ANALYSIS OF NONRESPONSE EFFECTS ON INCOME AND POVERTY TIME SERIES DATA FROM SIPP

Smanchai Sae-Ung and Franklin Winters
U.S. Bureau of the Census, Washington, D.C. 20233 smanchai.sae-ung@ccmail.census.gov

Key Words: SIPP, Time Series Data, Income, Poverty, Nonresponse Effects, Weight Adjustment

1.0 Introduction

The Survey of Income and Program Participation (SIPP) is a longitudinal survey that provides both cross-sectional and longitudinal estimates. The survey universe of SIPP consists of persons living in the United States households and group quarters. Persons living in military barracks and in institutions, such as prison and nursing homes, are excluded. Each SIPP panel is a multistage probability sample of the survey universe, designed to produce national estimates. A description of SIPP sample design can be obtained from Jabine, King, and Petroni (1990).

Missing data due to nonresponse in the income and poverty time series data from SIPP has been at least partially compensated by using a sample weight adjustment and imputation scheme. The weight adjustment generally consists of two principal components: (a) the noninterview adjustment and (b) the second stage adjustment based principally on raking to match a set of population controls. A detailed discussion of SIPP weight adjustment and imputation can be found in Jabine, King, and Petroni (1990).

Previous studies (Ryscavage, 1994; Winters, 1996; Rust, 1996) have identified the following concerns for these income and poverty time series data. Based on the quarterly estimate of low income household total (the total number of low income households in the SIPP universe) shown in Figure 1, panel data generally yielded (i) a low income household total time series estimate with a peculiar trend that decreased as the panel age increased, (ii) a somewhat doubtfully large reduction in the estimate of low income household total from the first quarter to the second quarter due to seasonal effect, and (iii) an unreasonably large difference in the estimates of low income household total between each pair of overlapping or abutting panels. The first concern is not apparent in Panels 1990 to 1993. This may be related to the recession early in the 1990 decade coupled with the widespread corporation downsizing from the early- to mid-1990's. This recession and corporation downsizing are more likely to affect middle and high income households which generally have obtained an adequate sample weight adjustment in SIPP. Therefore, the first concern (which virtually associates with the steady low income households) is masked by the middle and high income households temporarily having low income during this period.

It is generally believed that the troublesome results discussed in Concerns i to iii are attributable to the bias in the sample weight adjustment of the low income households and a seasonal effect in the first quarter of a year. The study on seasonal effect in Winters (1996) has found that the seasonal effect is strong and thus apparently explains Concern ii.

In this study, in order to provide a basis for modifying the sample weight adjustment procedure to mitigate Concerns i and iii, a method is developed to identify a cause of Concerns i and iii and to assess the corresponding bias magnitude in the estimate. This method uses a probabilistic approach to derive an estimate for a population characteristic time series based only on the response rate at every wave among the response (interviewed) sample units having the characteristic under consideration in Wave 1. Since Concerns i and iii may affect estimates of characteristics other than the low income household estimate, this method is developed such that it is applicable to estimates of any statistically well behaved characteristics in SIPP or in other longitudinal surveys similar to SIPP. The derivation of this method is provided in Section 2.0. To demonstrate the ability of the method to fulfill its goal, it is applied to calculate the time series of the low income household total estimate as described in Section 3.0. Based on the results of the calculation in Section 3.0, a suggestion is then made on how to modify the SIPP sample weight adjustment procedure to possibly mitigate Concerns i and iii.

2.0 Methodological Derivation

Consider a characteristic (x) of a unit or member (u) of the SIPP universe. For example, with regard to a low income household in the SIPP universe, a unit u in the SIPP universe is a household and the characteristic x of the unit is having low income.

Let S_1 = a set of sample units actually included (having final weights) in the sample of a SIPP panel in Wave 1. Namely, if the unit u is a household then each sample unit will be a household actually included in the sample in Wave 1. Sx_1 = a subset of S_1 containing only the response sample units that have characteristic x. Sxc_1 = a subset of S_1 containing only the response sample units that do not have characteristic x. Hereinafter, 'characteristic xc' will be used to denote 'not having characteristic x'.

Let S_j = a set of sample units having final weights in the sample of a SIPP panel in Wave j for $j \geq 2$. Sx_j = a subset of S_j containing only the response sample units that have characteristic x. Sxc_j = a subset of S_j containing only the response sample units that do not have characteristic x.

$P[(u \in S_1)\cap(u \in Sx_1)]$ = the probability of a unit in the SIPP universe actually included in the SIPP sample in Wave 1 and having characteristic x in Wave 1.

$P[(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}]$ = the probability of a sample unit in the SIPP universe actually included in the SIPP sample in Wave j and having either characteristic x or characteristic xc in Wave j.

$P[\{(u \in S_1)\cap(u \in Sx_1)\}\cap\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}]$ = the probability of a unit in the SIPP universe actually included in Wave 1 and having characteristic x in Wave 1, and remaining in response Wave j and having either characteristic x or characteristic xc in Wave j.

$P[\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}|\{(u \in S_1)\cap(u \in Sx_1)\}]$ = the conditional probability of a unit in the SIPP universe actually included in the SIPP sample and having either characteristic x or characteristic xc in Wave j given the unit actually included in the SIPP sample in Wave 1 and having characteristic x in Wave 1.

By definition of the conditional probability, $P[\{(u \in S_1)\cap(u \in Sx_1)\}\cap\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}]$ can be expressed as shown in Eq. 2.1 below.

$$P[\{(u \in S_1)\cap(u \in Sx_1)\}\cap\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}] = P[\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}|\{(u \in S_1)\cap(u \in Sx_1)\}] \times P[(u \in S_1)\cap(u \in Sx_1)], \quad \text{for } j \geq 2 \qquad \text{(Eq. 2.1)}$$

Theoretically, the three probabilities in Eq. 2.1 are related to the SIPP survey analysis components in the following manner.

(1) By definition, the inverse of $P[(u \in S_1)\cap(u \in Sx_1)]$ represents the final weight (after second stage weight adjustment) at Wave 1 of each the response sample unit in Wave 1 that has characteristic x in Wave 1.

(2) $P[\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}|\{(u \in S_1)\cap(u \in Sx_1)\}]$ represents the response rate (proportion) at Wave j ($j \geq 2$) of the response sample units in Wave 1 that have characteristic x in Wave 1, and have either characteristic x or characteristic xc in Wave j.

(3) By definition, the inverse of $P[\{(u \in S_1)\cap(u \in Sx_1)\}\cap\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}]$ represents the final weight at Wave j of each response sample units in Wave 1 that has characteristic x in Wave 1, remains in response in Wave j, and have either characteristic x or characteristic xc in Wave j.

3.0 Application to Low Income Household Total Time Series Estimate

To demonstrate (a) how to use the method derived in Section 2.0 to provide a basis for mitigating Concerns i and iii described in Section 1.0 in the SIPP sample weight adjustment procedure, and (b) how well the method provides such basis; an analysis on the estimate of low income household total time series based on Eq. 2.1 is performed using SIPP Panels 1984 to 1993 as shown in Parts 1 to 5 provided below.

To be consistent with the previous studies (Ryscavage, 1994; Winters, 1996), the definition for a low income household defined in Ryscavage's study (1994) will be also used in this study. Namely, a household is characterized as a low income household in a reference month only if its total household income in that reference month is less than the corresponding Federal government's official poverty threshold for the same month. More details can be found in Orshansky (1963, 1965), and Weinberg and Nelson (1998).

Part 1 - Original Estimates of the Low Income Household Total Time Series

The estimate of low income household total time series is calculated based on the original weights of the sample households in the SIPP cross-sectional data file. These weights are calculated based on the current SIPP sample weight adjustment procedures for Panel 1984 to 1993. Hereinafter, this estimate will be referred to as an original estimate. In this study, both the monthly and quarterly estimates of the low income household total time series are calculated, and the results are summarized below.

Figure 2 provides a plot for the original quarterly estimates of low income household total time series for each individual panel. It exhibits the three concerns described in Section 1.0 and they consistently indicate a seasonal effect in the first quarter in each year. Note that the monthly time series indicates that a seasonal effect generally occurs in February.

Part 2 - Wave 2+ Response Rate of Households Having Low Income in Wave 1

Let Wave 2+ denote Wave 2 or higher. For each reference month, the response rate in Wave 2+ among the respondent low income household sample units in Wave 1 is calculated for each individual panel. The response rates among the four reference months are approximately the same, and Figure 3 shows the response rates of Panels 1984 to 1993 in reference month 1. As exhibited in Figure 3, there is apparent but not large variation of the Wave 2+ response rates among all the panels.

Based on all the Wave 2+ response rate data above, a regression analysis is used to obtain an estimate of the response rate at Wave 2+ for a SIPP panel. Let W_j denote Wave j of a SIPP panel, and R_j denote the response rate at Wave j (W_j) for $j \geq 2$. As indicated by the plot of the relationship between R_j and W_j in Figure 3, the response rate R_j decreases exponentially as W_j increases (as the panel age increases). Therefore, a simple semi-log linear regression model as shown in Eq. 3.1 will be used.

$$\ln(R_j) = aW_j + b + e, \quad \text{for } j = 2, ..., n \qquad \text{(Eq. 3.1)}$$

Where ln (R_j) is the natural logarithm of R_j. The symbols a (slope) and b (intercept) are the model parameters to be estimated based on the data of response rates versus wave numbers. The symbol e is a random error which is assumed to have a normal distribution with zero mean and constant variance. The ordinary least square (OLS) estimation method is used to perform the regression analysis and the results are summarized in Table 1. Namely, the standard errors (S.E.) for parameter a and b estimates (Est.) are at most a few percents of their estimates, the adjusted R-squares at least 0.93 (one for perfect fitting), and the results of the F-test and T-test show that the regression model is highly significant. Thus the regression model well represents the response rate.

Each of the predicted response rates (conditional means) from the above regression analysis represents an estimate of the conditional probability, $P[\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}|\{(u \in S_1)\cap(u \in Sx_1)\}]$ defined in Section 2.0.

Part 3 - Comparison of the Response Rates between Low Income Households and All Households

The response rate of low or middle or high income households in Wave 1 is theoretically unknown. It is unlikely that the response rate of low income households is substantially different from that of middle or high income households in Wave 1. Thus it can be assumed that the response rate of low income households is the same as the response rate of all households in Wave 1. Consequently, the response rate (in percent) of the low income households in Wave 2+ can then be approximated by {(the response rate at Wave 2+ among the respondent low income households in Wave 1 calculated in Part 2) - (100 - the response rate of all households at Wave 1)}.

Based on this approach, the response rates of low income households in Wave 2+ are calculated for each reference month of Panels 1984 to 1993.

The difference between the response rates of the low income households and all households for reference month 1 of Panel 1991 is graphically exhibited in Figure 4. The shapes of the two curves in Figure 4 are typical of all other reference months and other panels.

The comparison between the response rates of the low income households and all households reveals the following two points. (1) The response rates of the low income households are all lower than the response rates of all households. Among all panels, the magnitudes of the difference range from about 2% to 6% at Wave 2, and from about 9% to 13% at Wave 8. Thus, the response rates of the low income households are substantially lower than the response rates of all households. (2) The difference between low income household response rates and all household response rates generally increases as the age of the panel increases as shown in Figure 4.

The sample weight adjustment procedures used for Panels 1984 to 1993 did not include the poverty/nonpoverty (low income/not low income) household indicator variable in creating cells (classifications) for the noninterview adjustment, and for the second stage adjustment. This coupled with the response rate of the low income households being substantially lower than the response rate of all households lead us to believe that the low income households are significantly under-weighted in these panels. As a result of the under-weighting of low income households and the increase in magnitude of under-weighting as the panel age increases, (a) the estimate of low income household total time series (Figure 1) of an individual panel spuriously decreases as the panel age increases (Concern i), and (b) between each pair of overlapping or abutting panels, the estimate of low income household total of the elder panel is spuriously smaller than the estimate of low income household total of the younger panel as shown in Figure 1 (Concern iii).

Based on the above analysis, we postulate that a significant cause leading to the bias in the estimate of low income household total time series as expressed in Concerns i and iii is the under-weighting of the low income households from Wave 2+ for Panels 1984 to 1993. This under-weighting has occurred because the low income households have substantially lower response rate than the response rate of all households and the sample weight adjustment procedures do not include the poverty/nonpoverty household indicator variable in the noninterview adjustment and the second stage adjustment.

Part 4 - Assessing the Magnitude of Bias Due to Under-weighting of Low Income Households

To assess the magnitudes of the bias in the original estimate of low income household total time series caused by the under-weighting discussed in Part 3, the low income household weights in Wave 2+ will be modified based on Eq. 2.1 and the estimate of low income household total time series for individual panels will be recalculated using the modified weights.

Since the estimates of low income household total time series for Panels 1984, 1985, and 1986 have more apparent bias associated with Concerns i and iii, the calculation of the modified estimate of low income household total time series was limited to these three panels. The modified estimate of low income household total time series for each individual panel was calculated in the following manner. (1) The estimate of low income household total for Wave 1 is based on the original final weights in Wave 1. (2) For the respondent low income households in Wave 1 which remain in response in Wave 2+, their final weights in Wave 2+ (to be used for the estimate of low income household total in Wave 2+) are calculated based on Eq. 2.1. Namely, the inverse of the probability, $P[\{(u \in S_1)\cap(u \in Sx_1)\}\cap\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}]$ represents the modified final weight at Wave 2+ of a respondent low income household in Wave 1 and remains in response in Wave 2+. The inverse of the probability, $P[(u \in S_1)\cap(u \in Sx_1)]$ represents the original final weight at Wave 1 of a respondent low income households in Wave 1. The conditional probability, $P[\{(u \in S_j)\cap\{(u \in Sx_j)\cup(u \in Sxc_j)\}\}|\{(u \in S_1)\cap(u \in Sx_1)\}]$ can be represented by the response rate at Wave 2+ among the respondent low income households in Wave 1 (Part 2). Therefore, based on Eq. 2.1, the final weight at Wave 2+ for a respondent low income household in Wave 1 which remains in response at Wave 2+ = (its final weight at Wave 1 as provided in the SIPP internal users' file)÷(the response rate at Wave 2+ among the respondent low income households in Wave 1 as calculated in Part 2).

Theoretically, the final weights at Wave 2+ for the respondent low income households which remain in response at Wave 2+ calculated as described above are only approximations because the second stage adjustment for Wave 2+ is implicitly based on the second stage adjustment for Wave 1. Namely, the SIPP universe generally changes somewhat during the life of the panel as reflected by slight changes in population controls used in the second stage adjustment from wave to wave.

The original and modified quarterly estimates of low income household total time series are plotted in Figure 5 for comparison. Based on the comparison exhibited in Figure 5, the following points can be drawn.

(1) By using the modified weight that fully accounts for the lower than overall response rates among the low income households, the modified estimates of low income household total time series for Panels 84 and 86 do not have the peculiar trend that decreases as the panel ages increase (Concern i), as described below using the four vertically shaded curves in Figure 5. Due to seasonal effect in the first quarter (discussed in Section 1.0), the low income household total substantially decreases from the first quarter to the second quarter in each of the two panels. Therefore, for comparing the trends of the original and modified estimates of the two low income household total time series, we pick the first second quarter (instead of the first first quarter) and then start with drawing four horizontal lines starting at the first second quarter to the end of the panel duration as shown Figure 5. To enhance the visualization of the characteristic of the trend of each time series, we vertically shade every region bounded by each time series and its corresponding horizontal line (e.g., Line RS for Panel 84 original time series). Based on the above graphic construction in Figure 5, the following observations are made. For panel 84 in Figure 5, the original time series stays virtually below the horizontal line RS, namely, the original estimates of low income household totals generally decrease after the first second quarter or equivalently as the panel age increases (Concern i). On contrary, the modified time series stays evenly above and below the horizontal line TU, namely, the modified estimates of the low income household totals do not generally decrease after the first second quarter or equivalently as the panel age increases. Similar observation also occurs in Panel 86 as shown in Figure 5. The result of the above comparison indicates that the under-weighting of low income households in the sample weight adjustment procedures is a significant cause of such bias (Concern i).

(2) By using the modified weight that fully accounts for the lower than overall response rates among the low income households, the modified estimates of low income household total time series for Panels 84, 85, and 86 do not have an unreasonably large difference in the estimates of low income household total between each pair of overlapping or abutting panels (Concern iii), as illustrated in Figure 5. For example, at the first quarter of 1985, the difference between the original estimates of the low income household total of Panels 1984 and 1985 is 662,160 households (Point A to Point B in Figure 5), but the difference between the corresponding modified estimates strongly decreases to 192,880 households (Point C to Point D in Figure 5).

Part 5 - Suggestion for Mitigating Concerns i and iii

On the basis of the results of the analysis and discussion in Parts 3 and 4, we make the following suggestions for mitigating Concerns i and iii.

(1) In the SIPP sample weight adjustment procedures for Panels 1984 to 1993, add a poverty/nonpoverty household indicator variable in creating cells for the noninterview adjustment and the second stage adjustment. For the second stage adjustment, it is likely that an adequately accurate population control for low income households is not attainable. Thus, if a population control for low income households is indeed not available, we further suggest that, for simplicity, determine whether excluding the low income households from the raking will yield reasonable estimates prior to pursuing any complex approach.

(2) The SIPP sample weight adjustment procedures for Panel 1996 have included (a) an indicator variable for being in a poverty PSU Stratum as a new variable for creating cells for noninterview adjustment for Wave 1, and (b) household income levels as a new variable for creating cells for noninterview adjustment for Wave 2+. However, a poverty/nonpoverty households indicator variable is not used for creating cells for the second stage adjustment for all waves of Panel 1996. Since a poverty/nonpoverty indicator variable is not directly used in creating noninterview adjustment cells and is not used

at all for creating cells for second stage adjustment, the sample weight adjustment procedure for Panel 1996 may not adequately mitigate Concerns i and iii. Therefore, we suggest that this should be verified later when the Panel 1996 data are available.

4.0 Conclusion

Missing data due to the nonresponse in the income and poverty time series data from SIPP has been at least partially compensated for using a sample weight adjustment and imputation scheme. Previous studies (Ryscavage, 1994; Winters, 1996) have identified the following concerns for these income and poverty time series data. The data of a panel generally yielded (i) a low income household total time series estimate with a peculiar trend that decreased as the panel age increased, (ii) a somewhat doubtfully large reduction in the estimate of low income household total from the first quarter to the second quarter due to seasonal effect, and (iii) an unreasonably large difference in the estimates of low income household total between each pair of overlapping or abutting panels. Concern ii has been explained by Winters (1996).

In this study, a methodology has been developed to identify a cause of Concerns i and iii and to assess its bias magnitude (Section 2.0). Since Concerns i and iii may affect estimates of characteristics other than the low income household total estimate, this method is developed to be applicable to estimates of any statistically well behaved characteristics of SIPP and other longitudinal surveys similar to SIPP. This methodology uses a probabilistic approach to derive an estimate for a population characteristic time series based only on the response rate at every wave of the sample units having the characteristic under consideration in Wave 1. The ability of this methodology to fulfill its goal has been demonstrated by applying it to calculate the estimate of low income household total time series (Section 3.0). As a result of the calculation, we found that a significant cause of the bias associated with Concerns i and iii is the under-weighting of the low income households in the SIPP sample weight adjustment procedures for Panels 1984 to 1993. The under-weighting of low income households happens because the response rate of low income households is substantially lower than the response rate of all households, and the sample weight adjustment procedures do not include a poverty/nonpoverty household indicator variable in the noninterview adjustment and in the second stage adjustment. Based on this finding, we make the following suggestion for mitigating Concerns i and iii for the estimate of low income household total time series.

A poverty/nonpoverty household indicator variable in creating cells for the noninterview adjustment and the second stage adjustment is needed in the SIPP sample weight adjustment procedures. For the second stage adjustment, it is likely that an adequately accurate population control for low income households is not available. Thus, if a population control for low income households is indeed not available, we further suggest that, for simplicity, determine whether excluding the low income households from the raking will yield reasonable estimates prior to pursuing any complex approach. For example, a complex approach may involve finding a set of demographic variables (such as, education level, geographic location, family structure, etc.) highly correlated with low income households to be used for the second stage weight adjustment.

The views expressed in this paper are attributable to the authors and do not necessarily reflect those of the Census Bureau.

References

Orshansky, M. (1963), "Children of the Poor", Social Security Bulletin, Volume 26, July, pp. 3-13.

__________ (1965), "Counting the Poor", Social Security Bulletin, Volume 28, January, pp. 3-29.

Jabine, T.B., King, K.E., and Petroni, R.J. (1990), "Quality Profile - Survey of Income and Program Participation (SIPP)", Bureau of the Census, May.

Ryscavage, P. (1994), "Monitoring the Economic Health of American Households", Current Population Report No. P70-35, Bureau of the Census, May.

Winters, F. (1996), "Income and Poverty Time Series Data from the Survey of Income and Program Participation", ASA/SRM Meeting on SIPP, September.

Rust, K. (1996), "Observations on Current and Future Research into Bias in Poverty Time Series Data from SIPP", ASA/SRM Meeting on SIPP, September.

Weinberg, D.H. and Nelson, C.T. (1998), "Changing the way the United States Measures in Income and Poverty" (Draft), the U.S. Bureau of the Census, January.

Table 1 Regression analysis result.

Ref. Mon.	Parameter a		Parameter b		Adj. R-Sq.
	Est.	S.E.	Est.	S.E.	
1	-.0285	.00092	4.563	.00505	.9398
2	-.0284	.00091	4.564	.00501	.9403
3	-.0284	.00090	4.564	.00497	.9410
4	-.0283	.00090	4.564	.00495	.9412

Figure 1 Concerns i to iii

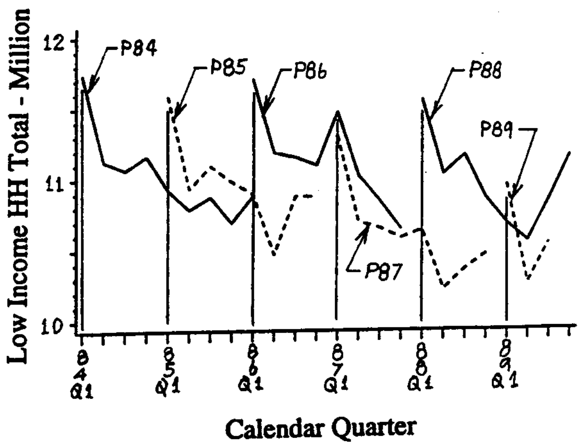


Figure 2 Original quarterly individual estimates for Panels 84 to 93

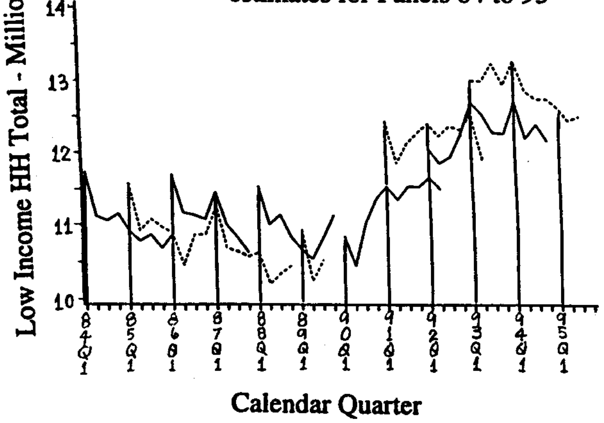


Figure 3 Response rate of Wave 1 low income households, Reference Month 1, Panels 84 to 93

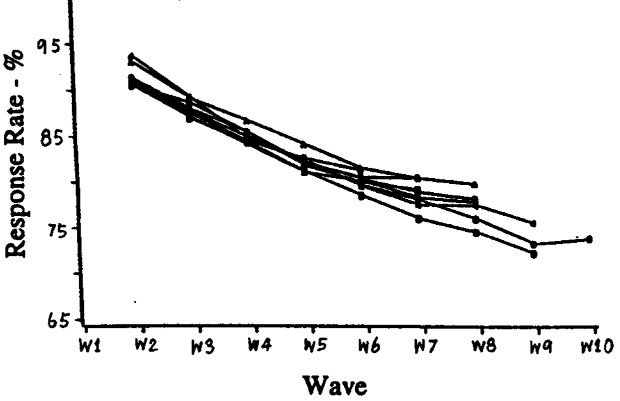


Figure 4 Comparing response rates between low income and all households, Reference Month 1, Panel 91

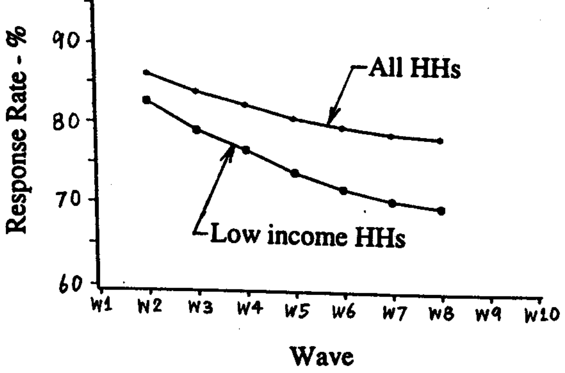
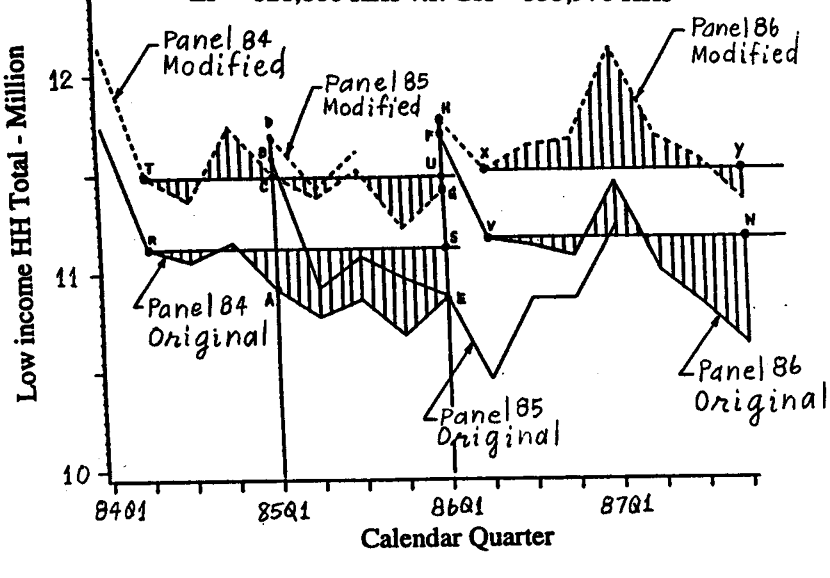


Figure 5 Comparing original and modified estimates of low income household totals

Estimate differences between abutting panels
AB = 662,160 HHs v.s. CD = 192,880 HHs
EF = 826,800 HHs v.s. GH = 353,970 HHs


Low income HH Total - Million
Panel 84 Modified
Panel 85 Modified
Panel 86 Modified
Panel 84 Original
Panel 85 Original
Panel 86 Original
T
U
X
Y
R
S
V
W
A
B
C
D
E
F
G
H
85Q1
86Q1
87Q1
Calendar Quarter