5 and 10 can be used to examine the effectiveness of the SIPP imputation system for married couples, and the relationship between income level and imputation quality. According to the two tables, the ratio of SIPP mean wages for those who were imputed wages to those who reported wages was 1.06 ($\$42,043/\$39,634$). IRS wage data show that the ratio of those who were imputed SIPP wages to those who reported SIPP wages was 1.16 ($\$50,361/\$43,078$), implying that the SIPP imputation system may not be going as far as it could in accounting for the differences between income reporters and non-reporters. For the top quintile, the SIPP ratio of reporters to non-reporters was 1.15 ($\$91,122/\$79,488$), while the IRS ratio was 1.35 ($\$122,748/\$90,740$). In the top 5 percent, the comparable SIPP and IRS ratios were 1.39 and 1.72, respectively.

Table 11, which shows the percentage of matched cases with SIPP/IRS ratios of less than .95, between .95 and 1.05, and greater than 1.05, demonstrates the extent of the tendency to under-impute wages, particularly at the top end of the distribution. For those with IRS wages in the top 5 percent, for example, 80 percent of the imputed SIPP wage amounts were under-

Table 11. Differences between SIPP and IRS Wages Among Those With IRS Wages in the Middle Quintile, Highest Quintile, and Top 5 Percent: 1990 Married-Joint Filer Returns

(Restricted to those with imputed SIPP earnings)

	SIPP/IRS RATIO:		
	Less than .95	.95 - 1.05	Greater than 1.05
	Percentage	Percentage	Percentage
	Mean absolute difference	Mean absolute difference	Mean absolute difference
All quintiles	52.9	18.3	28.8
	\$24,610	\$178	\$16,487
Middle quintile	53.4	21.5	25.2
	8,330	318	10,856
Highest quintile	67.9	14.7	17.4
	71,843	68	52,941
Top 5 percent	80.3	7.9	11.5
	188,809	2,197	71,325

imputed by more than 5 percent; the mean IRS/SIPP difference for these filers was \$188,809.

Single filers with imputed SIPP wages, in general, exhibit smaller SIPP/IRS differences than married couples with imputed wages (see table 12). The overall SIPP median wage for these filers was \$12,902, surprisingly close to their IRS median of \$13,055. The IRS mean was 8 percent higher than the SIPP mean (\$16,794 versus \$15,507), while the IRS Gini index (.456) was 7 percent higher than the SIPP Gini index (.426). By quintile, SIPP/IRS mean wage and salary ratios ranged from 1.05 (second quintile) to .88 (highest quintile). Thus we see the same type of widening of SIPP/IRS differences as we move up the income distribution as was observed for married couples with imputed wages, although the range of change appears narrower among single filers. (For married couples, the SIPP/IRS ratios by quintile ranged from 1.02 to .74.) There was a particularly large difference in SIPP/IRS ratios between married couples and single filers in the top 5 percent of their respective distributions. For single filers, the SIPP/IRS mean wage ratio was .85; among married couples, it was .61.

COMPARISONS OF CORRECTED SIPP WAGES WITH IRS DATA

In theory, if one uses an IRS/SIPP exact match to correct SIPP misreported wages (by substituting IRS wages for SIPP wages in

Table 12. IRS and SIPP Wage and Salary Summary Measures: 1990 Single Filer Returns
 (Restricted to those with imputed SIPP earnings)

	SIPP			IRS		
	Lower limit	Mean	Share	Lower limit	Mean	Share
Median		\$12,902		Median		\$13,055
Mean		\$15,507		Mean		\$16,794
Gini index		.426		Gini index		.456
Lowest quintile	X	\$ 2,777	3.6	X	\$ 2,684	3.2
Second quintile	\$ 5,107	7,568	9.7	\$ 4,847	7,170	8.5
Third quintile	9,901	12,834	16.7	9,864	13,081	15.6
Fourth quintile	15,423	18,663	24.0	16,240	20,423	24.3
Highest quintile	22,710	35,583	46.1	25,131	40,226	48.4
Top 5 percent	37,706	57,521	18.7	41,471	67,675	20.3

X Not applicable.

matched IRS/SIPP records), the resulting wage distribution should match relatively closely with an independently derived IRS wage distribution. Table 13 examines this issue for high-income filers by comparing a percent distribution (for returns with wages of \$50,000 or more) of the IRS wages of SIPP respondents to an IRS distribution of wages based directly on tax returns. Of course, there are at least two major reasons to suspect that this comparison may not be entirely valid. First, as pointed out in the Introduction, not all SIPP wage earners matched to tax returns. Some SIPP respondents, particularly those who refused to provide their SSN, would not have an IRS wage assigned to them. Second, some IRS tax filers are clearly not in the SIPP survey universe. Those living abroad who file U.S. tax returns and those living in institutions represent two components of the tax filing universe who would not have any possibility of inclusion in SIPP.

In constructing tables 13, the first of the two concerns was addressed by showing two percent distribution of SIPP respondents: one based on matched SIPP/IRS records and the other based on all SIPP respondents (including nonmatches). IRS wage amounts were assigned to unmatched SIPP records through a regression based on the relationship between IRS and SIPP wages among matched records. The second concern (SIPP/IRS universe differences) was not addressed, as it is unclear at this time how one might adjust the tax filing universe to conform to a Census

Table 13. Percent Distribution of Returns with Wages of \$50,000 or More in 1990: IRS Data and IRS Tax-Corrected SIPP Data (Before and After Inclusion of Nonmatches)

	IRS Data	IRS-Corrected SIPP Data	
		Before Nonmatch Inclusion	After Nonmatch Inclusion
\$50,000 or more	100.0	100.0	100.0
\$50,000-\$74,999	67.0	68.7	68.7
\$75,000-\$99,999	18.6	18.3	18.3
\$100,000-\$199,999	11.2	10.8	10.8
\$200,000 or more	3.2	2.3	2.2

Bureau survey universe. For that reason, the findings in Table 13 should be considered as extremely tentative.

The table indicates that, after correcting individual survey wage records for net underreporting, the distribution of wages still appears downwardly biased at the very top end of the distribution (those with incomes of \$200,000 or more). Data based directly on tax returns show that 3.2 percent of the returns with wages of \$50,000 or more had wages of \$200,000 or more. The IRS-corrected SIPP distribution shows 2.3 percent in this category. Including nonmatches does little to change the results, indicating that nonmatches appear to be distributed fairly evenly throughout this distribution.

Possible reasons that come to mind for the differences between the two distributions at the high end include: 1) a lack of sensitivity on the part of survey noninterview adjustment procedures to account for high-income households that refuse to participate in the survey; and 2) a lack on sensitivity on the part of survey weighting procedures to account for differential attrition rates of high-income households from SIPP. As SIPP is a longitudinal survey, adjustments to account for differential attrition rates are an important component of SIPP weighting procedures.

These findings indicate that it might be worthwhile to experiment

ENDNOTES

1. See Clark (1986) for an evaluation of CPS/IRS statistical match results.
2. Actually the differences shown here probably overstate the differences between IRS and survey data, since they also reflect census capital gains imputations (since capital gains are not part of the Census Bureau income concept). For more details see U.S. Bureau of the Census (1988).
3. For more information on independent estimate comparisons, see U.S. Bureau of the Census (1992).
4. Under an agreement with the IRS, the Bureau of the Census receives an extract of all Federal individual tax returns annually. This extract is used in the development of population estimates and for evaluating the quality of survey data. In accordance with Title 13, release of information which would allow the identification of individual survey respondents is prohibited. These linkages occur at the Bureau of the Census and the resulting linked files are maintained within the Bureau.
5. Although SIPP, CPS and the decennial census all employ topcoding, the CPS is used here to examine the effect of public use topcoding because it is the most commonly used income file produced by the Census Bureau, particularly for inequality research.

with another type of adjustment (probably a reweighting procedure), after adjusting individual wage records for misreporting, if one truly wants to come up with a survey wage distribution that matches an administrative record distribution at the upper tail.

TOPCODING

Even though there appear to be good reasons for data users to be aware of the limitations of high-income survey data resulting from reporting and imputation biases, one could make the strong argument that the topcoding of microdata is by far the most important limitation associated with the use of high-income data from the Census Bureau. Data from the March 1992 CPS, for example, show that the weighted estimate of persons with topcoded earnings (those with longest job earnings of over \$99,999) was 1.5 million, and the amount of earnings suppressed by the CPS public use processing system was \$62 billion, about 2.2 percent of aggregate longest job earnings.⁵ The 1991 Gini index on longest job earnings was .481 based on the non-topcoded CPS file and .470 based on the topcoded file.

Thus, income topcoding represents an important limitation to users who wish to use survey files to examine such critically important social issues as the rise in income inequality over

Table 14. Distribution of Longest Job Earnings Among CPS
Topcoded Respondents: 1991

(Numbers in thousands)

	Number	Cumulative percent
Total	1,509	-
\$100,000	410	27.1
\$100,001 to \$109,999	112	34.6
\$110,000 to \$119,999	146	44.3
\$120,000 to \$129,999	233	59.7
\$130,000 to \$139,999	78	64.9
\$140,000 to \$149,999	72	69.6
\$150,000	88	75.5
\$150,001 to \$174,999	64	79.8
\$175,000 to \$199,999	77	84.9
\$200,000	64	89.1
\$200,001 to \$249,999	35	91.5
\$250,000	40	94.1
\$250,001 to \$299,998	16	95.2
\$299,999 or more	73	100.0

time. Our goal at the Census Bureau in the long run is to work toward adapting our suppression procedures to provide users with more accurate data at the high end of the income distribution without compromising our disclosure rules. In the meantime, it is important for users to know how these topcoded cases are distributed on the internal Census Bureau files.

Table 14 shows the earnings distribution of respondents from the March 1992 CPS with topcoded earnings. The distribution shows that 410,000 of the 1.5 million topcoded cases (or about 27 percent of the total) reported or were imputed an earnings value of \$100,000, or \$1 more than the public use topcode. Thus, if the CPS public use topcode was changed from \$99,999 to \$100,000, the number of CPS respondents with topcoded earnings would be 1.1 rather than 1.5 million persons. Around three-fourths of the topcoded cases had longest job earnings of \$150,000 or less; 89 percent had longest job earnings of \$200,000 or less.

Seventy-three thousand persons, or about 5 percent of those with topcoded amounts, had longest job earnings on the CPS of \$299,999, which really means "\$299,999 or more" since the CPS questionnaire does not allow the reporting of longest job dollar amounts over this figure. This internal CPS topcode means that the \$62 billion figure quoted above as the aggregate suppressed CPS earnings figure is itself an under-estimate, since it also reflects the internal topcoding of CPS amounts. The internally

topcoded aggregate is not insignificant. For example, if the average earnings of persons with earnings of \$299,999 or more was \$450,000, the aggregate amount of earnings missed on the internal CPS file in 1992 would have been \$11 billion.

Coder (1990) found that internal topcoding had a significant effect on CPS-IRS comparisons. Based on married couples with reported amounts that were internally topcoded, he found that the suppression of earnings amounts over \$299,999, although it only affected 6 sample cases, accounted for 25 percent of the net underreporting for reported cases and 13 percent of overall net underreporting.

As surveys move to computer-assisted interviewing, we hope to build in a thorough verification process that will allow internal topcodes to be raised (and eventually phased out).

CONCLUSION

In examining the wage distributions of SIPP respondents who have been exact-matched to their tax returns, it has been shown that significant survey net underreporting exists, and that there is a positive correlation between net underreporting and income level. This was particularly true among married-couple filers. A comparison between married couples in the highest quintile and those in the middle quintile showed that those in both quintiles

were equally as likely to underreport wages, but that the mean underreported amount was much higher for the top quintile. Among those in the highest 5 percent, underreporting was much more common than it was for the rest of the distribution.

An examination of imputed wage amounts showed significant downward bias, particularly among married couples. As was true among cases with reported wages, the negative correlation between data quality and income level was strong, and those at the very top of the distribution showed significant under-imputation bias. Overall, the SIPP wage imputation system is probably not going as far as it should in accounting for the differences between income reporters and non-reporters. This appears to be particularly true among high-income married filers.

It has also tentatively been shown that it might be worthwhile to experiment with additional adjustments, after individual wage records have been corrected, if one wanted to construct a SIPP wage distribution that matched a distribution based on IRS wages.

Topcoding remains an important issue, and represents a real limitation for users wishing to examine income inequality and other issues that require detailed information at the upper tail. Internal topcoding will hopefully be lessened by the validation checks that computer-assisted interviewing will allow, and we will work with the user community to try to provide more

information about the upper tail of the distribution without compromising Census Bureau disclosure rules.

Obviously, there is much more research to be done in the area of merging administrative and survey data to examine issues related to data quality. Property income (interest, dividends, etc.) represents an important area (and one in which there is almost certainly a relationship between data quality and income level). It should be remembered that wages are considered one of the "better" income types in terms of survey data quality, and evaluations of other income types will surely provide important insights that will aid future survey questionnaire and imputation design.

Another extension of the research discussed here would be to examine the quality of high-income reporting and imputation over time. Since increasing inequality was one of the key income issues of the 1980's, it would be interesting, knowing how inequality measures are often driven by high-income households, to examine whether surveys understated the true increase in inequality over the decade (particularly during the early- to mid-1980s, when income inequality was growing at a fast pace).

REFERENCES

Clark, Philip, "Estimating After-Tax Income Distributions Using Matched IRS-CPS Data," paper presented at the Annual Meeting of the American Statistical Association, Chicago, Illinois, August 18-21, 1986, under the sponsorship of the Business and Economic Statistics Section.

Coder, John F., "Using Administrative Record Information to Evaluate the Quality of the Income Data Collected in the Survey of Income and Program Participation," paper presented at Statistics Canada Symposium 92, November 1992.

Coder, John F., "Exploring Nonsampling Errors in the Wage and Salary Income Data from the March Current Population Survey," paper presented at the Allied Social Sciences Association/Society of Government Economists Meetings, Washington, D.C., December 28-30, 1990.

Internal Revenue Service, Statistics of Income Bulletin, Spring 1992, Washington, D.C., 1992.

Lillard, Lee; Smith, James P.; and Welch, Finis; "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," Journal of Political Economy, Volume 94, Number 3, Part 1, June 1986, pp.488-506.

Nelson, Charles, "A Comparison of CPS and IRS Interest Amounts from the March 1983 CPS-1982 IMF Exact Match File," unpublished Census Bureau memorandum, 1983.

Pechman, Joseph A., Who Paid the Taxes, 1966-85, The Brookings Institution, Washington, D.C., 1985.

Scheuren, Frederick H.; Oh, Loch; Vogel, Linda; and Yuscavage, Robert; "Methods of Estimation for the 1973 Exact Match Study," Studies from Interagency Data Linkages, Report No. 10, U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, SSA Publication Number 13-11750, January 1981.

U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 180, Money Income of Households, Families, and Persons in the United States: 1991, U.S. Government Printing Office, Washington, D.C., 1992.

U.S. Bureau of the Census, Current Population Reports, Series P-23, No. 157, Household After-Tax Income: 1986, U.S. Government Printing Office, Washington, D.C., 1988.

Vaughn, Denton R., "Reflections on the Income Estimates from the Survey of Income and Program Participation," ORS Working Paper Number 39, U.S. Department of Health and Human Services, Social Security Administration, Office of Research and Statistics, September 1989.

Williams, Roberton, "Comparing CBO and Census Income Statistics," CBO Papers, Congressional Budget Office, June 1993.