Sampling for Household Financial Characteristics Using Frame Information on Past Income

Arthur B. Kennickell and Douglas A. McManus
Board of Governors of the Federal Reserve System

Paper presented at the 1993 Joint Statistical Meetings, Atlanta, GA

August 1993

The Survey of Consumer Finances (SCF) is designed to gather extensive data on the assets, liabilities, income and other financial characteristics of U.S. households. To be useful, a sample design for the survey must address three problems. First, the sample must provide representation of characteristics that are broadly distributed in the population, such as credit card debt and home mortgages. Second, because income and wealth overall are distributed in a highly skewed way in the population (see Avery, Elliehausen, and Kennickell [1988] and Kennickell and Woodburn [1992a]), a simple random sample would yield too few wealthy households to draw any conclusions about the distribution of many financial variables. Third, because it is generally believed that nonresponse is positively correlated with wealth, estimates of many financial variables that do not account for this fact in the nonresponse adjustments will be biased.

The SCF deals with these problems by using a dual-frame design. One part, a standard area-probability design, is included to ensure adequate representation of broadly distributed characteristics. The other part of the design is a list sample drawn from administrative records maintained by the Statistics of Income Division of the Internal Revenue Service. This second sample enables both differential sampling of households that are more likely to be wealthy and systematic corrections for nonresponse based on extensive frame information. The 1989 SCF list sample accounts for nearly half of the net worth measured by the survey and nearly all of the observations of the top five percent of the wealth distribution

The remainder of this paper will address the adequacy of the current list design for the measurement of financial characteristics. The next section briefly outlines the list design. The sections that follow deal with some of the potential problems with the design, including sampling based on tax units rather than households, coverage problems, use of lagged rather than current information in the design, and high levels of nonresponse. The final section makes some proposals for the future development of the SCF sample. Because the 1992 survey is at a preliminary stage of processing, this paper mixes data from the 1989 and 1992 surveys. A later revision will replace all 1989 data with comparable 1992 data.

■ Sample Design

If sufficient frame information were available, the SCF list sample would use a design that minimized the expected sampling error for net worth or, perhaps, a weighted combination of financial variables, where the weights would reflect the relative importance of the individual variables at the analysis stage. In practice, no comprehensive data on wealth are available to make this calculation. However, extensive information on income from assets is collected as a part of Federal tax administration. Because many assets generate income that would be reported on a tax return, it is plausible that a proxy for net worth could be developed using income flows.

Under the terms of an interagency agreement, the SCF sample is designed using income data obtained from a stratified sample of individual tax returns (ITF) maintained by the Statistics of Income Division (SOI) of the Internal Revenue Service. Data in the file derive from tax returns filed one year before the survey is executed [1].

The income data for each record in the ITF are used to compute a "wealth index" which is intended to stand as an indicator of household wealth. The index is computed as the sum of capitalized income flows, where the rates of return used for each income flow are intended to reflect average annual returns for assets underlying each income type that prevailed in the tax year. For example, if one assumed an interest rate of 10 percent for the assets underlying interest income, a value of $100 in interest income would imply an asset worth $1,000. An amount that is constant within the ITF design strata is added to the index as a proxy for home equity [2].

List cases are selected in two stages. At the first stage, cases are selected that fall in the geographic sampling units drawn for the area-probability

sample. At the second stage, cases are sorted by the value of the wealth index, sampling strata are created, and the cases are randomly selected at disproportionately higher rates in strata corresponding to higher levels of the wealth index [3].

The tax filers who are selected are approached in two stages. As an added protection of the privacy of these individuals, all potential respondents are sent a mailing including letters explaining the purposes of the survey, and a postcard to be returned if the person does not wish to be contacted further. Interviewers then attempt to obtain interviews from all respondents who do not return the postcard.

■ Conceptual Problems with the Wealth Index

There are many reasons to think that the wealth index might not measure household wealth perfectly. Indeed, the wealth index used in the 1989 SCF has a simple correlation of only .34 with net worth or gross assets, though this measure is dominated by a number of large values [4].

The correlation of the log of the index with the log of net worth, a more robust measure, is .66 and that with the log of gross assets is .75. The Spearman rank correlation of the index with net worth is .76 and that with gross assets is .77. A broader indication of the relationship is given by Table 1, which shows the distribution of net worth measured in the 1989 SCF by wealth index stratum. The data show considerable dispersion of net worth in each stratum, but the general connection between the index and net worth is evident. The pattern for gross assets shown in Table 2 is very similar. Table 3, which shows the unweighted mean and median net worth and gross assets by stratum, demonstrates clearly that the central tendency of these distributions increases with the wealth index. The comparison of the ordering implied by the wealth index and that implied by survey values is complicated by a number of factors, including conceptual errors in the calculation of the index, changes in the definition of the unit over time, changes in asset values, reporting errors, imputation errors, variations in taxable income that may not be correlated with asset values, changes in tax definitions of income components, and other problems.

Omitted Assets

It is clear that not all assets generate income that would be recorded on an individual tax return. For example, it is only for homeowners who have mortgages and who itemize their deductions that we know anything about an individual's house [5]. Using survey data from the 1989 SCF list sample, Table 4 gives an indication of the amount of non-income-yielding assets [6].

For each sampling stratum, the index omits a considerable fraction of assets for at least the top quarter of the population. The median fraction of non-income-producing assets ranges between 19 and 69 percent over the wealth index strata. When the principal residence is deleted from the numerator in this calculation, as given in Table 5, the median share of other omitted assets ranges from 1 to 15 percent and the 90th percentile of this distribution ranges between 26 and 54 percent over the strata. Thus, even if the adjustment for home equity in the construction of the wealth index were perfect, there are still substantial amounts of assets not included in the index. These results suggest that significant numbers of units are classified in "too low" a stratum.

Omitted Debt

The calculation of the index makes no direct allowance for variations in indebtedness across filers. Other than the case for mortgages on a principal or secondary residence, there is generally no systematic trace of any other borrowing in the tax file data. Using final data from the 1989 SCF list sample, Table 6 shows the distribution of household leverage [7]. There is very broad variation in the degree of household leverage within each stratum, and leverage is generally decreasing with increasing levels of the wealth index.

Variable Rates of Return

The rates of return used to capitalize income flows for the wealth index are assumed to be fixed. This assumption is particularly flawed in the case of closely-held businesses or newly started businesses, where current income flows may be a poor indicator of the value of a business. Table 7 shows the percentile distribution of the ratio of income from personal businesses measured in the 1989 SCF to the value of the businesses. One might expect the upper tail of this distribution to be large and variable, but even the median rate of return varies from nothing in list stratum 1, to a high of 28 percent for stratum 3, to only 3 percent for highest stratum. For other assets, rates of return also are highly variable and, on average, will tend to vary directly with the riskiness of the assets. Thus, on average, we will tend to overestimate the value of risky assets and underestimate that of more conservative investments. Unfortunately, we have no information to make systematic corrections to individual observations to account for the variability of returns.

Intertemporal Income Variability

Because the index is based on only one year of income, the flows may differ from the "permanent" or longer-run income of the tax units. This may be a particular problem for filers with more complicated finances who may bunch certain types of income, either for purposes of reducing their taxes or because of the nature of some of the assets they own. We are able to match two years of the ITF for a part of the original 1992 list sample. Table 8 is a transition matrix for the wealth index computed for 1990 and 1991, using the same formula and adjusting the 1991 values for inflation using the Consumer Price Index for Urban Consumers [8].

For strata 2 through 5, the implied 1991 wealth index stratum for 1990 filers is roughly symmetrically distributed about their 1990 stratum. Consistent with other research (e.g., Williams and Sammartino [1993]), limited downward mobility at the top appears greater, particularly in the case of stratum 7.

In every stratum except the eighth one, which is not sampled in the SCF, substantial fractions of filers did not appear in both years of the file, either under the social security number of the primary filer or that of the secondary filer. Several factors may account for this less than perfect match. First, some people may either not file returns every year or they may request an extension and file later. Second, some filers may die, leave the country, or otherwise move out of the legal scope of U.S. taxes. Third, the ITF is not a panel in the usual sense. The ITF has a partially overlapping panel structure, which tends to minimize the variance of estimates based on that file over time and which also reduces problems of sample selection. An ITF sample case will not be retained if it is no longer available, the sampling rate in one of the ITF strata changes, or the case changes ITF strata [9].

It seems likely that filers who change ITF strata are also relatively likely to change their wealth index strata. Based on other information reported later in this paper, it appears that this third reason explains most of the nonmatches. Thus, in analyzing the change data from the ITF, we need to be clear that the changes we observe are for filers that are, in a sense, the most stable ones [10].

Efficiency Loss

Overall, the problems with the wealth index detailed in this section will lower the efficiency of the SCF list sample design stratified on the basis of the wealth index, but they do not (generally) induce bias. The loss of efficiency can be serious if significant numbers of cases are misclassified — particularly if high wealth cases are classified into low wealth index strata and, thus, have large weights.

■ Conceptual Differences Between Tax Filers and Households

There are some important conceptual differences between the area and list frames. The elements of the area-probability frame are dwelling units at the time of the survey. The list frame differs in several key ways. First, because the list sample is based on tax returns, any household that did not file at least

one tax return would not be eligible for sampling. Using the area-probability sample from the 1992 SCF, we estimate that about 15 percent of the respondents did not file (or expect to file) a return for tax year 1991[11].

In the case of the SCF, the omission of non-filers does not create any serious problems, since the area-probability sample provides adequate coverage of non-filers, and when the two samples are merged through weighting, the non-filers are treated separately.

Multiple Economic Units Within a Household

The list sample may over-represent certain types of households. There may be multiple economic units within a household; members of couples may file separate returns and, thus, have multiple chances of being selected [12].

In the area-probability sample, such households would be counted only once. Table 9 shows the percent of interviewed households in the 1989 list sample that contained multiple economic units that could have filed at least one additional return [13].

In the lowest stratum, almost 18 percent of households have at least one family unit beyond the primary economic unit, and in the highest stratum, the figure is only 4 percent. In principle, it is possible to make a multiplicity correction at the weighting stage if enough is known about the other family units. As a practical matter, this will be a significant problem only where the additional family units are much wealthier than the primary one. As shown in Table 10, except for stratum 1 in 1989, the secondary family net worth is a small fraction of total respondent net worth in each stratum. However, as indicated by the 90th percentile of the ratio of secondary family net worth to respondent net worth, some secondary units have substantial wealth relative to the primary units.

Multiple Tax Returns

Even in households without secondary units, there is still a possibility that the primary unit could file more than one return. At the time of the sample selection, an adjustment was made for taxpayers who filed separately. The probability of selection for married cases that filed separately was halved, which assumes implicitly that both the husband and wife would have fallen in the same stratum. Using the 1992 survey data on the list respondents, Table 11 shows the types of returns that respondents filed for the tax year after that used to design the sample. Separate returns are relatively frequent for all the strata. It is an interesting reflection of the variability of underlying income that over 2.5 percent of the units in the bottom two strata did not file a return at all in 1991.

The 1992 SCF data also suggest that the AGI of husbands and wives who file separately tend to be quite dissimilar. Indeed, the level correlation of the AGI of husbands and wives is -.01, and in logs the correlation is -.18. Table 12 uses survey data to show the median by stratum of the ratio of the husband's adjusted gross income (AGI) to the sum of AGI for the husband and the wife in married couples that filed separate returns for 1990. It appears that AGI for spouses may tend to be more equal at the bottom and more different at the top [14].

Given the incentives in the tax code, one would expect the returns of separate filers to be quite different. It is possible that an adjustment could be made at the weighting stage for these differences, though information has only become available for this purpose in the 1992 survey.

Changes in Household Composition

Changes in household composition also complicate the interpretation of the list sample. A tax filing unit may divorce or marry, or a family member may die. From interviews with the 1989 SCF list sample, we learn that 24 cases had married since 1987 (the tax year on which that sample was based) and 35 cases had divorced or separated. The list sample design will also miss some newly formed households, in cases where there are immigrants, or some cases where young people have filed their first tax return.

■ Geographic Coverage of the List Sample

Because of the number of calls needed to complete a typical interview, it is a practical necessity to have interviewers living close to the respondents [15]. Consequently, the first stage in drawing the list sample is the subsetting of the frame by the geographic areas that have previously been drawn as a part of the design of the area-probability sample. Aside from self-representing areas such as the New York SMSA, these areas are selected with probability proportional to the size of the population living in the areas. In using the same areas for the list and area-probability samples, it is implicitly assumed that people with substantial assets are geographically distributed in approximately the same way as the rest of the population.

We have used the 1990 ITF to examine the distribution of members of wealth strata 6 and above (wealth index over 10 million) at a level of detailed geography [16]. From the wide variation in the ratio of the estimated number of these filers to the total population by state, it is apparent that people with high predicted wealth are not distributed geographically in the same way as other households. After reweighting all of the ITF cases in the selected Primary Sampling Units (PSUs) to account for the area selection probabilities, the estimated number of cases in the selected PSUs understates the full ITF estimate by 18.5 percent.

The geographic distribution of high wealth index filers relative to the total population of tax filers is illustrated graphically on a map of the United States in Chart 1 [17]. Clearly these high wealth index cases concentrate in certain regions — like along the coasts — which suggests that a strata-specific PSU adjustment may be more appropriate, at least for the two highest wealth strata.

The SCF is not designed to be representative at the state level. However, it is possible to get a more detailed idea of the coverage of the list sample by comparing the number of potential respondents in the ITF for each state with the number predicted by the area selections [18].

For most states that are included in the sample, the weighted number of eligible respondents tracks the potential pool for the states reasonably well. A notable exception is Florida, where the omission of a relatively small number of retirement communities and other enclaves of wealthy people appears to cause the eligible pool of sample members to underestimate by almost half the number of filers in strata 6 and above for the state. Given the history of retirement settlement in Florida, it is not surprising that people tend to cluster more by economic similarity than is the case in other locations. There appear to be signs of similar coverage problems in parts of Connecticut, Texas, and a few other areas, as well.

Only about 5.8 percent of filers in stratum 6 and above are estimated to live in states that do not contain primary sampling units for the 1992 SCF [19]. Of this amount, more than half—3.2 percent—is estimated to be accounted for by a small number of locations in Louisiana, Nevada, and New Mexico.

■ Nonresponse

The achieved cross-section sample from all parts of the 1992 design includes 3,906 families, of which 1,450 come from the list frame. The area-probability cases were approached directly by interviewers and the response rate for these cases was about 69 percent. As noted earlier, the list-sample cases were given a prior opportunity to refuse participation by returning a postpaid card. As Table 13 indicates, about 23 percent of the original sample of list cases refused participation at this stage by returning the card [20].

The postcard refusal rate is fairly uniform across the strata, ranging only from 19 percent to 26 percent. The remaining list cases were approached by interviewers, yielding an overall interview rate for the list sample of about 31 percent [21].

Overall, response rates by wealth index stratum range from 43 percent for the lowest wealth index stratum to 14 percent for the highest stratum. The decrease in response rates by wealth index stratum

indicates that the ultimate probability of observation differs systematically from the selection probability.

Given the high level of unit nonresponse in the list sample, one might reasonably question the representativeness of the sample and its usefulness for estimation. It is important to note that this is not a problem unique to the SCF. While the noninterview rate for the list cases is high according to usual criteria, this figure merely makes explicit what is latent in other household surveys which lack the auxiliary information to identify the problem. Moreover, unlike most other surveys, in the SCF we have a significant amount of frame information to use in making adjustments to compensate for nonresponse [22].

To get an broader sense of the nature of differential response in the survey, Table 14 gives the unweighted mean and median of the ITF-based wealth index and financial income for different response categories in the 1992 SCF list sample [23].

The overall level of the figures is very high, reflecting the oversampling of cases with higher levels of the wealth index. In terms of medians, there is a clear rank-ordering with taxpayers who could not be contacted having the highest levels of the wealth index and financial income, followed by those who refused at the door, those who refused to participate by returning the postcard they were sent, and those who were interviewed. A similar result holds for the means.

This difference may be seen more clearly from a plot of the density of the wealth index and of financial income for each response category, shown in Chart 2 [24]. If no differential response were present, these density estimates would be identical, up to random error. What is important here is the differences in the densities of the response categories — the overall shape of the densities reflects the differential sampling. The plots reveal that the differences in means and medians of the wealth index and financial income are largely determined by the top of the distribution. The "no contact" response is relatively most dense at the top of both distributions, least dense for a large part of the middle, and relatively important at the bottom.

Some of these differences suggest behavioral and other distinctions. The pattern of no contacts may be explained by the fact that wealthier people are more likely to have gatekeepers, who make it difficult for interviewers to contact respondents, and that the poorest taxpayers may be more transient and, thus, more difficult to contact. The pattern of postcard refusals suggests that wealthier people are more likely to treat the advance letter as junk mail or, perhaps, more likely never to have received it, in the case where the mailing address is that of an accountant, a lawyer, or other tax preparer. Somewhat more detailed distinctions can be made using the models reported below, that control simultaneously for a number of observables in the ITF.

To obtain a more systematic understanding of the nonresponse mechanism(s) at work, we estimated a number of models using the range a variables available in the ITF. Three events are modeled as binary choices: whether the taxpayer was interviewed (interviewed=1), whether the taxpayer returned the postcard
declining participation (postcard not returned=1), and, given that a taxpayer did not return the postcard, whether the taxpayer was interviewed (interview=1) [25].

The models are based on a logit specification. Other distributional assumptions were examined (probit, cloglog) and found to generate essentially the same predicted probabilities of response. The exogenous variables consist of financial variables and a limited number of demographic variables available in tax data, supplemented by age data obtained by the IRS from the Social Security Administration [26].

All continuous financial variables were subjected to a log (base ten) transformation of the following form:

$$l(x) = \log(x + 1) \text{ if } x>0$$
$$-\log(-x +1) \text{ otherwise.}$$

A model search routine was used to reduce the dimension of the basic nonresponse model. Cubic splines were initially specified for continuous variables, starting with 6 knot points placed at

quantiles, and reduced based on the Cp measure of predictive accuracy (See Mallows (1973) and Hastie (1992)). In general, most the estimated functional forms were quantitatively similar through the specification search process. Binary variables were dropped if their contribution was sufficiently small, again based on the Cp criterion. To control for the effects of the sample design, each model contains dummy variables for the wealth index strata used in sampling. Models for postcard refusals and refusal conditional on not receiving a postcard refusal were fit to these chosen variables.

Table 15 shows the summary estimated significance criteria for models of three events: a complete interview, passive response at the postcard stage, and response given the postcard outcome. More detailed information of the explanatory variables is given in Chart 3, which plots the effects of a select set of splined variables — financial income, real estate taxes, salary, local taxes, charitable contributions, and age — on the probabilities of the three events. The heavy line in each cell represents the effect of the variable on the overall probability that a taxpayer gave a complete interview [27].

The line bisecting the dark gray band represents the effect of the variable on passive postcard response, and the dark gray band is a 95% (pointwise) confidence interval. Similarly, the remaining line and the light gray band surrounding it represent the effect on refusal conditional on not refusing by postcard and its 95% confidence interval, respectively.

The single most important factor in explaining nonresponse in this model is financial income. There is a significant decline in response with increasing financial income, beginning at about $1000; the effect is approximately flat below this amount. Similarly, nonresponse increases with real estate taxes paid above about $1000. One explanation is that higher valued homes are more likely to be in neighborhoods to which there is limited access. The propensity to respond increases with charitable contributions, suggesting that altruism may be an important determinant in the willingness to respond. Somewhat surprisingly, higher levels of nontaxable interest income are associated with higher levels of response. In addition, higher levels of pension income, large capital gains or losses, large losses from a business or farm, large amounts of rents, royalties, or trust income tend to increase response. Having either estate or partnership income increases the likelihood of response somewhat. Response is approximately flat in salary income until about the million dollar level, at which point response declines, although the confidence band at that point is wide. A similar effect is seen for local taxes. Social security income has a slight depressive effect on response. Age has very little effect until about age 65, at which point the likelihood of response decreases. While this age effect probably reflects a complex mixture of circumstances, it is very likely influenced by the greater likelihood of hospitalization and death at older ages. Relative to people living in the North Central region, people in the Northeast are less likely to respond, but people in the South and the West — except California — are more likely to respond. In examining the effects of living in popular retirement states, separate effects were explored for Florida and California. While Florida was not significantly different from the rest of the South, Californians were far less likely to respond than other Westerners. Folklore from the interviewers had it that the sample there had an unusually high incidence of communities in California with limited access. The design terms are fairly weak except for a self-representing PSU variable, which was strongly associated with nonresponse.

The model for response conditioned on response at the postcard stage is very similar to that for unconditional response. The differences appear to be only in minor details, such as design dummy variable for stratum two, which changes sign, but in neither case is significantly different from zero.

The model for passive postcard response is significantly different from those of both overall response and response conditioned on passive postcard response. Unlike the earlier cases, postcard refusal is increasing throughout the range of financial income, but flattens out at about the $10,000 level; above about $10,000 of salary income, response is increasing in salary; between about ages 30 and 65, postcard nonresponse is

increasing with age and then flattening out. In terms of the geographic variables, the postcard response model differs in the fact that taxpayers in the Northeast are actually more likely to respond. The higher level of nonresponse in California, seen in the earlier models, disappears, supporting the hypothesis that it is the difficulty in actually reaching respondents that raises the nonresponse rate in California.

■ Future Research

There are several areas in which improvements could be made in the SCF list sample. One obvious possibility is to explore the calibration of a wealth index model by merging wealth data from the SCF data with data from the ITF. However, this simple technical matter raises a number of complex ethical and procedural questions. We expect to discuss this question in detail with SOI, NORC, and others. We also welcome comments from readers.

Use of a single year of ITF data adds noise to the sampling process, as some filers have abnormally high or low incomes. One strategy for dealing with this problem would be to make use of longitudinal tax records. Such a record system is under construction at SOI (see Czajka and Schirm(1992)). This file has the added advantage of including data on families.

Alternatively, the SCF list sample might be drawn in two stages. Start with a file dated three years before the survey and draw a sample. Match the sample with the data from the following year and subsample based on changes in income components. There are several problems with this approach. First, the additional time between the filing of the return and the ultimate interview would exacerbate the potential for change in the unit of observation. Second, not all cases could be expected to be in two successive ITFs; retrieval of information from the IRS master files could be technically difficult and might raise additional administrative problems.

Geographic coverage problems in the list sample might be dealt with effectively by sampling independently from the ITF, subject to a cost function. At the least, some simple adjustments could be made to include key omitted areas. Additional information might also be collected on the survey to be used in making adjustments to the data for changes in the household composition and for multiple possibilities of entering the sample.

■ Acknowledgments

The authors would like to thank Barry Johnson, Fritz Scheuren, Bill Wong, and Louise Woodburn, at SOI, for their invaluable comments and help in assembling the data underlying this research. The authors would also like to thank Marty Frankel and Steve Heeringa, for their important work on the SCF sample design. The contents of this paper reflects the views of the authors only and not the opinion of the Board of Governors of the Federal Reserve System.

■ Endnotes

[1] For a description of the Statistics of Income Individual Program , see *Individual Income Tax Returns, 1987* [1990]. Statistical and research uses of SOI data are closely regulated to guarantee that individuals (and other entities) will remain protected against any disclosure of their financial and tax data [e.g., Wilson and Smith, 1983]. For the 1992 SCF, contractual agreements between the Federal Reserve Board, the National Opinion Research Center (NORC), and SOI clearly specify the limitations on the use of the administrative data in order to guarantee the privacy rights of the individual taxpayers. The list sample for the 1992 SCF was drawn from the set of tax returns filed during calendar year 1991, mostly for tax year 1990. This set of returns may include returns filed for years preceding 1990 and may also include multiple returns for a given filer.

[2] The amount is the mean home equity for the group estimated from an earlier survey and adjusted for overall price changes. The gross amount of capital gains was also added to the index. Other adjustments are detailed in Heeringa and Woodburn [1992]. Briefly, the basic wealth index is given as Home equity of ITF Stratum + ABS (taxable interest income)/.1165 + ABS(nontaxable interest income)/.067 + ABS(dividends)/.057 + ABS (rents and royalties)/.115 + (ABS(S-corp. income) + ABS(estate and trust income))/.230 + (ABS(Schedule C gross) + ABS (Schedule F gross profit) + ABS (other fam. income))/.172 + ABS (long-term capital gains) + ABS (short-term capital gains), where ABS denotes the absolute value function.

[3] See Heeringa and Woodburn [1992] for a description of the design of the 1989 SCF and Kennickell, McManus, and Woodburn [1993] for a description of the design of the 1992 SCF. The 1989 sampling strata in terms of values of the wealth index are: below 100,000, 100,001 to 500,000, 500,001 to 1,000,000, 1,000,001 to 2,500,000, 2,500,001 to 10,000,000, 10,000,001 to 250,000,000, and over 250,000,000. The 1992 strata divide the top end of the wealth index more finely — 10,000,001 to 100,000,000, 100,00,001 to 250,000,000, and over 250,000,000. The highest stratum was not sampled. The number of people in the highest stratum is very small and it is unlikely that a significant number of those people would be willing to participate in the survey, even if we were able to contact them.

[4] This calculation and all others reported here using the 1989 SCF are based on multiply-imputed data. The imputation procedure incorporates sampling stratum indicators as one of the explanatory variables. Gross assets includes bank accounts, money market and other mutual funds, corporate stock, bonds, businesses that are not publicly traded, real estate, pension accounts from which withdrawals could be made, and other assets. Net worth is defined as gross assets less all debts, including credit card and store accounts, closed-end consumer credit, mortgages, and other loans.

[5] As noted earlier, the wealth index assumes a fixed house value for each stratum, based on mean values derived from earlier surveys.

[6] Assets that do not yield an immediately taxable return include IRAs and Keogh accounts, employer-sponsored thrift plans (e.g., 401(k) plans), the principal residence, and real estate or business investments that yielded neither income nor losses.

[7] Leverage here is defined as the ratio of total debt to gross assets (as defined in footnote 5), including withdrawable amounts in employer-sponsored thrift plans.

[8] Woodburn [1991] has performed a similar exercise for the 1989 SCF using data from the 1987 and 1988 ITF.

[9] The ITF strata are defined in terms of levels and types of income reported.

[10] In the continuation of this research, we expect to have access to more complete information for the missing cases. Note that no adjustment is made to the data for variations in filing status, since it is not clear what sort of adjustment is appropriate for this purpose. Over the sampling strata, between 92 and 97 percent of filers chose the same filing status in both years.

[11] The corresponding estimate from the 1989 SCF for tax year 1988 was about 17 percent.

[12] As discussed further, below, at the time of sample selection, an adjustment is made for taxpayers who filed separately.

[13] Many of the additional units contain couples and others who might have filed an independent return.

[14] The number of cases in each stratum is too small to provide a sense of the distribution by stratum, though this ratio varies considerably over the entire list sample. Another indication of the variability of the ratio may be obtained from the area-probability sample cases. In the 1992 SCF, there were 138 area-probability cases that filed separate returns and, for the group, the maximum of the ratio was nearly 100 percent at the 90th percentile, 51 percent at the median, 24 percent at the 25th percentile, 3 percent at the 10th percentile, and the minimum was less that 1 percent.

[15] The median number of calls needed to complete a list case in 1989 was five. Ten percent of the cases required as many as 12 calls. Cases in the top three strata required on average about a third more calls than cases in the bottom three strata.

[16] The address given in the ITF may be that of a tax-preparer, rather than the taxpayer's address. In addition, to make this calculation properly, we would need to reweight the ITF at the state levels and we do not have sufficient information to make this calculation. However, most of the cases in wealth index stratum six and above are sampled at a very high rate in the ITF. The median sampling weight is 1 and the 95th percentile is 7.4. Thus, unless there are unobserved geographic sampling criteria for the IRS, sampling error in the ITF estimates for the geographic distribution of these filers should be relatively small.

[17] To construct this graph the ratio of strata 6 and 7 filers to total

population counts was calculated for each 3 digit zip code region. This ratio was assigned to each of the 5 digit zip code locations within the 3 digit zip code region. (We thank Paul Tukey for providing us with a public domain data set relating zip codes to geographic location. For a more detailed discussion of plotting this type of data see Carr and Pickle (1993).) These ratios were smoothed over latitude and longitude coordinates, using the nonparametric regression method, locally weighted regressions (LOESS), described in Cleveland, Grosse, and Shyu (1992), and the smoothed fit interpolated over the convex hull of the data.

[18] The weight for the selected sample is computed as the original ITF weight times the inverse of the probability of selection of the area.

[19] States not in the sample at all include Alaska, Arizona, Delaware, Hawaii, Idaho, Louisiana, Maine, Montana, North Dakota, Nebraska, New Mexico, Nevada, South Dakota, and Vermont.

[20] This figure is down dramatically from the 36 percent level of returns in 1989. However, the overall response rate is only slightly higher.

[21] The category "other non-interview" appears to be both a collection of miscellaneous reasons for nonresponse, and a category used when interviewers did not provide precise information. For this reason, we do no emphasize the distinction between nonresponse categories in our examination of nonresponse beyond the postcard stage.

[22] These adjustments are discussed in more detail in the context of weighting in Kennickell and Woodburn [1992a and 1992b] and Kennickell, McManus, and Woodburn [1993].

[23] Financial income is defined here to include taxable and nontaxable interest income and dividend income.

[24] The scales of the chart have been subject to the transformation, log10 (x+1). Consequently, a 4 on the scale is associated with about $10,000, 6 with about $1,000,000, and 8 with about $1,000,000,000. Densities are produced for each of the four response categories using an unweighted kernel density estimator with the same smoothing parameter. Note that as a result of the addition of a constant estimate of mean home equity by wealth stratum, described above in the review of the sample design, the wealth index has a non-zero floor.

[25] The first two contingencies were estimated using the entire SCF sample from the ITF and the third was estimated using only that part of the sample that did not return the postcard. The no contact and refusal at the door categories were collapsed for purposes of this estimation for two reasons. The degrees of freedom available are substantially reduced by treating the two categories separately. In addition, it is not clear that the incentives for interviewers were such that the distinction between the two would always be clear in the choice of disposition codes.

[26] Simple imputations were made for cases with missing age data.

[27] The value on the y-axis represents the variable's contribution to the index — which is monotonically linked to the probability of the event modeled. A lower (higher) level for the index

represents a lower (higher) probability of response.

■ Bibliography

Avery, R.B., G. Elliehausen, and A.B. Kennickell (1988) "Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finances," *Review of Income and Wealth*, pp. 339-369.

Carr, D.B. and L.W. Pickle (1993) "Topics in Scientific Visualization: Plot Production and Details," *Statistical Computing and Graphics Newsletter*, V.4 No.2 pp.16-20

Cleveland, W.S., E. Grosse and W.M. Shyu (1992) "Local Regression Models" in *Statistical Models in S*, Eds. J. Chambers and T. Hastie, Wadsworth and Brooks: Pacific Grove, CA.

Czajka, J.L., and A.L. Schirm (1992) "Selection and Maintenance of a Highly Stratified Panel Sample," *Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys*. Ottawa, Canada.

Hastie, T. (1992) "Generalized Additive Models," in *Statistical Models in S*, Eds. J. Chambers and T. Hastie, Wadsworth & Brooks: Pacific Grove, CA.

Hastie, T. and C. Loader (1993) "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, V. 8, No. 2, pp.120-143.

Heeringa, S. and L. Woodburn (1992) "Sample Design Documentation, 1989 Survey of Consumer Finances," mimeo, Institute for Social Research, University of Michigan.

Internal Revenue Service (1990) *Individual Income Tax Returns, 1987*, Government Printing Office, Pub. No. 1304.

Kennickell, A.B., and R.L. Woodburn (1992a) "Estimation of Household Net Worth Using Model-Based and Design-Based Weights: Evidence from the 1989 Survey of Consumer Finances," Federal Reserve Board, mimeo.

Kennickell, A.B., and R.L. Woodburn (1992b) "Methodological Issues in the Estimation of the Distribution of Household Net Worth: Results from the 1989 Survey of Consumer Finances," Federal Reserve Board, mimeo.

Kennickell, A.B., D.A. McManus and R.L. Woodburn (1993) "Methodological Issues in the Construction of Weights for the 1992 Survey of Consumer Finances," Federal Reserve Board, mimeo.

Mallows, C.L. (1973) "Some Comments on Cp," *Technometrics*, Vol. 15, 661-676.

Williams, R., and F. Sammartino (1993) "Income Mobility of the Rich: Some Preliminary Findings," mimeo.

Wilson, O., and Smith, W.J. Jr. (1983) "Access to Tax Records for Statistical Purposes," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 595-601.

Woodburn, L. (1991) "Using Auxiliary Information to Investigate Nonresponse Bias," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 278-283.

■

Chart 1: Geographic Distribution of High Strata Cases Relative to the General Population

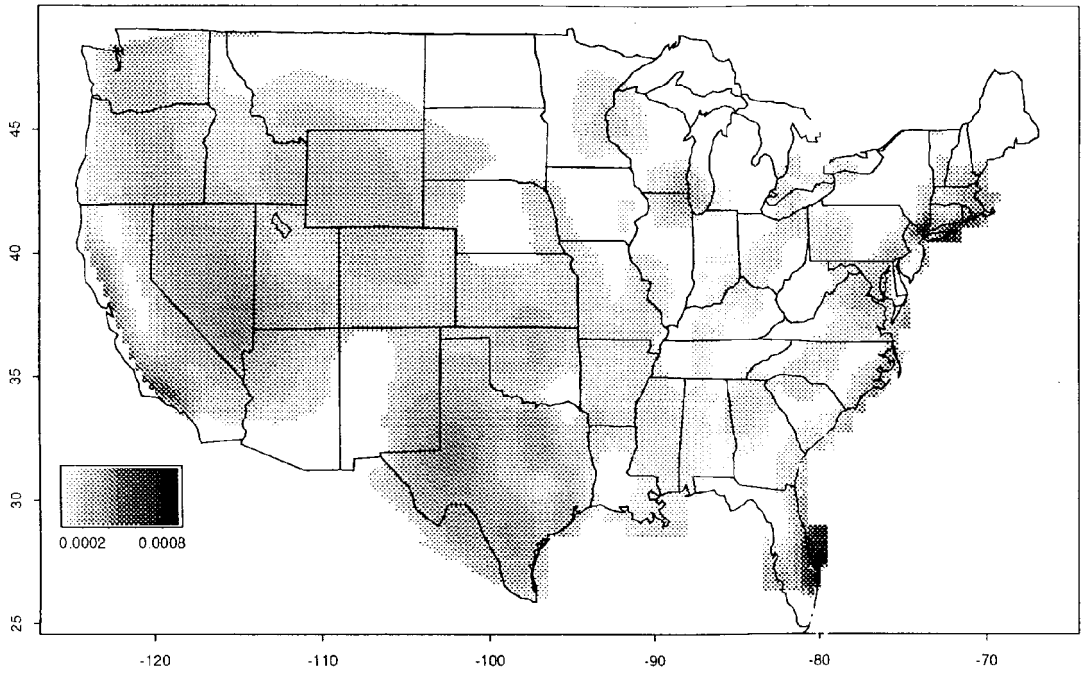
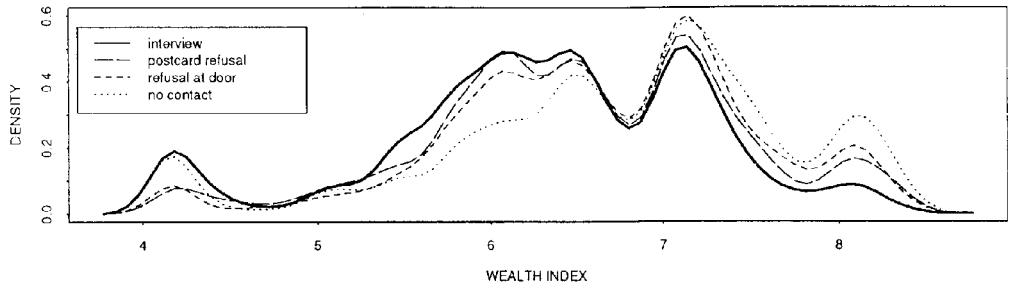


Chart 2: Density of Wealth Index and Financial Income



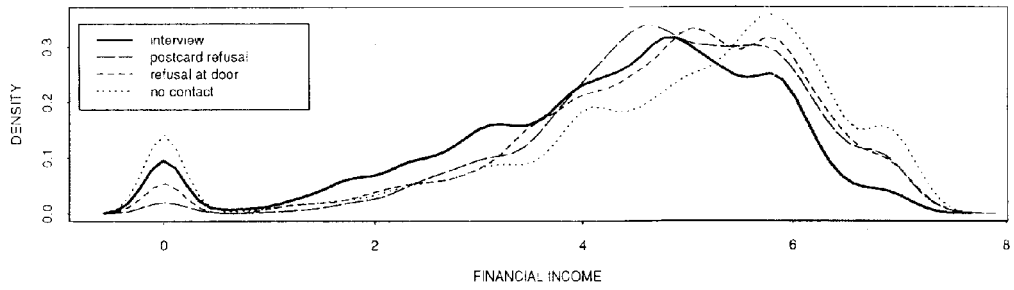
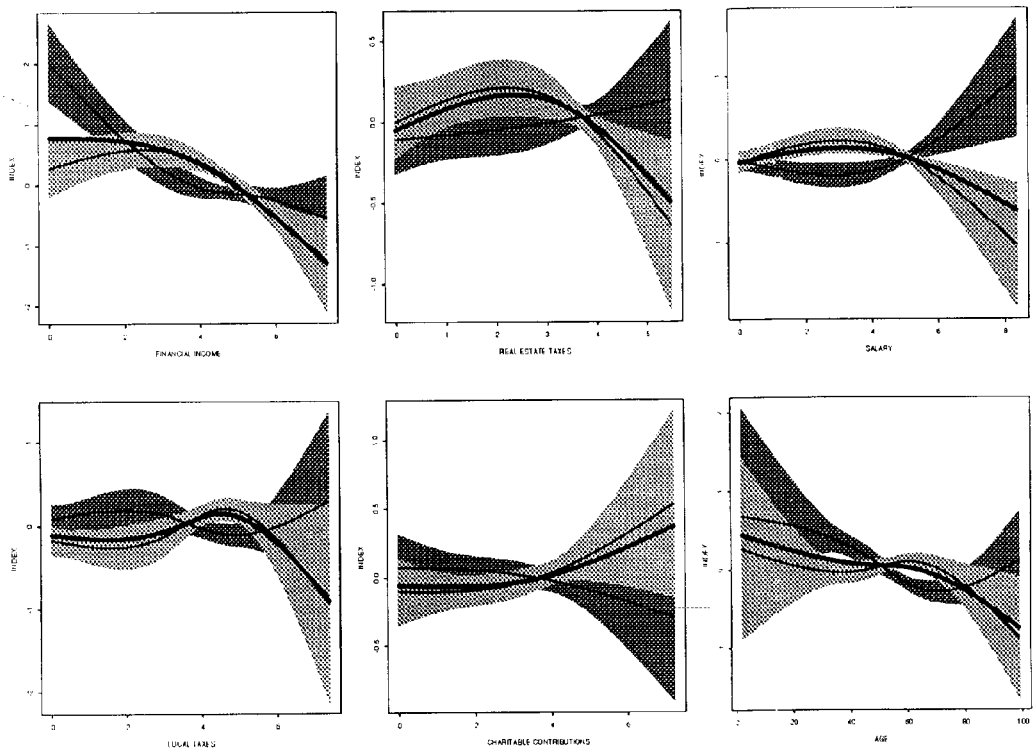


Chart 3: Effect of Select Variables on Response Probabilities.



The graphs in chart 3 show effects for (starting at the upper left-hand corner and moving in a clockwise direction) financial income, real estate taxes, salary income, age of primary filer, charitable contributions, and local taxes.

Table 1: Percent of Families in Wealth Index Strata Having Various Levels of Net Worth, 1989 SCF List Sample (Unweighted)

	Net worth						
Stratum	Under $100K	$100K-$500K	$500K-$1M	$1M-$2.5M	$2.5M-$10M	$10M-$25M	Above $25M
1	84.00	9.33	6.67	0.00	0.00	0.00	0.00
2	23.28	59.83	11.21	5.69	0.00	0.00	0.00
3	3.80	26.33	24.94	30.13	13.54	1.27	0.00
4	2.24	7.67	14.05	35.34	26.55	8.45	5.69
5	0.74	1.77	6.60	14.70	44.65	22.70	8.84
6	0.20	4.00	1.60	11.20	23.20	27.20	32.60

Table 2: Percent of Families in Wealth Index Strata Having Various Levels of Gross Assets, 1989 SCF List Sample (Unweighted)

	Gross assets						
Stratum	Under $100K	$100K-$500K	$500K-$1M	$1M-$2.5M	$2.5M-$10M	$10M-$25M	Above $25M
1	77.33	16.00	4.44	2.22	0.00	0.00	0.00
2	11.72	65.86	16.03	6.38	0.00	0.00	0.00
3	1.90	21.27	22.15	36.20	17.09	1.39	0.00
4	1.29	6.98	10.52	35.78	31.12	8.45	5.86
5	0.00	1.40	5.30	12.65	46.23	24.09	10.33
6	0.00	2.20	2.60	8.80	25.60	25.80	35.00

Table 3: Mean and Median Values of the Gross Assets and Net Worth, 1989 SCF List Sample (Unweighted), Thousands of Dollars

	Gross assets		*Net worth*	
Stratum	*Mean*	*Median*	*Mean*	*Median*
1	111	55	85	29
2	389	282	319	200
3	1799	1091	1558	877
4	6840	2238	6301	2064
5	11240	6551	9817	5870
6	30731	16238	26615	13752

Table 4: Percentile Distribution of Percent of Assets That Yield No Immediately Taxable Return, By Wealth Index Sample Stratum, 1989 SCF List Sample (Unweighted)

Stratum	*Percentile*				
	10th	*25th*	*50th*	*75th*	*90th*
1	0.0	0.9	59.8	84.8	94.9
2	24.0	39.2	68.6	86.0	94.2
3	16.1	27.6	49.2	71.7	90.4
4	7.8	20.5	36.8	59.5	77.8
5	4.6	12.7	26.8	48.3	65.2
6	3.0	6.2	19.3	38.1	60.1

Table 5: Percentile Distribution of Percent of Assets That Yield No Immediately Taxable Return, Excluding Principal Residence, By Wealth Index Sample Stratum, 1989 SCF List Sample (Unweighted)

Stratum	*Percentile*				
	10th	*25th*	*50th*	*75th*	*90th*
1	0.0	0.0	0.6	12.9	25.7
2	0.0	3.7	11.5	24.2	44.4
3	1.2	6.2	15.0	30.4	51.6
4	1.2	5.5	14.6	26.8	49.5
5	0.7	3.7	12.2	25.7	51.1
6	2.1	1.5	6.6	19.0	54.4

Table 6: Percentile Distribution of Total Debt as a Percent of Total Assets, By Wealth Index Sample Stratum, 1989 SCF List Sample (Weighted)

Stratum	*Percentile*				
	10th	*25th*	*50th*	*75th*	*90th*
1	0.0	3.5	29.2	68.6	102.2
2	0.0	2.6	17.6	38.9	56.2
3	0.0	1.4	10.6	26.2	37.5
4	0.0	0.0	6.6	17.4	32.5
5	0.0	0.1	4.0	13.3	29.1
6	0.0	0.0	1.1	11.1	29.7
All	0.0	1.4	13.1	34.7	58.9

Table 7: Percentile Distribution of Net Income from Personal Businesses as a Percent of the Value of the Businesses, 1989 SCF List Sample Business Owners (Weighted)

	Percentile		
Stratum	*25th*	*50th*	*75th*
1	0	0	12
2	0	6	8
3	4	28	197
4	1	19	95
5	1	9	24
6	0	3	25
All	1	13	91

Table 8: 1990 Wealth Index Strata by 1991 Wealth Index Strata, Full ITF Less Duplicates, Percent Distribution (Unweighted)

1990 wealth index stratum	*1991 wealth index stratum*								
	Missing	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	29.33	68.82	1.78	0.07	0.01	0.00	0.00	0.00	0.00
2	43.69	3.41	50.00	2.67	0.19	0.02	0.00	0.00	0.00
3	34.72	0.13	7.21	50.12	7.27	0.53	0.03	0.00	0.00
4	35.82	0.01	0.41	7.14	49.57	6.77	0.28	0.00	0.00
5	57.63	0.00	0.02	0.25	4.70	34.16	3.24	0.01	0.00
6	35.18	0.00	0.00	0.04	0.34	8.40	55.45	0.55	0.04
7	10.46	0.00	0.00	0.00	0.16	0.00	31.38	55.31	2.69
8	2.11	0.00	0.00	0.00	0.00	0.00	9.86	27.46	60.56

Table 9: Number of Economic Units Within Household, Percent of Households in Stratum, 1989 SCF List Sample (Unweighted)

	Number of economic units		
Stratum	*1*	*2*	*3+*
1	82.22	15.56	2.22
2	90.52	6.90	2.59
3	87.97	10.13	1.89
4	91.81	6.90	1.29
5	92.09	6.51	1.40
6	96.00	3.00	1.00

Table 10: Secondary Economic Units and Net Worth, 1989 SCF List Sample (unweighted)

Stratum	*% HH with secondary units w/ net worth^=0*	*Total secondary unit net worth as % total stratum net worth*	*Percentiles of ratio of secondary unit's net worth to primary unit's net worth*	
			50th	*90th*
1	17.78	22.0	>100.0	>100.0
2	9.49	0.0	1.0	16.6
3	12.02	0.6	0.4	42.7
4	12.19	0.1	0.1	15.9
5	7.91	0.1	0.1	8.1
6	4.00	0.0	0.1	0.6

Table 11: Types of Federal Income Tax Returns Filed in 1990 by Members of the 1992 SCF List Sample, by Wealth Index Stratum, Percent (Unweighted)

Stratum	*Type of return* *Non-filers*	*Single*	*Joint**	*Separate*
1	2.5	46.2	39.5	11.8
2	2.7	17.3	78.0	2.0
3	0.6	19.1	78.7	1.6
4	0.8	20.0	76.2	3.0
5	0.6	16.6	80.2	2.6
6	0.3	13.5	83.3	2.9
7	0.3	11.4	77.0	11.5

** Joint returns include cases where only the husband or wife filed.*

Table 12: Median of Ratio of Husband's AGI to Total AGI of Husband and Wife for Married Couples Filing Separately, By Stratum, 1992 SCF List Sample (Unweighted)

Stratum	1	2	3	4	5	6	7
Median	0.50	0.50	0.28	0.67	0.53	0.99	0.01

Table 13: Dispositions Codes, Number of Cases and Percent of Stratum, 1992 SCF List Sample, by Wealth Index Stratum (Unweighted)

	Wealth index stratum							*All*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*strata*
				Complete				
N	120	151	184	268	314	374	46	1457
% stratum	42.55	40.81	36.73	34.36	31.27	25.81	14.29	30.95
				Postcard refusal				
N	54	87	123	201	236	325	76	1102
% stratum	19.15	23.51	24.55	25.77	23.51	22.43	23.60	23.41
				R deceased				
N	1	5	7	4	3	9	2	31
% stratum	0.35	1.35	1.40	0.51	0.30	0.62	0.62	0.66
				Unlocatable				
N	33	11	14	18	37	71	21	205
% stratum	11.70	2.97	2.79	2.31	3.69	4.90	6.52	4.35
				Unavailable				
N	2	2	3	12	14	26	7	66
% stratum	0.71	0.54	0.60	1.54	1.39	1.79	2.17	1.40
				Out of scope				
N	0	0	3	2	0	3	0	8
% stratum	0.00	0.00	0.60	0.26	0.00	0.21	0.00	0.17
				Refusal by gatekeeper				
N	0	1	2	4	5	8	4	24
% stratum	0.00	0.27	0.40	0.51	0.50	0.55	1.24	0.51
				Language barrier				
N	1	2	4	6	2	0	1	16
% stratum	0.35	0.54	0.80	0.77	0.20	0.00	0.31	0.34
				Refusal, gov't involvement				
N	1	6	11	8	11	7	3	47
% stratum	0.35	1.62	2.20	1.03	1.10	0.48	0.93	1.00
				Other refusal				
N	51	75	123	205	272	433	97	1256
% stratum	18.09	20.27	24.55	26.28	27.09	29.88	30.12	26.68
				Other non-interview				
N	19	30	27	52	110	193	65	496
% stratum	6.74	8.11	5.39	6.67	10.96	13.32	20.19	10.54
				All dispositions				
N	282	370	501	780	1004	1449	322	4708
% stratum	5.99	7.86	10.64	16.57	21.33	30.78	6.48	100.00

Table 14: Mean and Median Wealth Index and Financial Income, By Response Type, List Sample, Unweighted

Response type	*Wealth index (thousands of $)*		*Financial income (thousands of $)*	
	Mean	*Median*	*Mean*	*Median*
Interview	10,902	2,528	390	32
Postcard refusal	17,834	3,364	799	73
Refusal at door	19,772	4,391	828	94
No contact	27,440	8,293	799	151

Table 15: Nonresponse Model Results

Variable	*df*	*Refusal*		*Postcard Refusal*		*Refusal Conditional*	
		Sign	*Incremental sum of sq*	*Sign*	*Incremental sum of sq*	*Sign*	*Incremental sum of sq*
Financial income	3	-	30.90	-	25.62	-	40.73
Nontaxable interest income	1	+	9.94	+	9.13	(+)	0.12
Pension income	1	+	7.32	+	11.25	(-)	0.65
Real estate income	2	-	5.03	-	6.27	(+)	1.21
Capital gains or losses	4	±	5.66	±	6.24	±	5.11
Rents, royalties, and estate income	3	±	6.48	(±)	3.73	±	9.30
Any partnership income	1	(+)	2.76	(+)	2.71	(+)	0.08
Any estate income	1	+	5.75	(+)	3.72	(+)	1.93
Salary	2	-	4.13	-	8.80	+	7.65
Local taxes	3	±	6.42	-	10.10	(±)	2.91
Wealth stratum 2	1	(+)	0.13	(-)	0.15	(+)	3.65
Wealth stratum 3	1	(-)	0.36	(-)	2.56	+	5.13
Wealth stratum 4	1	(-)	0.76	(-)	3.44	+	5.07
Wealth stratum 5	1	(-)	1.18	-	5.12	+	8.17
Wealth stratum 6	1	(-)	0.98	-	4.50	+	7.50
Wealth stratum 7	1	(-)	0.92	(-)	2.50	(+)	3.39
Business and farm income	3	(-)	2.58	-	4.19	(±)	1.98
Charitable contributions	2	(+)	1.61	(+)	3.15	(-)	1.29
Social Security income	1	(-)	0.56	(-)	0.64	(+)	0.17
Age	3	-	5.87	-	6.29	-	26.32
Western region	1	+	3.79	(-)	1.56	(+)	2.96
California	1	-	7.48	-	10.64	(+)	0.64
Northeastern region	1	(-)	1.41	-	6.86	+	10.56
New York	1	(-)	1.64	(-)	2.74	(+)	1.68
Southern region	1	+	6.78	(+)	2.28	+	8.46
Self-representing PSU	1	-	14.74	-	12.12	-	3.85
Sample size		*4708*		*4708*		*3606*	

The symbol "+" ("-") indicates increases in this variable increase (decrease) the probability of a response; "±" indicates that this variable has a mixed effect. A sign in parentheses indicates lack of statistical significance at the 5% level.