

RESEARCH ON CHARACTERISTICS OF SURVEY OF INCOME AND PROGRAM PARTICIPATION NONRESPONDENTS USING IRS DATA

Mark R. Hendrick, Karen E. King and Julia L. Bienias, U.S. Bureau of the Census¹
Mark R. Hendrick, U.S. Bureau of the Census, Washington, DC 20233

I. MOTIVATION FOR NONRESPONSE RESEARCH

Nonresponse is a critical problem for the Survey of Income and Program Participation (SIPP). The SIPP is a nationally representative longitudinal survey conducted by the Census Bureau. The survey collects information about the financial situation of persons, families, and households in the noninstitutionalized population of the United States. Being a longitudinal survey, SIPP requires multiple interviews over a period of years. Nonresponse increases with successive interviews, causing the sample to become less representative of the population it is approximating. This issue creates concerns about the size of household and person nonresponse bias in the cross-sectional and longitudinal estimates of the SIPP.

Previous SIPP panels include six to eight interviews for up to a two and a half-year period. The upcoming 1996 panel will include twelve interviews over a four year period. This increase in the number of interviews is expected to increase nonresponse over time leading to an increase in nonresponse-related bias. Learning more about the nonrespondent group is considered a major nonresponse research goal.

IRS earnings data can be used as a benchmark for comparing SIPP earnings estimates. Throughout this research, we have treated the IRS data as truth. Through the use of IRS earnings data as a tool to compare respondents to nonrespondents and IRS earnings to SIPP earnings, a better understanding of SIPP nonrespondents can be gained. The availability of 1990 IRS earnings data and the ability to match these data with the 1990 SIPP panel allowed us to conduct nonresponse research.

With IRS earnings as a benchmark, we can determine if SIPP is obtaining accurate earning amounts or is under-reporting earnings through field interviews. A major research topic at the current time is whether SIPP is actually under-reporting income totals compared to other surveys and various administrative benchmarks.

This paper focuses on earnings comparisons for various respondent groups and IRS filing classification groups. We will define the methodology for various components of the research, present results, and then give conclusions and possible areas of future research.

¹

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. We wish to thank Vicki Huggins, Dave Ferraro, and Dale Garrett for their comments on this paper; and Sandy Carnegie for her help in preparing the final version.

II. METHODOLOGY

In this study, an IRS file containing IRS information from April 1991 filings for the tax year 1990 was matched by social security number to the 1990 SIPP panel first interview respondents. Approximately 51% of SIPP respondents (28,046) matched to an IRS tax return. The IRS file contains income variables whose composition depends on filing status. For purposes of analysis, we combine the filing statuses into three filing groups, based somewhat on marital classes:

1. The "single" group consists of single filers or surviving spouse with dependent child filers.
2. The "married" group consists of married/joint filers or married/separate filers.
3. The "head of household group" consists of head of household filers or husband filing separate (with a wife exemption) filers.

The IRS file contains several income variables, including wages (earnings), total income, adjusted gross income, social security income, and total interest income. We had to determine which of these variables was comparable to a variable on the 1990 panel SIPP file. There are many differences between the IRS definition of total income and the SIPP definition of total income. "Earnings" is the only income variable that has a similar definition from the IRS file to the SIPP file, therefore, it was used for our research.

Due to differences between SIPP and IRS earnings, other adjustments were also necessary. Individuals under the age of 15 (a total of 114) had to be dropped from the research because SIPP does not collect earnings data on them, yet they can still have earnings and file a tax return. Also, individuals with negative IRS earnings (a total of 9) were dropped because SIPP does not collect negative earnings data. Finally, individuals with earnings greater than or equal to \$1,000,000 (a total of 7) were dropped due to earnings limits on the SIPP data sets.

Approximately 3,872 of the matched cases (88% of earnings for married individuals are combined for a joint earnings figure) had zero earnings during 1990. Table 1 depicts zero and positive SIPP earnings cross-classified by zero and positive IRS earnings.

Table 1. (IRS/SIPP zero Earnings)

	IRS zero	IRS > zero
SIPP zero	2580	872
SIPP > zero	424	17056

Despite some of these individuals (424) having non-zero SIPP earnings, we decided to drop the 3,872 cases because earnings should not be modeled for people with zero earnings. In addition, these 3,872 cases would create a spike in the data causing regression assumptions to fail. To more clearly understand zero earnings individuals, we observed the relationship between SIPP and IRS reported earnings, in terms of whether those earnings are zero or not. Based on McNemar's Test, we can conclude a strong relationship exists ($p < 0.0001$). Therefore, these 3,872 cases should be studied separately as future research.

We classified the total number of IRS/SIPP matched individuals into the three filing groups discussed above. The largest filing group, with 7,458 units of analysis is single filers. The married filing group contains 7,072 units of analysis. It is important to note that most (95%) earnings for the married filing group are combined or paired data. The married filing group contains combined SIPP earnings for the married/joint filers, and individual SIPP earnings for married/separate filers. The head of household filing group consists of 2,526 units of analysis. Therefore, the total units of analysis equals 17,056 (7,458 + 7,072 + 2,526). The total number of individuals used to compute earnings for the 17,928 total units of analysis follows: 7,458 single + 7,072 married + (7,072 * .95) married joint + 2,526 head of household = 23,788.

In order to study IRS and SIPP earnings for nonrespondents, the total number of IRS/SIPP matched individuals were classified into a respondent group and a nonrespondent group. It is important to note that all IRS/SIPP matched individuals are first interview respondents. For this study, individuals are classified as nonrespondents if they became nonrespondents during any month of calendar year 1990. The remaining individuals, who were respondents during all of 1990, are classified as respondents. 1990 is used as the time period to determine response/nonresponse classification because IRS earnings data, for this research, exists for 1990 exclusively. Therefore, the first calendar year of the 1990 SIPP panel is used for this research. Of the 17,056 total units of analysis, 15,098 (89%) are respondents and 1,958 (11%) are nonrespondents. By filing group, the married group has the highest response rate (90%), followed by the single group (88%) and the head of household group (86%).

In order to compare IRS and SIPP earnings for respondents as well as nonrespondents, annual 1990 earnings for nonrespondents had to be approximated by the following procedure. Approximate calendar year earnings for those people we classified as "nonrespondents" were determined by weighting up the total of their reported earnings to represent 12 months. This procedure assumes these individuals accumulated earnings at the same rate during their periods of nonresponse as during their periods of responding to the SIPP. It should be noted that although we have confidence in this adjustment, due to the wide distribution of months of nonresponse, it is only an approximation. Graph 1 shows the distribution of the total number of months where nonrespondents had missing data during 1990.

III. RESULTS

A. Overview

For this research, the four earnings-respondent categories (SIPP earnings for respondents, SIPP earnings for nonrespondents, IRS earnings for SIPP respondents, and IRS earnings for SIPP nonrespondents) were cross-classified by the three filing status groups (married, single, and head of household). Further breakdowns of the filing groups and respondent groups were done by race. For the analysis of earnings, weighted mean and range frequencies were used. The 1990 calendar year longitudinal weight was used for analysis of the weighted mean. For additional background regarding SIPP weighting please refer to King (1990a and 1990b).

The results of this research are divided in three parts. The first part addresses the question "How do IRS earnings data differ for respondents and nonrespondents?" An appropriate way to address this question is to fit the following regression model (which will be referred to as Model I) (R/NR refers to respondent or nonrespondent):

$$\text{IRS} = \mu + \beta_1 \text{"single"} + \beta_2 \text{"married"} + \beta_3 \text{"R/NR"} + \beta_4 (\text{"single"} \text{ by } \text{"R/NR"}) + \beta_5 (\text{"married"} \text{ by } \text{"R/NR"}) + \epsilon$$

In this model and all following models, "single" equals 1 if single and 0 if not single, "married" equals 1 if married and 0 if not married, "R/NR" equals 1 if respondent and 0 if nonrespondent.

The second part of this research addresses the question, "Does the relationship between SIPP and IRS earnings differ for respondents and nonrespondents, and if so, how?" A logical way to answer this question is by fitting the following regression model, which will be referred to as Model II:

$$\text{SIPP} = \mu + \beta_1 \text{"single"} + \beta_2 \text{"married"} + \beta_3 \text{"R/NR"} + \beta_4 (\text{IRS}) + \beta_5 (\text{IRS by } \text{"R/NR"}) + \beta_6 (\text{IRS by } \text{"single"}) + \beta_7 (\text{IRS by } \text{"married"}) + \epsilon$$

This model also allows us to answer the following questions: Are SIPP earnings lower than IRS earnings? Are SIPP earnings lower for nonrespondents? Are SIPP earnings affected by filing status? Does the relationship between IRS and SIPP earnings differ for respondents and nonrespondents or for filing status?

The two regression models were expanded to include race and related interactions after study of the original models. Dummy variables were used for all of the variables in the regression models except for IRS earnings and SIPP earnings.

Due to the skewness of the data, we decided to initially take the log of both IRS and SIPP earnings, to protect against violations of homoscedasticity.

The third part of the research involves comparing the distributions of SIPP reported earnings and IRS reported earnings, where the distributions are based on collapsing the reported values. IRS and SIPP earnings were divided into three ranges. The following earnings ranges were selected to simulate those used in SIPP longitudinal weighting (King, 1990a): (I) \$0-\$14,399, (II) \$14,400-\$47,999, and (III) \$48,000+. Throughout this research, we treated IRS earnings as "truth." Accordingly, we performed 1-sample Pearson's Chi-squared tests with the SIPP counts in each cell as the "observed" and the IRS counts as the "expected." The distributions were analyzed separately within each filing status-response status combination, as well as within the larger response status groupings.

B. Model I

$$\log(\text{IRS}) = \mu + \beta_1 \text{"single"} + \beta_2 \text{"married"} + \beta_3 \text{"R/NR"} + \beta_4 (\text{"single"} \text{ by } \text{"R/NR"}) + \beta_5 (\text{"married"} \text{ by } \text{"R/NR"}) + \epsilon$$

The above model served as a starting point for the regression analysis. The goal was to arrive at the most parsimonious model and use that model to determine how IRS earnings data differ for respondents and nonrespondents.

The first step of the analysis was to test whether the two interaction terms could be dropped from the model. The resulting F (2,17050) statistic of 4.21 was significant (p = 0.015) at the $\alpha = 0.1$ level. Therefore, the interaction terms were not dropped from the model and the original model was used for analysis. The final model, including the estimated parameter values, follows:

$$\log(\text{IRS}) = 9.5118 - 0.3192 \text{ "single"} + 0.7122 \text{ "married"} + 0.0284 \text{ "R/NR"} - 0.0764 (\text{"single"} \text{ by "R/NR"}) + 0.0771 (\text{"married"} \text{ by "R/NR"}) + \epsilon$$

Considered sequentially, there is a significant main effect for filing status ($F(2,17052) = 1487.72$, $p < 0.0001$) and for response status after filing status is taken into account ($p = 0.0342$). The β_4 term, "single" by "R/NR" was significant ($p = 0.0069$) indicating single respondents had different IRS earnings than other people. To analyze the interaction further, Scheffé multiple comparison tests were conducted to determine if differences existed by response status within the three IRS filing groups. The weighted means and related standard errors (in parentheses) are provided in table 2. When controlled at the $\alpha = 0.1$ level, none of the three multiple comparison tests was significant.

Table 2. (Log(IRS) Earnings)

	Single	Married	Head of Household
Respondents	9.1391 (1.1507)	10.3458 (0.8867)	9.5499 (0.8517)
Nonrespondents	9.1644 (1.0415)	10.2245 (0.9315)	9.5042 (0.7931)

The second step of the analysis included extending the model to include a race main effect term and an "interaction of race by respondent" term. The race category defined filers as black or nonblack ("B/NB"). The resulting full model is:

$$\log(\text{IRS}) = \mu + \beta_1 \text{ "single"} + \beta_2 \text{ "married"} + \beta_3 \text{ "R/NR"} + \beta_4 (\text{"single"} \text{ by "R/NR"}) + \beta_5 (\text{"married"} \text{ by "R/NR"}) + \beta_6 (\text{"B/NB"}) + \beta_7 (\text{"B/NB"} \text{ by "R/NR"}) + \epsilon$$

In this model and all following models, "B/NB" equals 1 if black and 0 if nonblack.

Again, a parsimonious model was desired. First, we dropped the β_7 interaction term from the model and calculated an F statistic. The F (1,17048) statistic was not significant ($p = 0.1821$), indicating that the term should be dropped from the model. The race main effect term was left in the model. Thus, the final model was:

$$\log(\text{IRS}) = 9.5771 - 0.3622 \text{ "single"} + 0.6653 \text{ "married"} + 0.0120 \text{ "R/NR"} - 0.0673 (\text{"single"} \text{ by "R/NR"}) + 0.0840 (\text{"married"} \text{ by "R/NR"}) - 0.1581 (\text{"B/NB"}) + \epsilon$$

The race main effect was significant ($F(1,17049) = 39.91, p < 0.0001$) when considered sequentially indicating that race affects IRS earnings. We conclude that response status affects earnings differently based on filing status.

C. Model II

$$\log(\text{SIPP}) = \mu + \beta_1 \text{ "single"} + \beta_2 \text{ "married"} + \beta_3 \text{ "R/NR"} + \beta_4 \log(\text{IRS}) + \beta_5 (\log(\text{IRS}) \text{ by "R/NR"}) + \beta_6 (\log(\text{IRS}) \text{ by "single"}) + \beta_7 (\log(\text{IRS}) \text{ by "married"}) + \epsilon$$

The first step in analyzing this model was to test whether the β_6 and β_7 interaction terms could be dropped from the model. These two interaction terms were tested simultaneously because they both involve IRS filing status. The resulting $F(2,17048)$ statistic was significant ($p < 0.0001$), therefore, β_6 and β_7 were not dropped. Next, the β_5 interaction term was tested to observe whether it should be dropped from the model. The test $F(1,17048)$ was significant ($p < 0.0001$) allowing us to leave the interaction term in the model. The estimate of the final model was:

$$\log(\text{SIPP}) = 2.3842 - 0.0491 \text{ "single"} + 0.9299 \text{ "married"} - 1.5985 \text{ "R/NR"} + 0.7478 \log(\text{IRS}) + 0.1617 (\log(\text{IRS}) \text{ by "R/NR"}) + 0.0043 (\log(\text{IRS}) \text{ by "single"}) - 0.0831 (\log(\text{IRS}) \text{ by "married"}) + \epsilon$$

For this model, the intercept was significant ($p < 0.0001$) indicating that, in general, SIPP earnings are lower than IRS earnings. The R/NR main effect was significant ($p = 0.0001$), showing that SIPP earnings are lower for nonrespondents when compared to respondents. Also, we can conclude that SIPP reported earnings are affected by IRS filing status by observing that the "married" main effect is significant ($p < 0.0001$). In addition, we can conclude that the IRS earnings by response status interaction and the IRS earnings by married interaction both significantly affect SIPP earnings ($p < 0.0001$, for both).

The second step of this analysis involved expanding the model to include a race main effect and a race by $\log(\text{IRS})$ earnings interaction. A nonsignificant ($p = 0.7083$) $F(1,17048)$ value was calculated after dropping the race by $\log(\text{IRS})$ interaction. The full extended model, including the parameter estimates, follows:

$$\log(\text{SIPP}) = 2.4747 - 0.1283 \text{ "single"} + 0.8564 \text{ "married"} - 1.5989 \text{ "R/NR"} + 0.7413 \log(\text{IRS}) - 0.7990 (\text{"B/NB"}) + 0.1613 (\log(\text{IRS}) \text{ by "R/NR"}) + 0.0108 (\log(\text{IRS}) \text{ by "single"}) - 0.0774 (\log(\text{IRS}) \text{ by "married"}) + \epsilon$$

This model leads to the same conclusions found in the main effects model from above. In addition we find race affects the relationship between IRS and SIPP earnings data, as the race main effect term is significant ($p < 0.0001$).

D. Earnings Range Analysis

Graphs 2-5 provide visual models of the relationship between the distribution of IRS earnings and the distribution of SIPP earnings within the three earnings groups (labeled as low, middle, and high on the graphs). The earnings are cross-classified by response status and filing status groups. Chi-squared goodness of fit tests were used to check for differences between the distributions.

For respondents, a significant difference was found overall ($p < 0.0001$), for the single group ($p < 0.0001$), for the married group ($p < 0.0001$), and for the head of household group ($p < 0.0001$). For nonrespondents, a significant difference was found overall ($p < 0.01$), for the married group ($p < 0.0001$), and for the head of household group ($p < 0.0899$). However, the single nonrespondent group failed to show a significant difference ($p = 0.5144$).

Analysis of graphs 2-5 appears to indicate a general shift of SIPP earnings towards the low earnings group in comparison to IRS earnings. This shift occurs within both response statuses and across filing groups. This appears to represent SIPP overestimation at the low earnings level and/or SIPP underestimation at the high earnings level.

IV. CONCLUSION

It is important to note that results for this research can only be generalized to people who respond in the first interview, as well as people who are similar to those who were matched. The fact that the unmatched population (49%) is large creates biases in the results which are difficult to measure.

From model I we can conclude that IRS earnings differ by filing status as we would expect. Also, the significant effect of response status on earnings is mediated by filing status. In addition, IRS earnings vary by race and for single respondents.

Results from model II indicate that SIPP earnings are lower than those for IRS. Also, SIPP earnings for nonrespondents are lower than SIPP earnings for respondents when all filing statuses are combined, and SIPP earnings are affected by IRS filing status as well as by race. This relationship between IRS and SIPP earnings depends on the existence of filing status, race, and response status in the model. Therefore, any IRS-based imputation scheme would require the usage of those variables.

This research also appears to verify an under-reporting of earnings in SIPP, based on the assumption that the IRS information is the truth and that the concept of earnings is the same between the two sources.

V. FUTURE RESEARCH

Several other SIPP characteristics, (such as marital status, number of imputations, Census division, and average household income), can be used to define cells for respondent/nonrespondent and SIPP/IRS earnings comparisons in order to shed more light on the differences.

Also, the SIPP noninterview adjustment procedure can be simulated based on new SIPP defined cell criteria taken from this and other related research. After classifying the matched individuals (from this study) into appropriate cells, response bias can be estimated by using the IRS earnings data and SIPP response status. The cell criteria can be modified with the goal to minimize response bias. This analysis may provide information valuable to SIPP nonresponse adjustment.

Future analysis involving other SIPP panels is contingent upon the availability of IRS information.

REFERENCES

King, Karen E., 1990a, "SIPP 87: Specifications for Panel File Longitudinal Weighting of Persons," Internal Census Bureau Memorandum from Waite to Courtland, June 1, 1990

King, Karen E., 1990b, "SIPP 90: Cross-Sectional Weighting Specifications for Wave 1, "Internal Census Bureau Memorandum from Waite to Courtland, September 26, 1990.

