

Bias from Unit Non-Response in the Measurement of Income in Household Surveys*

C. Adam Bee, U.S. Census Bureau, Social, Economic, and Housing Statistics Division
Graton M. R. Gathright, U.S. Census Bureau, Center for Economic Studies
Bruce D. Meyer, University of Chicago and NBER

Draft: August 4, 2015

Abstract

Declining response rates to surveys is a widespread and troubling problem. Unit non-response (when a household is not interviewed at all) has been rising in most surveys. For example, unit non-response rates rose by 3-12 percentage points over the 1990s for six U.S. Census Bureau surveys (Atrostic et al. 2001). Many recent papers have raised the concern that this increased non-response has led to bias in key statistics. In light of this concern, we propose a new method to evaluate and correct bias from unit non-response. We apply this method to the Current Population Survey (CPS), the most used economic survey and the source of official employment, income, poverty, inequality and health insurance coverage information. Specifically, we use addresses to link the 2011 CPS Annual Social and Economic Supplement (CPS ASEC) to IRS Form 1040 records. This link allows us to compare several characteristics of respondents and non-respondents, including income and some of its components, self-employment status, marital status, number of children, and the receipt of social security. We find little evidence of differences between the percentiles of the income distribution of the linked respondents and non-respondents. We also find little difference between the respondent distribution, conventionally adjusted for non-response, and the combined respondent and non-respondent distributions. Significant differences between respondent and non-respondents are found for the number of children, and some other characteristics. We then compare our new method of assessing unit-nonresponse bias to prior methods.

JEL CODES: C81, D31, I32

KEY WORDS: Non-response, bias, income distribution, administrative data.

*This work was done at the Census Bureau Headquarters in Suitland, Maryland, and at secure research data centers by employees or those with special sworn status. The results have been through disclosure review to insure that no individual information is disclosed. Bee: U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, charles.adam.bee@census.gov; Gathright: U.S. Census Bureau, Northwest Census Research Data Center, University of Washington, Box 353412, Seattle, WA 98195-3412, graton.m.gathright@census.gov; Meyer: Harris School of Public Policy Studies, University of Chicago, 1155 E. 60th Street, Chicago, IL 60637, bdmeyer@uchicago.edu. The authors thank Tracy Mattingly for assistance obtaining address data, Mahdi Sundukchi for statistical review, Ed Welniak for comments and support, and Frank Limehouse for support at the Chicago Research Data Center.

I. Introduction

Large and nationally representative surveys are arguably among the most important innovations in social science research of the last century. Household surveys are the source of official rates of unemployment, poverty, health insurance coverage, inflation and other statistics that guide policy. They are also a primary source of data for economic research and are used to allocate government funds. Unfortunately, a decline in survey response rates is widespread and raises the possibility that our key data are no longer accurate. Unit non-response, which occurs when a household in a sampling frame is not interviewed at all, has been rising in most surveys. Unit non-response rates rose by 3 to 12 percentage points over the 1990s for six U.S. Census Bureau surveys (Atrostic et al. 2001). In non-Census surveys, the rise in unit non-response is also evident, and in some cases even sharper (Steeh et al. 2001; Curtin, Presser and Singer 2005; Battaglia et al. 2008; Brick and Williams 2013). The National Research Council (2013) report provides a thorough summary for U.S. surveys, but the pattern is apparent in surveys in other countries as well (de Leeuw and de Heer 2002). In light of this pattern, we propose a new method of using administrative records to evaluate and correct bias from unit non-response. We apply the method to recent Current Population Survey (CPS) data.

Indeed, the problem of rising unit non-response in major surveys and its potential to bias key statistics has been a heavily discussed topic in the survey research community. Unit non-response was the subject of two National Research Council reports and a special issue of a major journal (National Research Council 2011, 2013, Massey and Tourangeau 2013). The federal government, through its Office of Management and Budget (2006), has set a target response rate for federal censuses and surveys, and recommends analysis of non-response bias when the unit response rate is less than 80 percent. The editorial policy of at least one influential journal, the *Journal of the American Medical Association*, restricts publication of research using surveys with low response rates (Davern 2013).

In Figure 1, we report the unit non-response rate for five prominent household surveys during the 1984-2013 period: the Current Population Survey Annual Demographic File/Annual Social and Economic Supplement (CPS), the Survey of Income and Program Participation (SIPP), the Consumer Expenditure (CE) survey, the National Health Interview Survey (NHIS), and the General Social Survey (GSS). The surveys in Figure 1 show a pronounced increase in unit non-response over time, reaching rates in recent years that range from 11 to 33 percent.

Between 1997 and 2013 the unit non-response rate in the CPS rose from 7.2 to 10.7 percent while the rate in the NHIS rose from 8 to 24 percent. The National Research Council (2013) reports a general decline in response rates for a long list of surveys. The decline in response rates seems to be even more pronounced for public opinion surveys (Pew 2012).

However, the rate of unit non-response is not particularly informative about the accuracy of statistics from a survey. Unit non-response only leads to bias if it is non-random, with the exact requirement depending on the statistic in question and the weighting method. Evidence on the extent to which unit non-response leads to bias differs by survey and question. While there are examples of substantial bias, in other cases the resulting bias is small or can be mitigated by appropriate weighting, in which certain demographic variables in the survey are weighted to correspond to the total population (National Research Council 2013, p. 42-43). In their survey of bias estimates, Groves and Peytcheva (2008) found that bias magnitudes differed more across statistics (such as mean age or gender) within a survey than they did across surveys.

The standard method used now to assess unit non-response bias is some form of comparison of survey respondents to other survey respondents who were likely to be non-respondents in other circumstances. This procedure may suggest a direction of bias, but relies on the strong assumption that those reached after several contact attempts are identical to those never reached. Other recent research has used information on the characteristics of the location of respondents and non-respondents to examine any bias due to non-response.

A potentially more accurate approach is to match administrative data to the addresses of respondents and non-respondents and compare the characteristics available in the administrative data. As long as the matching to respondents and non-respondents is done in a parallel fashion, this approach seems likely to be more convincing.¹ We propose to directly examine potential non-response bias in Census Bureau surveys by linking both respondents and non-respondents to administrative data sources.

Specifically, we will link the entire 2011 CPS Annual Social and Economic Supplement (ASEC) frame to IRS Form 1040 for tax year 2010 using address information from both sources.

¹ A version of this type of linking, but only to part of the sample frame (the unit frame) was done on a small scale at the Census Bureau (Mah and Resnick 2009), by matching CPS data to Medicaid enrollment data from the MSIS file. This research linked to administrative data using the StARS administrative records system and Medicaid data from MSIS. The authors found little bias in Medicaid receipt statistics due to non-response.

We examine several ways of linking the two datasets. The resulting link allows us to compare several characteristics of respondents and non-respondents, including income, self-employment status, marital status, presence and number of children, and the receipt of pensions and certain government benefits. We focus on the overall or unconditional distribution of income in the CPS, with a secondary emphasis on the relatively few demographic variables available in the tax data.

The CPS is by far the most used household survey in the U.S. and is the source of official income and poverty statistics that are regularly reported in Census Bureau publications (see Census Bureau 2014a,b for some of the most recent examples). As well as examining bias due to non-response, this approach may provide a tool for survey design. In particular, it may provide the information to improve post-stratification weighting adjustments currently used to adjust base survey weights for differential non-response by geography and demographic characteristics.

Our results indicate little evidence of differences between the percentiles of the income distribution of the linked respondents and non-respondents. Not surprisingly then, we find little difference between the respondent distribution, conventionally adjusted for non-response, and the combined respondent and non-respondent distributions. There are significant differences between respondent and non-respondent households in the number of children, and some other characteristics. We then evaluate past methods of assessing unit-nonresponse bias in light of our results and find that in an important case they give misleading results.

In section II we describe the literature and in section III the data. In section IV we examine link rates and indications of the degree of randomness in any non-linking. We then in section V compare respondents and non-respondents along several characteristics, and compare estimates for weighted respondents to the entire population. In section VI, we compare our results to those obtained by earlier bias assessment methods that rely on late responders, attriters or zipcode level characteristics. Section VII concludes.

II. Past Work on Non-response and Non-response Bias

The literature on nonresponse bias has several strands. In their survey of bias estimates, Groves and Peytcheva (2008) found that bias magnitudes differed more across statistics, such as mean age or gender, within a survey than they did across surveys.

The standard method used now to assess unit nonresponse bias is some form of comparison of survey respondents to other survey respondents who were likely to be non-respondents in other circumstances. The likely non-respondents might be respondents who were only reached after many attempts, or they might be respondents who later left a panel survey. These methods are versions of “double sampling” or “two-phase sampling”. These methods are summarized in Chapter 4 of Groves (2004) and King et al. (2009) provides an application to Consumer Expenditure Survey nonresponse bias. Such procedures may suggest a direction of bias, but relies on the strong assumption that those reached after several contact attempts are identical to those never reached.

Other recent research has used information on the location of respondents and non-respondents to examine the likely bias due to nonresponse. For example, Sabelhaus et al. (2015) compare the zipcode level average income from Statistics of Income (SOI) data for respondents and non-respondents to the Consumer Expenditure Survey. They find that the nonresponse rate for the top quintile is 8 percentage points lower than the 26 percent rate found for the middle 90 percent of the zipcode-level AGI distribution. Consequently, they expect that the income of the top quintile is understated. A similar sized difference in reporting in the other direction is found for the bottom quintile where the response rate is higher than average.

Moving to studies of the CPS and those using linked administrative data, there are studies that have examined item nonresponse and imputation, but not unit-nonresponse. To study bias due to item non-response, several early papers linked tax data to CPS earnings data, including Herriot and Spiers (1975), Greenlees, Reece and Zieschang (1982) and David et al. (1986) who examined the bias in CPS due to item nonresponse for earnings and the resulting imputations. More recently, Bollinger et al. (2015) match the 2006-2011 ASEC to the DER. They find that item nonresponse to the earnings question is much higher in the tails, roughly 10 percentage points higher in the bottom 10 percent and 5 percentage points higher in the top 5 percent on a base nonresponse rate of about 20 percent. They also find that whole imputes have much lower

earnings through the bulk of the distribution, but particularly at the bottom, than do supplement respondents. The implications for income distribution estimates are not highlighted.

Other work by a subset of these authors (Hokayem, Bollinger and Ziliak forthcoming) and by Turek et al. (2012) examines the effect of earnings item nonresponse and imputation on the poverty rate. Both authors link item non-respondents to the DER, Hokayem et al. for reference years 1997-2008, and Turek et al. for 2005. Hokayem et al. find that item nonresponse leads the poverty rate to be understated by about 1 percentage point over their entire period, while Turek et al. find an understatement of 0.2 percentage points. The difference between these results seems to be how families with multiple earners and un-linked families are handled.

There is a very slim literature examining unit nonresponse bias in the CPS. Korinek et al. (2007) uses cross-state differences in response rates to estimate how the response rate varies with income. They then use these estimates to provide a corrected distribution of income from the survey. Their results indicate greater income inequality than is currently reported in the CPS.

III. Data

Our survey data come from the 2011 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC). The survey interviewed 75,178 households (those living at the same address) in-person or by telephone. The income reference year for the survey was calendar year 2010. The sample for the 2011 ASEC consisted of 96,944 households. Of the household interviews attempted, 81,725 were determined to be eligible and 75,178 completed the survey. Our focus is on the eligible households, those who were either interviewed, or who refused, were temporarily absent, or were unavailable for other reasons. Units that were vacant, demolished, or converted to uses other than residential are excluded from our analysis. In the parlance of the ASEC, we include type A non-responding units (non-responding eligible units), but we exclude non-responding units of types B (vacant units) and C (demolished units). We do not treat as non-respondents those who respond to the basic CPS survey but refuse to answer the supplement with the earnings question. These units are called “whole imputes” as their responses are imputed, while for non-respondents the weights are adjusted to account for nonresponse.

The CPS sample is comprised of units from four different frames: the unit frame, the area frame, the permit frame, and the group quarters frame. We obtained residence-level address information for all of these units that we use in linking to the administrative data as described later. We also employ indicators for rural vs. urban residence and for four regions of the country. Since the CPS sampling is based on residential addresses, the addresses are likely to be of high quality.

Our administrative data source is the universe of IRS Form 1040s filed during the 2010 calendar year.² We do not have all items that appear on the Form 1040; rather we have access to a subset of variables that includes adjusted gross income (AGI), number of children, filing status, receipt of social security, and indicators for filing various schedules, including schedule A, C, D, and E. Since tax units are not the same as households, we consider two strategies below to aggregate 1040s to the household level.

We also make use of tabulations from the IRS Statistics of Income program of mean AGI at the zipcode-level. These publicly-available data served as the key measure of income for respondents and non-respondents in the Sabelhaus et al. (2015) evaluation of bias from unit non-response in the Consumer Expenditure Survey. We employ these zipcode-level data both in examining the identifying assumptions for our comparisons between respondents and non-respondents and also to investigate how our method compares to previous methods.

IV. Linking

Theoretical Issues

The question of when a test using administrative data linked to survey data is useful in assessing non-response bias can be described in the following terms. We would like to know under what conditions a null of equality of respondent and non-respondent distributions we would like to test is implied by a null of equality of the respondent and non-respondent distributions we are able to test. We would also like to know under what conditions a test on this

² Surprisingly, the 1040 records do not indicate the tax year of the filing, so further data work is warranted to determine, if possible, which records represent late filings from previous years. Similarly, the file pools both original and corrected filings, where applicable.

second pair of distributions has power to reject the null when the null for the original distributions is violated.

To be more precise, let Y_i^S be a survey report of a variable for unit i . In our case Y_i^S is CPS measured income. Unfortunately, we do not observe Y_i^S for all i because not all units respond. Let $D_i = 1$ when i is a respondent, and 0 when i is a non-respondent unit. We would like to be able to test the null hypothesis that the distribution in the respondent population ($Y_i^S | D_i = 1$) is the same as that in the non-respondent population ($Y_i^S | D_i = 0$). The first goal of linking to administrative data is to find a situation where a null hypothesis of no difference in distributions is implied by this original null hypothesis.

Suppose we observe Y_i^a , administratively reported income. For simplicity, let us suppose that $Y_i^a = Y_i^a(Y_i^S)$, that administratively reported income is a function of reported income. This assumption would imply that equality of the distributions in the survey would imply equality in the administrative data. It is probably more plausible to assume that both $Y_i^S = Y_i^S(Y_i)$ and $Y_i^a = Y_i^a(Y_i)$, where Y_i is true income, but this complication is not crucial to the key issues.

We also assume that, fortunately, administratively reported income Y_i^a is available for ASEC respondents and non-respondents, but only for a subpopulation of each, those that we can link to the survey frame. What conditions on these subpopulations, i.e. the linking process, are required for the distributions of the subpopulations of respondents and non-respondents to be equal under the null of the distributions being the same in the full populations of respondents and non-respondents? Precisely, we would need that the subsample selection indicator L_i which equals one when i is included in the subsample, satisfy the following condition:

If $(Y_i^a | D_i = 1)$ equals $(Y_i^a | D_i = 0)$ then
 $(Y_i^a | D_i = 1, L_i = 1)$ equals $(Y_i^a | D_i = 0, L_i = 1)$.³

This condition requires that selection into the subsample depend on something besides D_i . This condition allows the subset of respondents for whom $L_i = 1$, to be a non-random subset, but it must be non-random in a way that it selects observations from respondents and non-respondents in a similar fashion. For example, if $L_i = 1$ drops all values of $Y_i^a < k$, such a

³ Alternatively, suppose that whether the distribution of Y_i^a is the same for respondents and non-respondents is of interest in itself. We may believe that Y_i^a suffers from less error conditional on a report, for example, so is of more interest than reported income. Then, the same condition would be applicable.

restriction would satisfy the condition. This condition also makes clear that administrative data on a different variable, say an indicator for receipt of a transfer program, T_i , that is only received by a subset of low-income units, could also satisfy this condition as long as L_i is not associated with D_i . It should also be noted that L_i can be related to both Y_i^s and Y_i^a and still satisfy this condition.

Besides unbiasedness, we would also like a test using sample values of a variable from the subpopulation distributions of respondents and non-respondents to have power against violations of our original null hypothesis. The test will have higher power when the administrative variable is more highly correlated with the survey variable. It will also have higher power when the subsamples are larger, i.e. the link rate between the administrative and survey data is higher. The power against certain violations of the null will also depend on the relationship between the survey and administrative variables more generally. Suppose that the administrative data is T_i above, an indicator for receipt of a transfer program to low-income units. In this case, the equality of the administrative data distributions (just the proportion receiving the transfer) would have power against differences between respondents and non-respondents at the bottom of the income distribution, but be uninformative to differences at the top.

Practical Considerations

We were able to link the ASEC units to tax data in two alternative ways. First, we used IDs from the Master Address File (MAF). The MAF is the Census Bureau's official inventory of all known living quarters and selected nonresidential units in the United States. The file includes address information, geographic location codes, other attribute information about each living quarter. The Census Bureau continually updates the MAF using the United States Postal Service Delivery Sequence File and various automated, computer assisted, clerical, and field operations. MAFIDs were attached to both the ASEC records and the IRS 1040 records based on by the Census Bureau Center for Administrative Records Research and Applications (CARRA) using the Census Bureau's MAFMatch process which includes standardizing and parsing of address information for each record and performing probabilistic linking that is calibrated to be very conservative (minimizes false positive matches to the MAF at the risk of lower match rates).

These MAFIDs were then employed as linking keys to perform a many-to-many merge between the two files.

Our second method of linking survey units and tax records was to merge the files together directly using the address information on each source. The address information from each source was cleaned and parsed using the SAS DQ module and then a many-to-many merge on these address fields was performed between the files. We draw the same conclusions regardless of the linking method. The appendix presents the results from the direct-linking of ASEC units to the 1040s.

Linked Data Characteristics

We first provide evidence that comparing respondents and non-respondents using linked 1040s is a sensible strategy, as link rates are quite similar for respondents and non-respondents, both overall and conditioning on various observable characteristics. The results suggest that the linking process selects a subsample of respondents and non-respondents in a parallel fashion that does not depend on whether a unit is a respondent or not.

Table 1 reports the share of responding and non-responding units whose addresses were assigned a MAFID, i.e. those that can be linked to the MAF. The link rate for responding and non-responding units is the same, 94 percent.^{4,5} We find no differences in link rates for respondents and non-respondents within geographic and frame subgroups.

Using zipcode-level AGI for ASEC units, we examine the relationship between AGI and assignment of MAFID for both responding and non-responding ASEC units. This provides evidence that the relationship is not very different between respondents and non-respondents. Figure 2 presents rates of MAFID assignment, separately for respondents and non-respondents, for each vingtile of zipcode-level AGI.⁶ The assignment of MAFIDs to ASEC units across the income distribution does not appear to be related to response to the ASEC.

⁴ These percentages are base-weighted to reflect the probability of selection of each unit due to the sample design.

⁵ The fact that ASEC units do not all find a match in the Census Bureau's Master Address File is a puzzle that merits further investigation. The unit frame part of the CPS sample is actually drawn from an earlier version of the MAF, so the non-matching of any of these units is surprising.

⁶ We note that that the relationship between zipcode-level AGI and unit-level differs between respondents and non-respondents. We discuss this finding further in Section VI.

The rate at which MAFIDs can be assigned to the addresses on IRS Form 1040 records is 89 percent, somewhat lower than the link rate to the ASEC.⁷ For the 1040s, there are more explanations for non-linking to the MAF. A 1040 can legitimately be filed with a post office box instead of a residential address.⁸ A rural-route address is another type that might be difficult to link to the MAF. Further, residents of apartments where the name and the address of the apartment building are sufficient for mail delivery might omit the apartment number. Typographical errors or illegible handwriting could also be factors.⁹

Since some individuals do not appear on any IRS Form 1040s, AGI may be unobservable for some ASEC units. Approximately 10-12 percent of individuals do not appear on any 1040, either as a filer, spouse, or dependent (Mortenson et al. 2009, Heim et al. 2014). The share of tax units that do not file is even higher, at around 17 percent. The share of CPS households that have no members that file a 1040 could be higher or lower than either of these numbers. On the one hand, for a household to have a 1040 only one member of the household needs to file. On the other hand, the non-filing households may disproportionately be single individuals who make up a small share of all individuals, but a large share of households. Thinking of the CPS household filing percentage as a ratio of filers/population, other factors that could matter but are likely to be less important are the likely inclusion of foreigners and part-year residents in the numerator but not the denominator (which would mean that the true non-filing percentage would have to be higher), and the differential inclusion of illegal immigrants in the numerator and the denominator (effect uncertain).

Table 2 reports the rate at which CPS units are linked to at least one 1040. Overall, we link at least one 1040 to 79 percent of respondent addresses and 76 percent of non-respondent addresses. Given that we are starting with 94 percent of each type of CPS units having MAFIDs, the additional rates of non-linking due to the 1040 step is 15 percentage points for respondents and 18 percentage points for non-respondents. Since these percentages are close to the rates of non-filing indicated earlier (10-12 percent at the individual level, 17 percent at the tax unit level), it seems likely that the main incremental source of non-linking in this case is non-filing. The 3

⁷ This rate is for records of filings of IRS Form 1040 from the 50 US states and the District of Columbia, excluding military mail.

⁸ The 2014 instructions for Form 1040 are that a PO Box is to be used only if the post office does not deliver mail to the residence.

⁹ A single MAFID was assigned by CARRA to each 1040. There were multiple 1040s for some MAFIDs as we discuss later.

percentage point difference in linking between respondents and non-respondents is low, but the difference is statistically significant, and the null hypothesis that respondents and non-respondents are the same would suggest that the number should be zero. The only significant difference by region is for those from the West. In terms of the components of the sample frame the only noticeable differences are for the permit and unit frames. Differences are statistically significant for units in both rural and urban areas.

As we did to examine the linking of respondents and non-respondents to the MAF, we can use the zipcode-level AGI to examine how linking of CPS units to 1040s varies with income, for both respondents and non-respondents. Figure 3 displays the rate of linking of ASEC units to 1040s, separately for respondents and non-respondents, by vingtile of zipcode-level AGI. The linkage rates for respondents and non-respondents track each other fairly closely through the 15th vingtile and then diverge somewhat beginning at that point in the distribution. The largest difference in link rates between respondents and non-respondents is under 8 percentage points and occurs for the 20th vingtile. The similarity in the link rate gradients for respondents and non-respondents suggests that comparing linked tax records between the two groups will be informative about the differences in income distributions between them.

Having argued for the validity of our comparisons of respondents and non-repondents based on linked tax records, we also want to describe how observability of unit level AGI for CPS units is related to income in order to document the alternatives over which our tests of equality between the income distributions for respondents and non-respondents have power. Table 3 examines the selection process that determines which 1040s are assigned a MAFID and which are not. The unlinked 1040s include a disproportionate share of extreme values of AGI which means that our tests may have slightly lower power against small differences between respondents and non-respondents in the tails of the income distribution. Table 3 also reports that unlinked 1040s are slightly less likely to list a spouse, have slightly fewer dependents, and have slightly different likelihood for some income sources and schedules. Assignment of MAFIDs to 1040s appears to be well-distributed among sub-groups, supporting comparisons of respondents and non-respondents for all of these subgroups when linked to ASEC units.

Figure 4 depicts the relationship between AGI and assignment of MAFIDS among 1040s. Rates of MAFID assignment are fairly high and constant across most of the distribution of AGI in the universe of 1040s. The four highest and four lowest percentile groups do show lower match

rates. This pattern confirms that non-assignment of MAFIDs to 1040s does not preclude useful comparison between ASEC respondents and non-respondents.

Since the CPS ASEC data include measures of income, we can compare reported total income of households--linked and not linked--in the respondent sample. Table 4 shows the sample distribution of survey-reported household income for linked and unlinked respondents. The distribution for unlinked respondents has lower income percentiles than the linked respondents. This difference is not surprising as we expect individuals who do not file a tax return to be disproportionately those with incomes sufficiently low that filing was not a requirement for them. In line with this expectation, those respondents whose addresses could not be directly linked, a sample in which the share of non-filers is almost certainly smaller than in the sample of respondents not linked to a 1040 via the MAF, has a distribution that is less shifted to the left, i.e. has higher values that are more like those of the linked respondents. If anything, it is a surprise that the sample distribution of income of unlinked respondents is not shifted further to the left. This pattern suggests that non-filers are less concentrated at very low income than we might have expected. Mortenson et al. (2009) find that in 2003 the median income of non-filers was about \$10,000.

In Table 5, we compare survey-reported demographic characteristics between linked and unlinked CPS respondent units. The comparison indicates that characteristics of the linked units also differ somewhat from the characteristics of the unlinked units. These differences do not threaten the validity of our comparisons of linked data for respondents and non-respondents but simply delineate certain subgroups and portions of the income distribution for which non-linking means our comparisons will be based on a slightly smaller sample.

When multiple 1040s link to a given ASEC units, we use the average across the linked 1040sas our unit-level measure of AGI and other characteristics. Average income is justified if the modal case of multiple 1040s occurs because one household moves out and another moves in within the year and both use the same address on their 1040. The use of the average would also be appropriate when we have multiple years of 1040s for a given household.

A second option we consider is summing the AGI for all of the 1040s linked to a CPS unit with the idea that filers at a given address are multiple tax units within the same household. For example, a household might consist of a married couple filing separate returns, a couple and an elderly parent, a couple and an adult child, or a couple and a working teenager. In each of

these cases, two returns might be filed. Summing is also appropriate in the situation where multiple 1040s are filed from an address because a housing unit is occupied by several unrelated individuals such as roommates, or by several families.

V. Main Results

Our main results directly compare the distribution of income and other characteristics from linked 1040s for respondents and non-respondents. Comparing the unconditional distributions of unit 1040-reported adjusted gross income (AGI) between responding and non-responding units (base-weighted) provides very little evidence that responding units are drawn from a different distribution than non-responding units. We do find evidence that responding units differ from non-responding units on the proportion of units with a married filer, on the number of dependents, and on the receipt of income from certain sources.

In Table 6, we present estimates of AGI and other characteristics based on linked 1040s for respondents and non-respondents using base-weights that have been adjusted by us for non-linking of ASEC units to 1040s. *p*-values presented are for tests of equality of the given estimate between respondents and non-respondents and are based on standard errors calculated using replicate weights for the units.

We find no statistically significant differences between respondents and non-respondents in the mean or percentiles of the distribution of AGI. The estimated mean AGI for respondents is \$61,868 and the estimate for non-respondents is \$63,546. The estimated percentiles of AGI differ by no more than \$1050 for the 1st, 5th, 10th, 25th, 50th, 75th, and 90th percentiles. At the 95th and 99th percentiles, the differences in the estimates are larger, but represent a small fraction of those larger income estimates.

We find that responding units are 5.8 percentage points more likely to file a 1040 as married (i.e. a spouse was listed on the return) and that the mean count of exemptions for children at home is higher by 0.084 (a little less than one-tenth of a dependent) for responding units. Responding units are less likely to have zero dependents and more likely to have 2, 3, or 4 dependents. Responding units are less likely to have wage and salary income, more likely to have dividends, and more likely to have income from social security. Responding units are less likely to itemize deductions, and more likely to have capital gains/losses, and more likely to have

a profit or loss from farming. The differences in non-income characteristics seem consistent with expectations about responding units being disproportionately available at home: retired or caring for dependents.

In Table 7, we present the distribution of AGI from linked 1040s for respondents only, weighted using published final survey weights that include conventional adjustments for unit-nonresponse. We compare this distribution to that of the base-weighted combined samples of respondents and non-respondents. Since the overall level of non-response is low it is not surprising that these distributions differ little. We find statistically significant differences for only three of the presented income percentiles, and the difference between estimates is less than \$400 in each case. For demographic characteristics and income sources we find statistically significant but small differences between the two sets of estimates.

In Figure 5, we pool the linked respondents and non-respondents, and then plot the response rate by vingtile of this pooled distribution. The figure suggests that there is no relationship between income and non-response. The non-response rate averages about 8 percent and varies across the individual vingtiles from about 9 percent to 6.5 percent with no visually apparent tendency at the top or bottom of the distribution for the rates to be higher or lower.

VI. Comparisons to Earlier Response Bias Estimation Methods

A natural way to examine non-response bias when linking to individual information on non-respondents is not possible is to use some of the information available on the sample frame. One sensible approach that has been taken is to use zipcode information, in particular zipcode-level average income (Sabelhaus et al. 2015). A natural expectation would be that zipcode-level income analysis of non-response would provide the same pattern as individual level income, but likely in muted form since zipcode-level income is correlated with unit income, but imperfectly. We examine this same approach in our case, comparing the pattern of non-response by percentiles of zipcode-level income to the non-response pattern by percentiles of unit level income for the same units that we are able to link to 1040s. We find a different pattern of response rates by vingtile when we perform this substitution of zipcode-level mean AGI for unit-level AGI.

In Figure 6 we depict the response rate by vingtile of zipcode-level mean AGI. The response rate does not systematically vary with income until around the 15th vingtile when it appears to drift down slightly. These differences are not pronounced; the non-response rate is about 8 percent through most of the distribution, drifting down to about 10 percent at the 20th vingtile. In terms of percentage changes in nonresponse with income, these changes are comparable to those that Sabelhaus et al. find for the Consumer Expenditure Survey. We find smaller percentage point differences in the CPS which has a lower overall non-response rate.

This zipcode-level AGI analysis though differs from what we previously saw with the unit level AGI analysis. We are exploring the reasons for this difference further, but logically it would seem that response rates are especially low for low income residents of high income zipcodes. Our findings suggest that the use of geographically-aggregated data for evaluating bias from unit non-response warrants some caution .

VII. Conclusions

Our results from linking AGI information from IRS Form 1040s by address to respondent and non-respondent units from the 2011 CPS-ASEC do not suggest bias in the unconditional distribution of income. This result should reduce one concern about the accuracy of measured income, poverty and inequality, some of the principal statistics obtained from the CPS. One caveat is that we find some indication that mean demographic characteristics such as marital status and number of children differ across respondents and non-respondents. Thus, it is possible that distributions for subgroups of the population might differ between the respondent sample and full population. We hope to address this question as we extend our study.

We also find that non-respondents are more likely to be in the lower portion of their zipcode income distribution (at least for high-income zipcodes), a finding that indicates a complication in the use of geographically-aggregated information as a proxy for individual-level data in evaluating and remediating bias from unit non-response.

References

- Abowd, John and Martha Stinson. 2013. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data." *Review of Economic Statistics*, December, (95)5: 1451-1467.
- Atrostic, B. K., Nancy Bates, Geraldine Burt, and Adrian Silberstein. 2001. "Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights." *Journal of Official Statistics*, 17:209-226.
- Battaglia, Michael P., Mina Khare, Martin R. Frankel, Mary Cay Murray, Paul Buckley, and Saralyn Peritz. 2008. "Response rates: How have they changed and where are they headed?" In *Advances in Telephone Survey Methodology*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japiec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, 529–60. New York, NY: Wiley.
- Behaghel, et al. 2012. "Please Call Again: Correcting Non-Response Bias in Treatment Effect Models." IZA working paper.
- Bollinger and Hirsch. 2013. "Is Earnings Nonresponse Ignorable?" *Review of Economic Statistics*, May, (95)2: 407-416.
- Bollinger and Hirsch 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching," *Journal of Labor Economics*, (24)3.
- _____. 2007. "How Well are Earnings Measured in the Current Population Survey? Bias from Nonresponse and Proxy Respondents," Working Paper.
- Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2015. "Trouble in the Tails? Earnings Non-Response and Response Bias across the Distribution." Working Paper, University of Kentucky.
- Brick and Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys," *Annals of the American Academy of Political and Social Science*, 645, January.
- Citro, Constance F. 2014. "From Multiple Modes for Surveys to Multiple Data Sources for Estimates." National Research Council working paper. Presented the 2014 International Methodology Symposium of Statistics Canada, Ottawa, Canada.
- Curtin, Richard, Stanley Presser, and Elinor Singer. 2005. Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly* 69(1): 87–98.
- Davern, Michael. 2013. Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research. *Health Services Research* 48(3): 905-912.
- David, Martin, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest. "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association* 81 (March 1986): 29-41.

- De Leeuw, Edith, and Wim de Heer. 2002. "Trends in Households Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, ed. by Groves, et al.
- DiNardo, McCrary, and Sanbonmatsu. 2008. "Constructive Proposals for Dealing with Attrition: An Empirical Example." Unpublished working paper.
- Greenlees, John, William Reece, and Kimberly Zieschang. "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association* 77 (June 1982): 251-261.
- Groves, Robert M. and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly* 72: 167-189.
- Groves, Robert M. 2004. *Survey Errors and Survey Costs*. Hoboken, NJ: John Wiley & Sons.
- _____. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly* 70: 646-675.
- Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York, NY: John Wiley.
- Groves, Robert M. and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly* 72: 167-189.
- Heim, Bradley T., Ithai Z. Lurie and James Pearce. 2014. "Who Pays Taxes? A Dynamic Perspective." *National Tax Journal*, December 2014, 67(4): 755-778.
- Herriot, R. A. and E. F. Spiers. "Measuring the Impact on Income Statistics of Reporting Differences between the Current Population Survey and Administrative Sources," *Proceedings, American Statistical Association Social Statistics Section* (1975): 147-158.
- Hokayem, Charles, Christopher Bollinger, and James P. Ziliak. forthcoming. "The Role of CPS Nonresponse in the Measurement of Poverty." *Journal of the American Statistical Association*, forthcoming.
- King, Susan L., Borian Chopova, Jennifer Edgar, Jeffrey M. Gonzalez, Dave E. McGrath and Lucilla Tan. 2009. "Assessing Nonresponse Bias in the Consumer Expenditure Interview Survey." *JSM, Section on Survey Research Methods*.
- Mah, Ming-Yi and Dean Resnick. 2009. "Preliminary Analysis of Medicaid Enrollment Status in the Current Population Survey." U.S. Census Bureau.
- Massey, Douglas S. and Roger Tourangeau, editors. 2013. "The Nonresponse Challenge to Surveys and Statistics." *The ANNALS of the American Academy of Political and Social Science*. 645(1): 6-236.

- Mortenson, Jacob A., James Cilke, Michael Udell, and Jonathon Zytneck. 2009. "Attaching the Left Tail: A New Profile of Income for Persons Who do not Appear on Federal Income Tax Returns." *National Tax Association Proceedings*.
- National Research Council. (2011). "The Future of Federal Household Surveys: Summary of a Workshop." K. Marton and J.C. Karberg, rapporteurs. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- _____. (2013). "Nonresponse in Social Science Surveys: A Research Agenda." Roger Tourangeau and Thomas J. Plewes, Editors. Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Office of Management and Budget. 2006. "Standards and Guidelines for Statistical Surveys." September.
- _____. 2014. "Guidance for Providing and Using Administrative Data for Statistical Purposes," Memorandum for the Heads of Executive Departments and Agencies, M-14-06, February.
- Pew Research Center. 2012. "Assessing the Representativeness of Public Opinion Surveys" Washington, D.C.
- Peytchev, Andrew. 2013. "Consequences of Survey Nonresponse". *Annals of the American Academy of Political and Social Science*. 645(1): 88-111.
- Ravallion, Martin, Anton Korinek, and Johan Mistiaen. "An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys" *Journal of Econometrics* 136(2006): 213-235.
- Sabelhaus, John, David Johnson, Stephen Ash, David Swanson, Thesis Garner, John Greenlees and Steve Henderson. 2015. "Is the Consumer Expenditure Survey Representative by Income?" In *Improving the Measurement of Consumer Expenditures*. University of Chicago Press.
- Steeh, Charlotte, Nicole Kirgis, Brian Cannon, and Jeff DeWitt. 2001. "Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century." *Journal of Official Statistics* 17(2): 227-47.
- Turek, Joan, Kendall Swenson, Bula Ghose, Friz Scheuren and Daniel Lee. 2012. "How Good Are ASEC Earnings Data? A Comparison to SSA Detailed Earning Records." Paper presented at the Federal Committee on Statistical Methodology (FCSM).
- U.S. Census Bureau. Various years-a. "Current Population Survey: Annual Social and Economic (ASEC) Survey Codebook," Washington D.C., United States Department of Commerce. Bureau of the Census.

- U.S. Census Bureau. Various years-a. “Survey of Income and Program Participation”, Washington D.C., United States Department of Commerce, Bureau of the Census.
- U.S. Census Bureau. 2012. “Census Bureau Releases Estimates of Undercount and Overcount in the 2010 Census.” Washington D.C., United States Department of Commerce. Bureau of the Census. http://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html.
- U.S. Census Bureau. 2014a. “Income and Poverty in the United States: 2013.” P60-249, September 2014.
- U.S. Census Bureau. 2014b. “The Supplemental Poverty Measure: 2013.” P60-251, October 2014.
- Wooldridge, Jeffrey M. 2007. “Inverse Probability Weighted Estimation for General Missing Data Problems.” *Journal of Econometrics*, 141, 1281–1301.

Table 1: Proportion of CPS units that link to the Master Address File

	Respondent	Type A Non-Respondent	p-value
Overall	.940	.941	.827
Region			
Northeast	.910	.919	.738
Midwest	.953	.950	.909
South	.942	.942	.989
West	.947	.952	.649
Urban	.952	.947	.320
Rural	.889	.906	.150
Frame			
Area	.892	.908	.615
Group Quarters	.807	n/a	n/a
Permit	.897	.875	.255
Unit	.952	.955	.410
Observations	75,178	6,547	

Note: Rates of linking to the MAF are base-weighted: statistics are the base-weighted proportions of 2011 CPS ASEC units with the specified characteristics that are probabilistically linked to the Master Address File by the Census Bureau's MAFMatch linkage algorithm using reference files available as of September 2013. By construction, no Group Quarters units are considered Type-A non-respondents. Tests of equality employ replicate weights to account for sample design.

Table 2: Proportion of CPS Respondent and Non-respondent units that link Tax Records

	Respondents	Non-Respondents	p-value
Overall	.786	.761	.003
Region			
Northeast	.790	.774	.286
Midwest	.805	.797	.613
South	.772	.755	.141
West	.787	.732	<.001
Urban	.793	.767	.005
Rural	.757	.726	.038
Frame			
Area	.695	.677	.335
Group Quarters	.771	n/a	n/a
Permit	.854	.820	.030
Unit	.787	.758	.011
Observations	75,178	6,547	

Note: The proportions are the share of ASEC units that link to at least one Form 1040 using the MAF to link the two sources. Statistics are the base-weighted proportions of 2011 CPS ASEC units with the specified characteristics that are linked to at least one Tax Year 2010 IRS Form 1040 record. By construction, no Group Quarters units are considered Type-A non-respondents. Tests of equality employ replicate weights to account for sample design.

Table 3: Characteristics of Tax Records that Link to the Master Address File

<i>Form 1040 variable</i>	1040s that do not link to the MAF	1040s that Link to the MAF	All 1040s
Mean AGI	\$ 84,859	\$ 53,439	\$ 56,794
Percentiles of AGI			
1	\$ -46,299	\$ -177	\$ -1,400
5	2	2,718	2,422
10	3,475	5,995	5,749
25	12,193	14,526	14,269
50	29,250	32,390	32,049
75	65,117	66,817	66,662
90	131,637	112,499	113,897
95	220,679	153,367	157,959
99	836,339	320,034	360,505
Married	.361	.391	.388
Mean count of exemptions for children at home	.531	.599	.592
Number of dependents			
0	.661	.625	.630
1	.153	.165	.164
2	.117	.133	.132
3	.048	.053	.053
4	.021	.022	.022
Receipt of income sources			
Wages, salary	.750	.833	.825
Interest	.340	.392	.392
Dividends	.217	.195	.198
Rent and royalties	.128	.069	.075
Social Security	.183	.177	.178
Schedules filed			
A: Itemized deductions	.341	.334	.335
C: Profit/loss from business	.228	.151	.159
D: Capital gains, losses	.246	.208	.212
E: Supplemental income, loss	.240	.118	.130
F: Profit or loss from farming	.021	.013	.014
SE: Self-employment tax	.183	.114	.122
Number of records	14,750,764	123,390,820	138,141,584

Note: The universe of “All 1040s” includes only the 50 states and the District of Columbia, excluding 1040s filed from foreign countries, island areas/territories, and military addresses. “MAF-Linked 1040s” refers to the set of Form 1040 records that are assigned a MAF ID by the Census Bureau’s MAFMatch probabilistic linking algorithm.

Table 4: CPS Characteristics of CPS-Tax Record Linked Households: Income

<i>ASEC-reported household income</i>	MAF-Linking			Direct-Linking			
	(1) Linked Respondents	p-value	(2) Unlinked Respondents	(3) Linked Respondents	p-value	(4) Unlinked Respondents	(5) All Respondents
Mean	\$ 74,573	<.001	\$ 42,341	\$ 75,203	<.001	\$ 53,253	\$ 67,371
Percentiles							
1	\$ 0	n/a	\$ 0	\$ 0	n/a	\$ 0	\$ 0
5	9,605	<.001	2,157	9,456	<.001	4,436	7,524
10	15,500	<.001	7,280	15,000	<.001	8,837	11,842
25	30,000	<.001	13,157	30,000	<.001	17,000	24,001
50	56,080	<.001	26,000	56,500	<.001	35,914	49,100
75	96,020	<.001	53,288	97,187	<.001	69,744	88,000
90	147,904	<.001	94,208	150,000	<.001	114,832	138,757
95	191,680	<.001	126,899	194,964	<.001	151,335	180,484
99	338,100	<.001	239,067	345,343	<.001	271,429	321,332
Observations	58,910		16,268	48,380		26,798	75,178

Note: Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. All statistics in this table are weighted using final ASEC household weights. Observations are at the household level. “Linked” respondents are respondent units that we successfully linked to at least one Form 1040 record. Statistics for “All Respondents” differ from published statistics because the published income statistics are produced using a linearized income distribution with bin widths of \$2,500, and above tables include group quarters units whereas the universe for published statistics consists only of household units. Respondent count (75,178) differs from published unit count (75,188) because 10 respondent units were lost due to miscoded case serial numbers.

Table 5: CPS Characteristics of CPS-Tax Record Linked Households: Demographics

<i>Proportions of ASEC respondents, by ASEC-reported characteristic of householder</i>	(1) Linked Respondents	p-value (1)=(5)	(2) Unlinked Respondents	(5) All Respondents
Age				
Under 65 years	.804	<.001	.721	.785
15 to 24 years	.044	<.001	.079	.052
25 to 34 years	.163	.546	.161	.162
35 to 44 years	.188	<.001	.148	.179
45 to 54 years	.217	<.001	.171	.207
55 to 64 years	.193	<.001	.161	.186
65 years and older	.196	<.001	.279	.215
Education (age ≥ 25)				
Less than HS	.093	<.001	.178	.112
HS grad	.267	<.001	.305	.276
Some college	.270	<.001	.236	.262
College grad	.205	<.001	.130	.188
Post-grad	.121	<.001	.070	.110
Race and Hispanic origin				
White, not Hispanic	.717	<.001	.657	.703
Black, not Hispanic	.113	<.001	.160	.123
Asian, not Hispanic	.049	<.001	.041	.047
Other, not Hispanic	.005	.003	.009	.006
Hispanic (any race)	.117	<.001	.133	.120
Marital status				
Married	.556	<.001	.334	.506
Widowed	.084	<.001	.152	.100
Divorced	.144	<.001	.186	.154
Separated	.025	<.001	.051	.031
Never married	.190	<.001	.277	.210
Nativity				
Native born	.861	.495	.859	.861
Foreign born				
Naturalized citizen	.075	<.001	.060	.071
Not a citizen	.064	<.001	.081	.068
Observations	58,910		16,268	75,178

Note: All statistics in this table are weighted using final ASEC household weights. Observations are at the household level. “Linked” respondents are respondent units that we successfully linked to at least one Form 1040 record. Education universe for MAF-linking consists of 56,521 linked units and 15,084 unlinked units, while 3,573 units are out-of-universe due to a householder aged under 25. Education universe for exact-linking consists of 46,490 linked units and 25,115 unlinked units, while 3,575 units are out-of-universe due to a householder aged under 25. Test for equality employ standard errors calculated using published replicate weights.

Table 6: Characteristics of CPS Respondents and Non-Respondents as Recorded in Tax Records, 2011 CPS ASEC

1040 Characteristic	Respondents	Non-Respondents	p-value
AGI			
Mean	\$ 61,868	\$ 63,546	.510
Percentiles			
1	\$ 0	\$ -1,050	.580
5	6,810	6,952	.777
10	11,179	11,747	.145
25	21,074	21,771	.110
50	39,854	40,887	.277
75	73,983	73,595	.802
90	120,806	120,416	.909
95	167,344	172,467	.475
99	362,223	397,074	.439
Married filing jointly	.464	.406	<.001
Mean count of exemptions for children at home	.787	.703	<.001
Number of dependents			
0	.539	.573	<.001
1	.188	.189	.852
2	.174	.157	.009
3	.071	.059	.003
4	.028	.022	.025
Receipt of income sources			
Wages, salary	.838	.863	<.001
Interest	.417	.409	.408
Dividends	.207	.189	.014
Rent and royalties	.080	.077	.437
Social Security	.177	.131	<.001
Schedules filed			
A: Itemized deductions	.385	.406	.012
C: Profit/loss from business	.173	.178	.435
D: Capital gains, losses	.221	.204	.013
E: Supplemental income, loss	.137	.134	.633
F: Profit or loss from farming	.014	.009	.003
SE: Self-employment tax	.132	.135	.619
Observations	53,559	4,500	

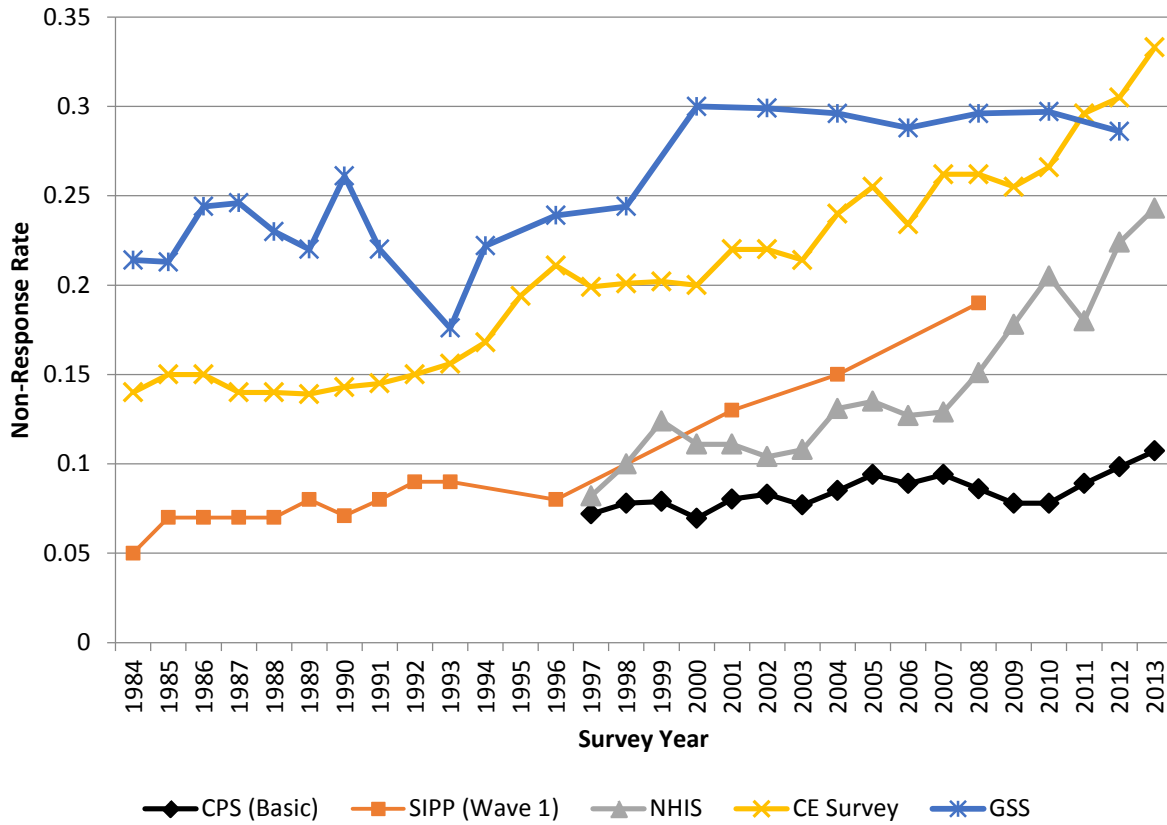
Note: Mean tax record characteristics are weighted by CPS base-weights adjusted for linkage probability. Statistical testing accounts for non-independent sampling of units due to complex survey design, using replicate weights reflecting the ASEC sample design. The tabulated unit-level statistic is the mean of Adjusted Gross Income values across all Form 1040s linking to each unit. Means and percentiles are rounded to the nearest dollar. The MAF-linked units are selected such that MAFIDs are unique; within each group of units with duplicate MAFID values, one unit is chosen at random, with equal probability. Reweighting for non-linkage is carried out by probits run separately for respondents and non-respondents, including frame, state/urban, presence of PO box or rural route, and type of living quarters (house/apartment, mobile home, or other). Listed *p*-values are generated from tests of equality of estimates in the designated pairs of columns. The tests are based on standard errors calculated using replicate weights to account for the sample design.

Table 7: Comparing TY2010 IRS Form 1040 Adjusted Gross Income between Linked Respondents and Linked Eligible Units of the 2011 CPS ASEC, after reweighting for non-linkage

<i>1040 Characteristic</i>	Final-Weighted Respondents	Base-Weighted Combined	p-value
AGI			
Mean	\$ 62,167	\$ 62,001	.704
Percentiles			
1	\$ 0	\$ 0	1.000
5	6,451	6,835	<.001
10	10,924	11,216	<.001
25	21,122	21,128	.934
50	40,393	39,923	<.001
75	74,142	73,936	.256
90	120,359	120,732	.295
95	167,311	167,577	.714
99	362,056	365,635	.485
Married filing jointly	.455	.459	<.001
Mean count of exemptions for children at home	.643	.780	<.001
Number of dependents			
0	.612	.541	<.001
1	.168	.188	<.001
2	.142	.173	<.001
3	.057	.070	<.001
4	.022	.027	<.001
Receipt of income sources			
Wages, salary	.817	.840	<.001
Interest	.447	.416	<.001
Dividends	.229	.205	<.001
Rent and royalties	.084	.080	<.001
Social Security	.211	.174	<.001
Schedules filed			
A: Itemized deductions	.389	.387	.121
C: Profit/loss from business	.168	.173	<.001
D: Capital gains, losses	.243	.220	<.001
E: Supplemental income, loss	.144	.137	<.001
F: Profit or loss from farming	.016	.013	<.001
SE: Self-employment tax	.128	.132	<.001
Observations	53,559	58,059	

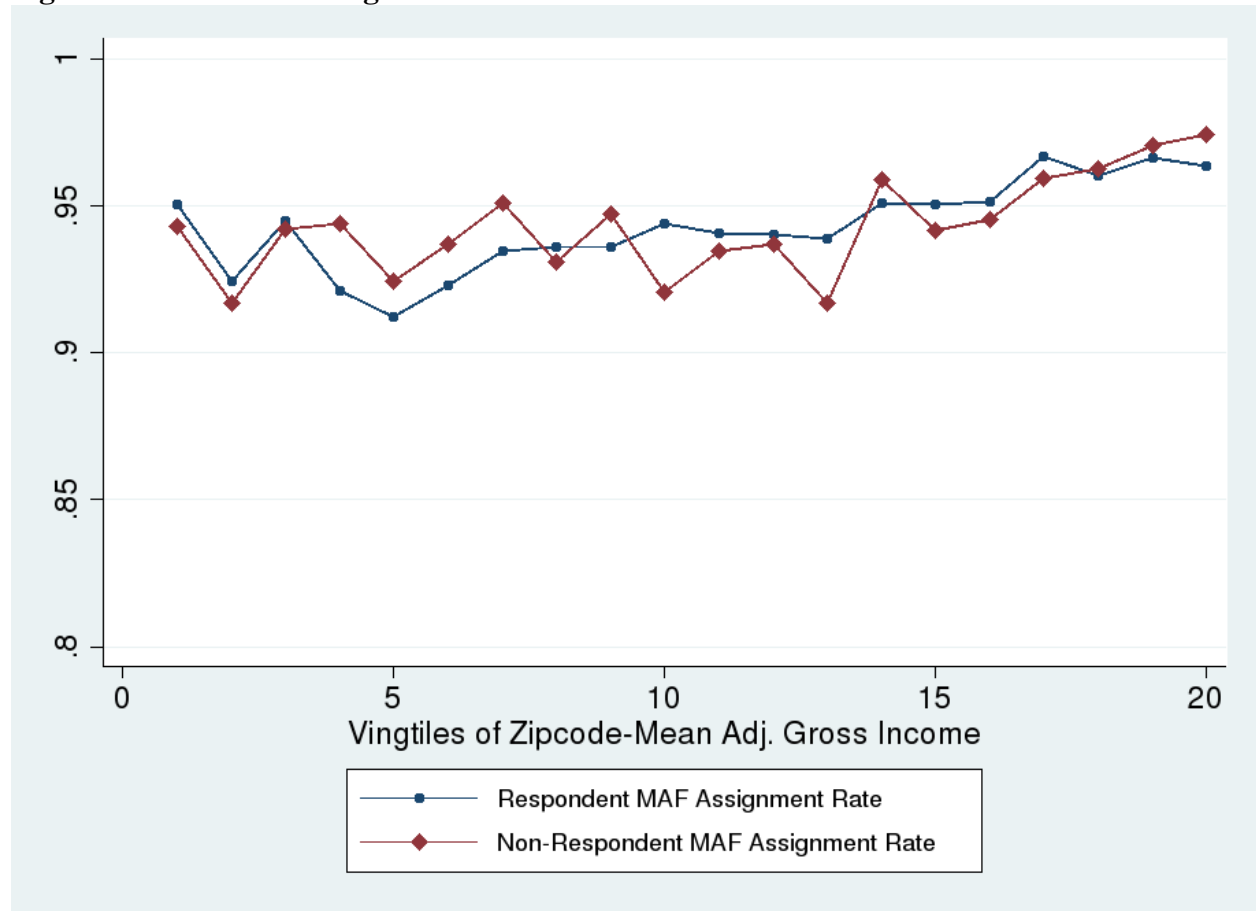
Note: Listed *p*-values are generated from tests of equality of estimates in the designated pairs of columns. Statistical testing accounts for non-independent sampling of units due to complex survey design, using replicate weights reflecting the ASEC sample design. Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. The tabulated unit-level statistic is the mean of Adjusted Gross Income values across all Form 1040s linking to each unit. Means and percentiles are rounded to the nearest dollar. The MAF-linked units are selected such that MAFIDs are unique; within each group of units with duplicate MAFID values, one unit is chosen at random, with equal probability. Reweighting for non-linkage is carried out by probits run separately for respondents and non-respondents, including frame, state/urban, presence of PO box or rural route, and type of living quarters (house/apartment, mobile home, or other).

Figure 1: Unit Non-Response Rates of Major Household Surveys



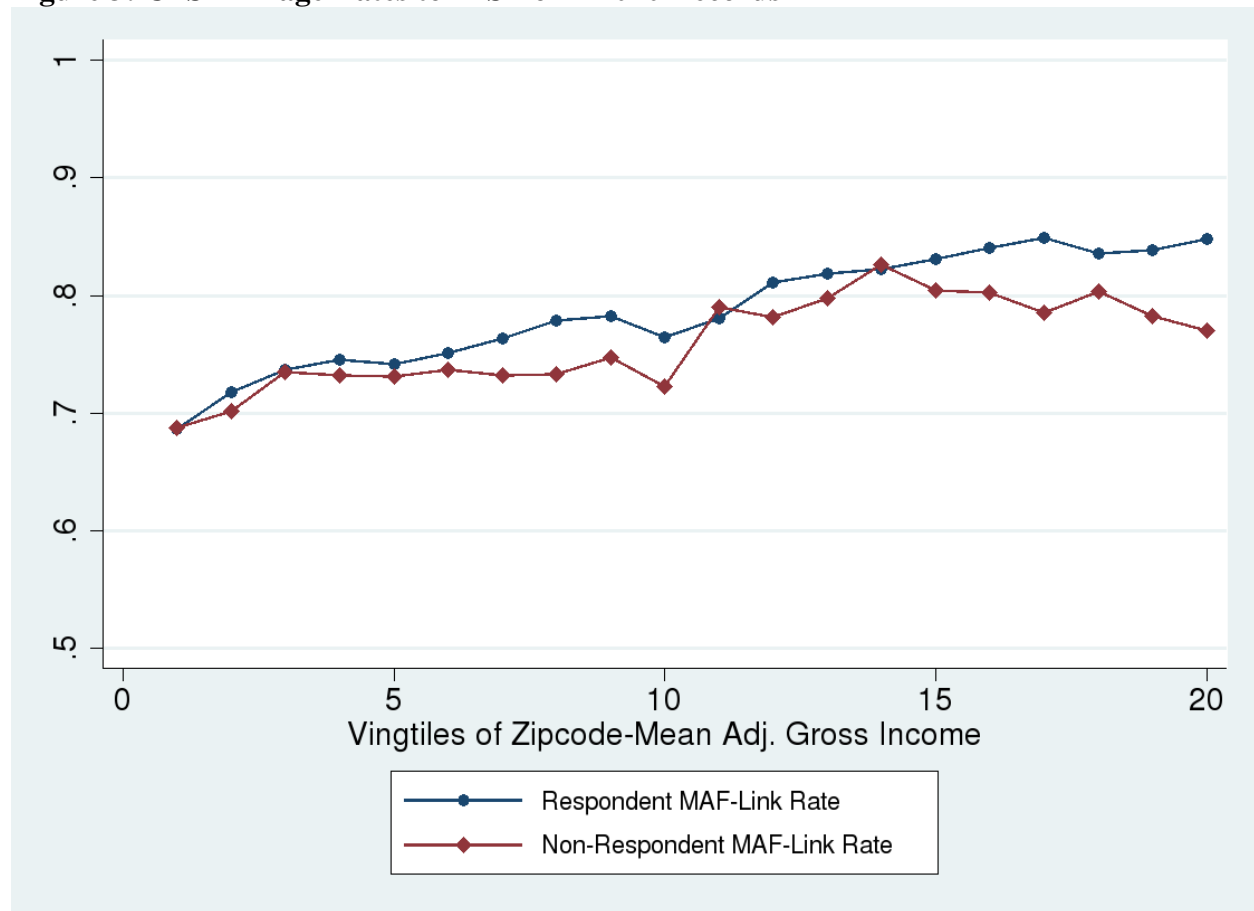
Sources: For CPS, see Appendix G of U.S. Census Bureau (Various years-a). For SIPP, see Source and Accuracy Statement of U.S. Census Bureau (Various years-b). For NHIS, see Table 1 of U.S. Department of Health and Human Services (2014). For CE Survey, see U.S. Department of Labor (various years). For GSS, see Table A.6 of Appendix A – Sampling Design and Weighting in Smith et al. (2013).

Figure 2: CPS Unit Linkage Rates to the Master Address File



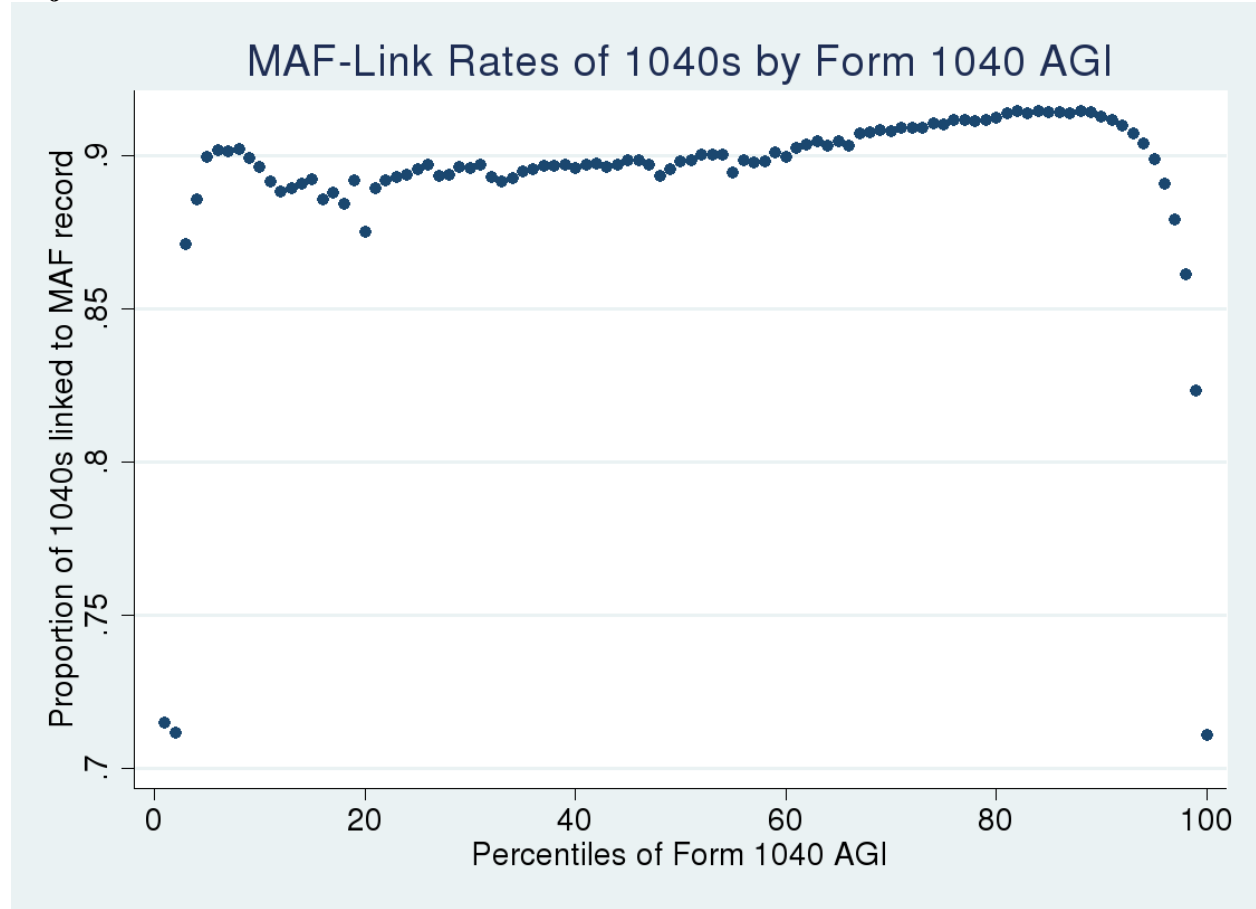
Note: Quantiles are assigned such that each contains an approximately equal sum of base weight. Zipcode-level mean AGI was assigned to ASEC units using tabulations from the Statistics of Income office at the IRS.

Figure 3: CPS Linkage Rates to IRS Form 1040 Records



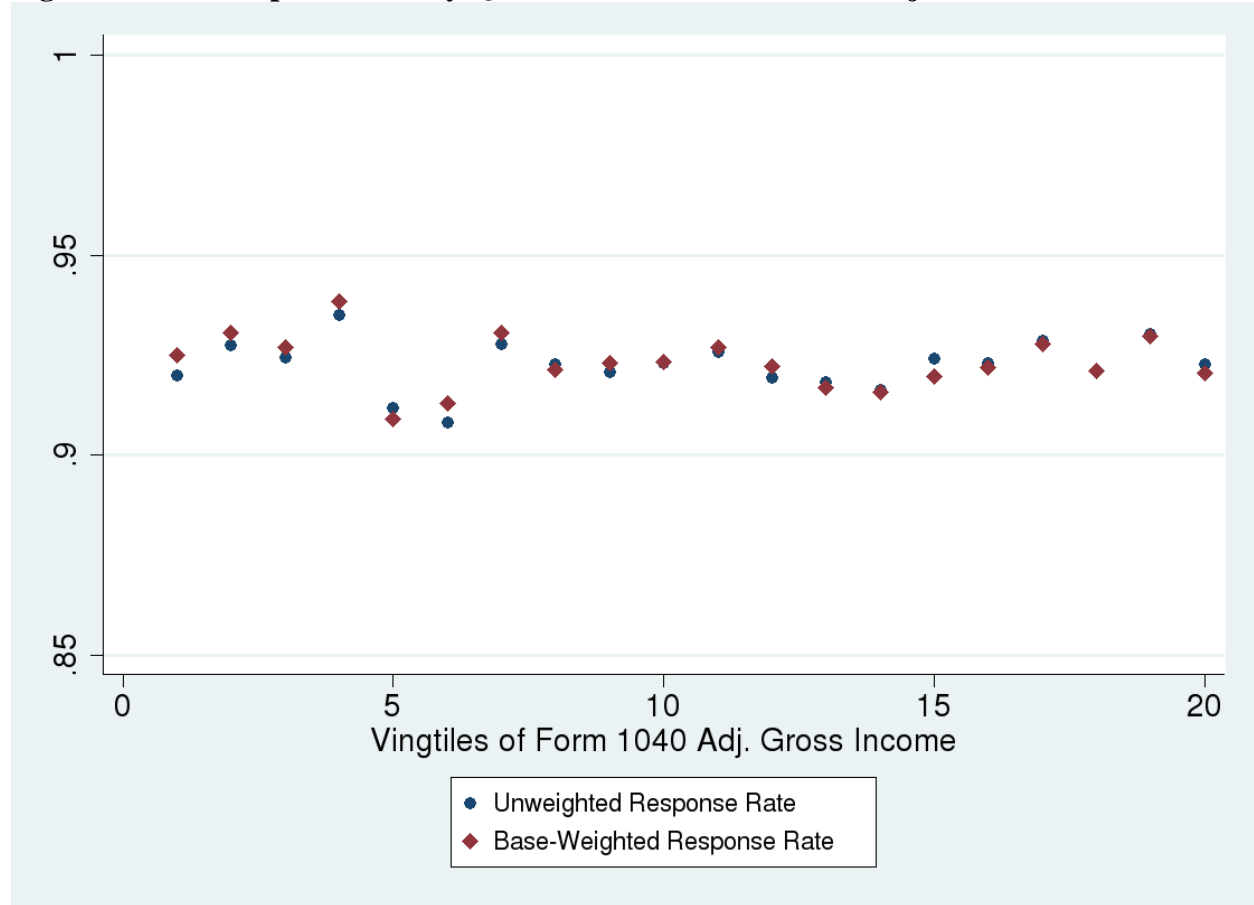
Note: The two data sources are linked using the MAF. Quantiles are assigned such that each contains an approximately equal sum of base weight. Zipcode-level mean AGI was assigned to ASEC units using tabulations from the Statistics of Income office at the IRS.

Figure 4: Tax Record Linkage Rate to Master Address File by Percentile of Adjusted Gross Income



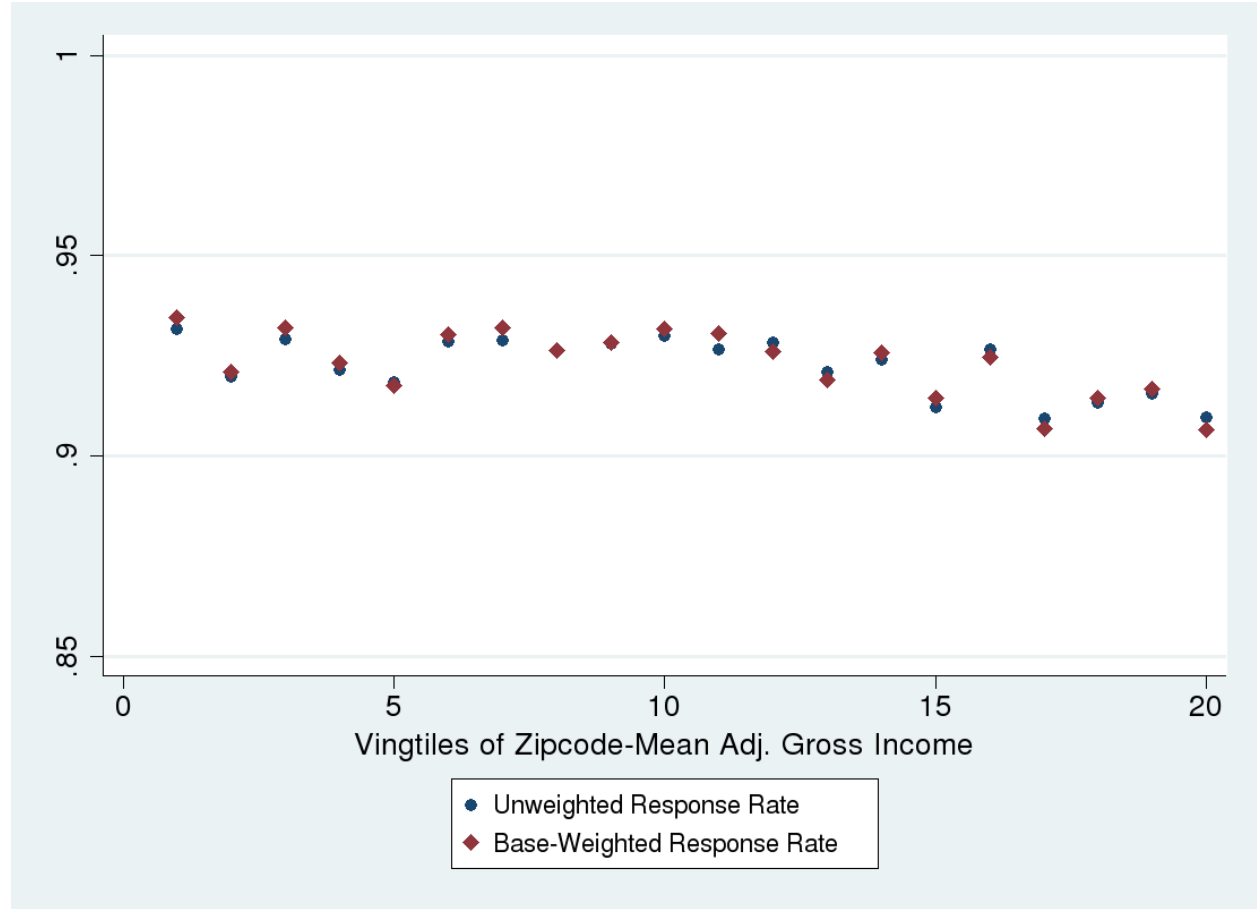
Note: Quantiles are assigned such that each contains an approximately equal sum of base weight. The distribution and link rates are for records of IRS Form 1040 filed in 2011 only from the 50 states and the District of Columbia and excludes returns filed from foreign countries, island areas/territories, and military addresses.

Figure 5: CPS Response Rate by Quantile of Household Level Adjusted Gross Income



Note: Sample is MAF-linked 2011 CPS ASEC households. AGI is the mean of records linked to a household. Quantiles are assigned such that each contains an approximately equal sum of base weight. Unit-level mean adjusted gross income (AGI) is calculated by summing the AGI of all Form 1040 records linked to each unit, then dividing by the number of linked Form 1040s.

Figure 6: 2011 CPS Response Rate by Quantile of Zipcode-level Mean Adjusted Gross Income



Note: Sample is MAF-linked 2011 CPS ASEC households. Quantiles are assigned such that each contains an approximately equal sum of base weight. Zipcode-level mean AGI was assigned to ASEC units using tabulations from the Statistics of Income office at the IRS.

Appendix

Direct Address Based Linking Results

Table A1: Trend in Unit Non-Response Rates of Major Household Surveys

	CPS (Basic)	SIPP (Wave 1)	NHIS	CE Survey	GSS
Trend	0.16 (0.05) ^a	0.52 (0.05) ^a	0.90 (0.16) ^a	0.62 (0.06) ^a	0.33 (0.07) ^a
N	17	14	17	30	19
R-squared	0.441	0.934	0.566	0.760	0.791

Notes: For each cell, we report the year coefficient from a regression of the percentage non-response rate on a constant and year, with its standard error underneath, followed by the sample size and R-squared. The regressions correct for first order autocorrelation using the Prais-Winsten procedure. The superscripts a, b and c, indicate that the coefficient is statistically significantly different from zero at the 1%, 5%, and 10% levels, respectively.

Table A2: Proportion of ASEC units that link to at least one IRS Form 1040 record

<i>Proportions of ASEC units that are linked to at least one Form 1040</i>	MAF-Linked			Directly Linked		
	(1)	p-value (1) = (2)	(2) Type A Non-Respondent	(4)	p-value (4) = (5)	(5) Type A Non-Respondent
	Respondent			Respondent		
Overall	.786	.003	.761	.649	.002	.618
Region						
Northeast	.790	.286	.774	.667	.387	.645
Midwest	.805	.613	.797	.665	.308	.640
South	.772	.141	.755	.636	.038	.604
West	.787	<.001	.732	.641	.003	.598
Urban	.793	.005	.767	.673	.004	.634
Rural	.757	.038	.726	.548	.231	.530
Frame						
Area	.695	.335	.677	.462	.595	.484
Group Quarters	.771	n/a	n/a	.454	n/a	n/a
Permit	.854	.030	.820	.506	.032	.464
Unit	.787	.011	.758	.692	.022	.657
Observations	75,178		6,547	75,178		6,547

Note: Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. Statistics ("link rates") are the base-weighted proportions of 2011 CPS ASEC units with the specified characteristics that are linked to at least one Tax Year 2010 IRS Form 1040 record. By construction, no Group Quarters units are Type A non-respondents.

Table A3: Comparing ASEC-Reported Demographic Characteristics Between Linked Respondents and Unlinked Respondents

1040 Characteristics	MAF-Linked		Direct-Linked		p-value (3)=(5)	(4) Unlinked Respondents	(5) All Respondents
	(1) Linked Respondents	p-value (1)=(5)	(2) Unlinked Respondents	(3) Linked Respondents			
Age							
Under 65 years	.804	<.001	.721	0.796	<.001	0.766	.785
15 to 24 years	.044	<.001	.079	0.043	<.001	0.068	.052
25 to 34 years	.163	.546	.161	0.156	<.001	0.175	.162
35 to 44 years	.188	<.001	.148	0.184	<.001	0.170	.179
45 to 54 years	.217	<.001	.171	0.219	<.001	0.184	.207
55 to 64 years	.193	<.001	.161	0.195	<.001	0.169	.186
65 years and older	.196	<.001	.279	0.204	<.001	0.234	.215
Education (age ≥ 25)							
Less than HS	.093	<.001	.178	0.093	<.001	0.145	.112
HS grad	.267	<.001	.305	0.268	<.001	0.291	.276
Some college	.270	<.001	.236	0.269	<.001	0.250	.262
College grad	.205	<.001	.130	0.205	<.001	0.158	.188
Post-grad	.121	<.001	.070	0.123	<.001	0.087	.110
Race and Hispanic origin							
White, not Hispanic	.717	<.001	.657	0.722	<.001	0.670	.703
Black, not Hispanic	.113	<.001	.160	0.114	<.001	0.140	.123
Asian, not Hispanic	.049	<.001	.041	0.047	0.721	0.047	.047
Other, not Hispanic	.005	.003	.009	0.004	<.001	0.010	.006
Hispanic (any race)	.117	<.001	.133	0.114	<.001	0.132	.120
Marital status							
Married	.556	<.001	.334	0.554	<.001	0.420	.506
Widowed	.084	<.001	.152	0.090	<.001	0.117	.100
Divorced	.144	<.001	.186	0.144	<.001	0.171	.154
Separated	.025	<.001	.051	0.025	<.001	0.041	.031
Never married	.190	<.001	.277	0.187	<.001	0.251	.210
Nativity							
Native born	.861	.495	.859	0.866	<.001	0.851	.861
Foreign born							
Naturalized citizen	.075	<.001	.060	0.074	0.010	0.067	.071
Not a citizen	.064	<.001	.081	0.060	<.001	0.081	.068
Observations	58,910		16,268	48,380		26,798	75,178

Note: Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. All statistics in this table are weighted using final ASEC household weights. Observations are at the household level. "Linked" respondents are respondent units that we successfully linked to at least one Form 1040 record. Education universe for MAF-linking consists of 56,521 linked units and 15,084 unlinked units, while 3,573 units are out-of-universe due to a householder aged under 25. Education universe for exact-linking consists of 46,490 linked units and 25,115 unlinked units, while 3,575 units are out-of-universe due to a householder aged under 25.

Table A4: Comparing AGI and Other Characteristics Between Linked and All TY2010 IRS Form 1040 Records

Form 1040 variable

	ASEC-MAF-Linked 1040s	Direct-Linked 1040s	All 1040s
Mean AGI	\$ 56,853	\$ 52,826	\$ 56,794
Percentiles of AGI			
1	\$ -207	\$ -125	\$ -1,400
5	2,500	2,358	2,422
10	5,736	5,526	5,749
25	14,617	13,585	14,269
50	33,399	30,426	32,049
75	69,403	63,866	66,662
90	116,754	110,103	113,897
95	160,015	151,822	157,959
99	356,186	338,726	360,505
Married filing jointly	.396	.369	.388
Mean count of exemptions for children at home	.702	.660	.592
Number of dependents			
0	.581	.588	.630
1	.179	.180	.164
2	.154	.148	.132
3	.062	.060	.053
4	.024	.024	.022
Receipt of income sources			
Wages, salary	.851	.839	.825
Interest	.385	.358	.392
Dividends	.192	.189	.198
Rent and royalties	.072	.068	.075
Social Security	.148	.150	.178
Schedules filed			
A: Itemized deductions	.346	.319	.335
C: Profit/loss from business	.158	.155	.159
D: Capital gains, losses	.201	.188	.212
E: Supplemental income, loss	.123	.115	.130
F: Profit or loss from farming	.014	.011	.014
SE: Self-employment tax	.122	.119	.122
Number of records	84,521	88,328	138,141,584

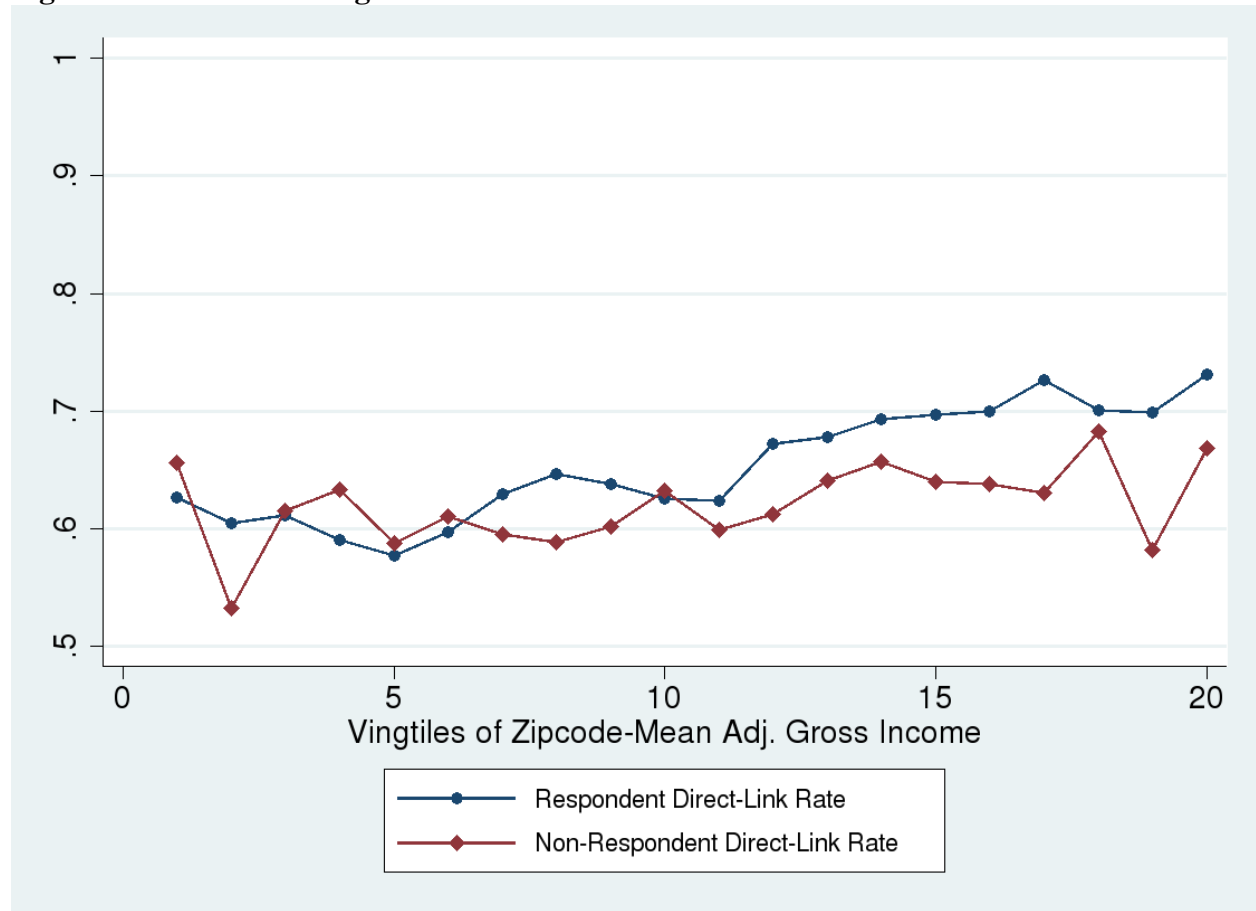
Note: Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. The universe of "All 1040s" includes only the 50 states and the District of Columbia, excluding 1040s filed from foreign countries, island areas/territories, and military addresses. "MAF-Linked 1040s" refers to the set of Form 1040 records that are assigned a MAF ID by the Census Bureau Center for Administrative Records Research and Applications (CARRA) MAFMatch probabilistic linking algorithm.

Table A5: Comparing TY2010 IRS Form 1040 AGI between Linked Respondents and Linked Non-Respondents to the 2011 CPS ASEC, after reweighting for non-linkage

Unit-mean 1040 variables	MAF-Linked		Direct-Linked			
	(1)	(2)	(3)	(4)		
	Base-Weighted Respondents	p-value (1)=(2)	Base-Weighted Non-Respondents	Base-Weighted Respondents	p-value (5)=(6)	Base-Weighted Non-Respondents
AGI						
Mean	\$ 61,868	.510	\$ 63,546	\$ 60,802	.760	\$ 61,678
Percentiles						
1	\$ 0	.580	\$ -1,050	\$ 1	.430	\$ -2,580
5	6,810	.777	6,952	6,748	.878	6,830
10	11,179	.145	11,747	11,112	.526	11,366
25	21,074	.110	21,771	20,633	.786	20,778
50	39,854	.277	40,887	38,998	.890	39,128
75	73,983	.802	73,595	73,265	.234	71,413
90	120,806	.909	120,416	120,360	.463	117,907
95	167,344	.475	172,467	166,802	.690	169,984
99	362,223	.439	397,074	365,635	.787	380,352
Married filing jointly	.464	<.001	.406	.467	<.001	.419
Mean count of exemptions for children at home	.787	<.001	.703	.779	.002	.720
Number of dependents						
0	.539	<.001	.573	.538	.022	.556
1	.188	.852	.189	.188	.721	.190
2	.174	.009	.157	.173	.421	.167
3	.071	.003	.059	.072	.094	.064
4	.028	.025	.022	.029	.006	.022
Receipt of income sources						
Wages, salary	.838	<.001	.863	.833	<.001	.854
Interest	.417	.408	.409	.412	.185	.400
Dividends	.207	.014	.189	.204	.011	.186
Rent and royalties	.080	.437	.077	.081	.642	.079
Social Security	.177	<.001	.131	.179	<.001	.136
Schedules filed						
A: Itemized deductions	.385	.012	.406	.384	.084	.400
C: Profit/loss from business	.173	.435	.178	.176	.275	.183
D: Capital gains, losses	.221	.013	.204	.218	.003	.197
E: Supplemental income, loss	.137	.633	.134	.138	.689	.135
F: Profit or loss from farming	.014	.003	.009	.015	.012	.010
SE: Self-employment tax	.132	.619	.135	.134	.346	.139
Observations	53,559		4,500	48,380		3,983

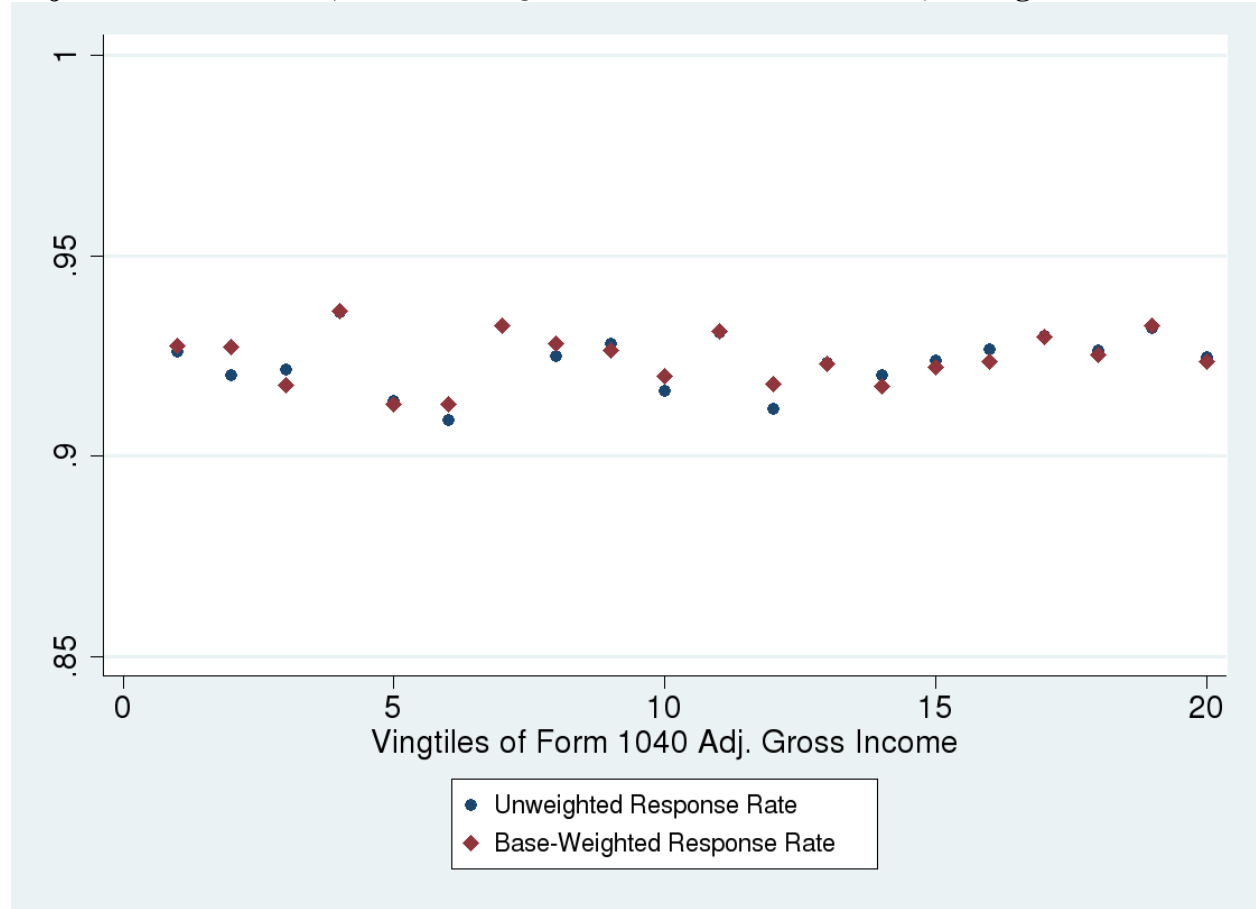
Note: Listed *p*-values are generated from tests of equality of estimates in the designated pairs of columns. Statistical testing accounts for non-independent sampling of units due to complex survey design, using replicate weights reflecting the ASEC sample design. Direct linking is a merge of the ASEC to the Form 1040 extract on addresses that were cleaned and parsed using the SASDQ module. The tabulated unit-level statistic is the mean of Adjusted Gross Income values across all Form 1040s linking to each unit. Means and percentiles are rounded to the nearest dollar. The MAF-linked units are selected such that MAFIDs are unique; within each group of units with duplicate MAFID values, one unit is chosen at random, with equal probability. Reweighting for non-linkage is carried out by probits run separately for respondents and non-respondents, including frame, state/urban, presence of PO box or rural route, and type of living quarters (house/apartment, mobile home, or other).

Figure A1: Rate of linking ASEC to 1040s via direct address match



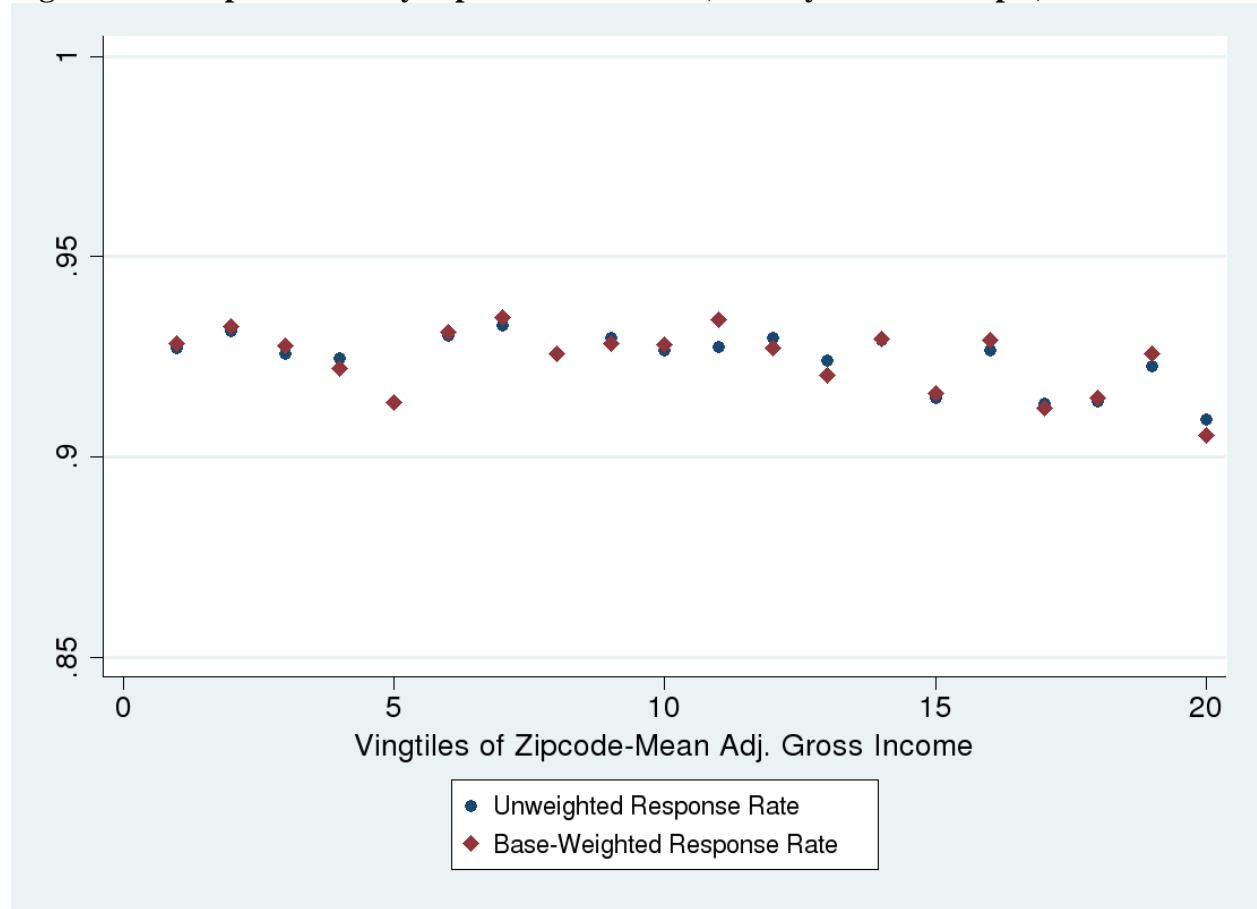
Note: Quantiles are assigned such that each contains an approximately equal sum of base weight. Zipcode-level mean AGI was assigned to ASEC units using tabulations from the Statistics of Income office at the IRS.

Figure A2: 2011 CPS ASEC Response Rates by Quantiles of Unit-Mean Form 1040 Adjusted Gross Income, from SASDQ-Parsed Address Direct Link, among Linked Units



Note: Quantiles are assigned such that each contains an approximately equal sum of base weight. Unit-level mean adjusted gross income (AGI) is calculated by summing the AGI of all Form 1040 records linked to each unit, then dividing by the number of linked Form 1040s.

Figure A3 : Response Rate by Zipcode-Mean AGI (Directly Linked Sample)



Note: Quantiles are assigned such that each contains an approximately equal sum of base weight. Assignment of units to quantiles is constant across series. Zipcode-mean Form 1040 adjusted gross income is published annually by the Internal Revenue Service Statistics of Income program. For weights available for both respondents and non-respondents (GESBWGT, HWWCWGT, HMNIWT), the response rate is calculated as the sum of the weight for the respondents over the sum of the weight for all units in the quantile. For the other (respondent-only, householder) weights, the weights are first rescaled to sum to the overall sum of the base-weight, then the response rate is the sum of the weights for the respondents over the sum of the base weight for all units in the quantile.