CARRA Working Paper Series
Working Paper 2018-01

Investigating the Use of Administrative Records in the Consumer Expenditure Survey

Quentin Brummet, Denise Flanagan-Doyle, Joshua Mitchell and John Voorheis
U.S. Census Bureau

Laura Erhard and Brett McBride
Bureau of Labor Statistics

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: March, 2018

Investigating the Use of Administrative Records in the Consumer Expenditure Survey[1]

Quentin Brummet, Denise Flanagan-Doyle Joshua Mitchell, and John Voorheis

U.S. Census Bureau

Laura Erhard and Brett McBride

Bureau of Labor Statistics

February 14, 2018

DRAFT—NOT FOR CIRCULATION OUTSIDE OF CENSUS

1 Introduction

The Consumer Expenditure Survey (CE) is the most important data source on the income and expenditure habits of American households. CE data are used, among other purposes, to update the basket of goods and services underlying the Consumer Price Index (CPI). The CPI in turn is used to calculate cost of living adjustments both for workers at many firms and for recipients of several government transfer programs (such as Social Security). Maintaining the high quality of this survey data is thus of the utmost importance.

Given rising costs of survey administration and falling response rates, there is substantial interest in investigating if alternative data sources–such as administrative records from other federal government agencies, or data provided by commercial vendors–can be used to improve surveys such as the Consumer Expenditure Survey. These alternative data sources have the potential to reduce respondent burden, reduce measurement error, and improve processes for imputation and non-response adjustment. Of particular interest is the possibility of enhancing CE questions about income with administrative income records from the Internal Revenue Service (IRS).

In this paper, we report results from a series of exploratory analyses that represent the first step in examining whether administrative records can be used to enhance CE income questions. To do perform this analysis, we leverage the Census Bureau's data linkage infrastructure, which allows for the direct person and address based linkage of survey responses to administrative records, and is thus ideally suited for these types of analyses. In this project, responses to the Consumer Expenditure Survey from 2013 and 2014 are linked to three sources of administrative records

[1] This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the Bureau of Labor Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

from the IRS: the universe of Form 1040 tax returns, the universe of Form W-2 wage reports, and the universe of other information returns.

We perform two types of record linkage: a person-based linkage, using anonymous individual identifiers, and a household-based linkage, using address and location information. Using these linked datasets, we perform a series of analyses that provide information about measurement error, non-response bias, and the quality of CE imputation and weighting processes. These analyses compare CE data about wages, Social Security payroll deductions, self-employment income and retirement income to administrative data on these same topics.

The remaining analysis proceeds as follows. We first describe the survey and administrative data, the matching processes used to link the survey and administrative data, and the results of the matching processes, summarized by match rates. We then provide further analysis relevant for the matching process to examine how match rates would differ if the CE collected more fine-grained information on date of birth. We then compare individual CE responses to income questions with linked administrative data on the same income concepts. We link responding and non-responding households to administrative records by address and analyze non-response bias. We conclude with a summary of how these linked data can inform current and future efforts to improve the quality of the CE.

2 Data and Matching

To analyze survey measurement error, we need to compare survey responses to administrative records covering the same content and the same individuals. Doing so requires internal CE research files, high quality administrative records on the universe of people, and a method for linking survey responses to administrative records. In this section, we describe each of these three items.

2.1 Survey Data

The Consumer Expenditure Survey program consists of two surveys: the Quarterly Interview Survey, which is designed to measure income and expenditures for lower-frequency purchases; and the Diary Survey, which is designed to measure expenditures on high frequency purchases. We focus on the Quarterly Interview Survey, which asks more detailed questions about income. During 2013-14, the Quarterly Interview Survey contacted approximately 12,000 sample units each quarter, and responding consumer units (CUs) were interviewed 5 times in total.[2]

Questions about income (including wage receipt, presence of social security payroll deductions, self-employment, and retirement income) received in the past twelve months were asked of all CU members over age 14 on the second interview. These responses were carried over to the third and fourth interviews for individuals who remained in the CU; individuals who entered a CU

[2] For more information about the survey design of the CE see: https://www.bls.gov/opub/hom/cex/pdf/cex.pdf

after the second interview or who previously reported not working, but now were working, were asked the income questions in the third or fourth interviews. All CU members were then asked the income questions once more in their fifth and final interview. If a CU member did not report an amount to an income question, while acknowledging receipt of such income to a previous screening question in the second interview, BLS imputed a response using a multiple regression based imputation method for interviews 2-4. Our analysis will proceed by comparing responses from the first time an individual responded to an income question with administrative records (for a vast majority of CU members, this occurs in the second interview).

In performing this analysis, we received two CE files from BLS. The first file contains personally identifiable information and location information collected by the Census Bureau on behalf of BLS in the administration of the CE survey (frame data). This file contains PII for all individuals in CUs who participated in all interviews conducted in July-December 2013 and all first interviews conducted in 2014. The second file, delivered by BLS, is an extract of the CE research microdata file, containing non-topcoded (but edited and imputed) income responses, as well as a variety of other demographic and family-level variables for all interviews conducted between July 2013 and December 2014. Both files contain unique identifiers for CUs and individuals within CUs, and so our analysis in section 4 will proceed using the set of individuals present in both files (the address-matching based analysis in section 5, however, is conducted using all sample units in the frame data).

2.2 IRS Administrative Records

The administrative records used in our analysis are collected by the Internal Revenue Service (IRS) from three types of tax filing forms–Form 1040 tax returns, Form W-2 wage statements, and Form 1099 information returns (filed in reference to various types of non-wage payments). Forms W-2 and 1099 are filed by third parties (employers, pension funds or other payers) and sent both to the IRS and to the taxpayer, while Form 1040 is filed by a taxpayer and sent to the IRS. These three types of administrative records represent slightly different universes. Form W-2 data represents the universe of individual wage earners, while the Form 1099 data represents the universe of other non-wage payees (e.g. retired individuals) for payments covered on these forms. These two datasets contain individual level data, while the Form 1040 data represents the universe of tax filing units, and contains tax unit-level data.

While the IRS extracts contain the universe of individuals (W-2 and 1099) or tax filing units (1040) for a given year, CARRA only receives a subset of the fields reported on these forms. Relevant for this analysis, CARRA's Form 1040 data extract contains information on filing status, the amount of wage and salary income and adjusted gross income, the amount of interest and dividend income, and flags for whether a tax unit filed schedule C or SE, which we use as an indicator for the receipt of self-employment income. CARRA's Form W-2 data contain an employer identifier, the amount of wages and tips, the amount of wages subject to FICA taxes (which we use as an indicator for whether an individual has Social Security or Medicare deductions taken out from their paycheck), as well as the amount of deferred compensation.

CARRA's Form 1099 data contain only an indicator for whether a form was filed for most types of 1099 filings, with the exception of Form 1099-R for retirement income. CARRA's Form 1099-R data contain information on the amount of pension or retirement fund distribution received by a taxpayer, as well as the type of distribution (pension, IRA, etc.). We deduplicate IRS administrative records to the extent it is possible given the information available on CARRA's data within each file.

2.3 Matching Procedure

The analysis requires that we match the CE to IRS records. The two matching processes we will use are a person-based match, where Protected Identification Keys (PIKs) are assigned to individual survey or administrative records based on available PII, and an address based match, where Master Address File identifiers (MAFIDs) are assigned to survey households or administrative tax units based on available address information. Once a file has been assigned these identifiers, it is possible to link this file with any other file that has been assigned MAFIDs and or PIKs.

CARRA's person-based matching process, the Person Validation System (PVS), is a probabilistic matching algorithm that assigns protected identification keys (PIKs) to individuals in administrative and survey data based on available personally identifying information. If an individual's Social Security number is included in a dataset, then records can be directly linked to the SSA Numerical Identification File (Numident) file. If no SSN is available on a file (as in CE), then information including name, address and date of birth (or year of birth if exact date is unavailable) is used to probabilistically match survey records to the SSA Numident. If only age is available (as in CE), PVS searches over a range of years of birth consistent with the response. See Wagner and Layne (2014) for an in-depth discussion of PVS.

CARRA's address matching methodology (MAF match) consists of an initial file edit process, then processing addresses through a commercial product to clean the addresses. The first step in the MAF match process is to standardize and clean addresses using SAS Dataflux, which updates zip codes, adds zip4 if needed, returns a preferred street name, corrects minor misspellings, and outputs the address in one string. The cleaned addresses are then parsed and standardized using the US Census Bureau's Geography Division (GEO) Address Standardization routine. The match process uses a generalized probabilistic matching routine to compare information from the parsed fields to the Census Bureau's Master Address File. Successful matches are assigned a MAFID.

As part of regular CARRA operations, Form 1040 tax returns, Form W-2 wage statements, and Form 1099 information returns have all undergone the PVS and MAF match processes (and new deliveries of the IRS data are subsequently processed upon receipt). CE data had not been processed through PVS before the start of this project, although the 2013 CE had previously been assigned MAFIDs as part of a separate project. CARRA received a file from the CE office at Census which contained detailed survey frame information, including address information for both respondents and non-respondents, and PII including name, age at survey response, sex, race

and ethnicity, for all CE interviews conducted from July-December 2013 and all wave 1 interviews conducted (or attempted) in 2014. CARRA staff then processed this file through PVS and MAF match.

Once PIKs and MAFIDs have been assigned to the CE, it is possible to match CE observations to administrative records using these unique identifiers. As noted, BLS sent CARRA a file containing information on income and demographics. This file, and the frame data sent by the CE office at Census, both contain internal CE identifiers. We proceed with our analysis using the first interview in which the respondent provides income information for all individuals who appear in both CE files.[3]

We create two linked datasets. We first link the CE data to the three IRS administrative records datasets by PIK to create a PIK-linked dataset. Since it is possible for an individual to appear on more than one Form 1040 or Form W-2, we deduplicate these cases as follows. If an individual appears on multiple 1040s as both a dependent and a primary filer, we select the 1040 in which the individual appears as a primary filer. If an individual appears on multiple unique W-2s, we treat these as distinct income sources, and sum all W-2 wages, FICA wages and deferred compensation to form total W-2 amounts for these fields. We also create a MAFID-linked dataset. Again, since multiple 1040s or W-2 can be matched to a given MAFID, we deduplicate matched records. We omit any CE MAFIDs which are matched to 10 or more 1040s, and then aggregate 1040 records for a given MAFID by averaging across all linked 1040s for each 1040 field. Since CE income questions ask about income received in the past twelve months, we match CE records to IRS administrative records from the previous tax year (i.e. matching the 2013 CE to tax year 2012 IRS data).

Table 1 presents PIK rates and PIK-based match rates to each of the three IRS administrative records datasets, broken out by CE-derived demographic characteristics. PIK assignment rates are a bit lower than MAFID assignment. Overall, about 77 percent of CE respondents can be assigned a PIK. About 70 percent of CE respondents can be matched to the universe of 1040s, 39 percent can be matched to a W-2, and about 59 percent can be matched to the 1099 data. The relatively lower match rates to the universe of W-2s are to be expected, as not everyone is employed in a given year.[4] There are patterns across race, ethnicity and age groups that are consistent with other linkage research (O'Hara et al., 2012, Luque and Bhaskar, 2014). Non-whites and Hispanics are less likely to be assigned a PIK, as are younger individuals and individuals with lower incomes.[5] Thus, these groups are also less likely to be matched to the administrative records.

[3] Throughout, we refer to each CU member as a respondent, regardless of whether they personally responded to an interview question, or a proxy response was recorded.

[4] For comparison, 100*(0.39/0.766)=50.9 percent of individuals with PIKs can be matched to the universe of W-2, while the employment-to-population ratio was 58.5 percent in January 2013. For the prime age 25-44 age group, 78.9 percent of individuals with a PIK can be linked to the W-2's, compared to a prime age EPOP of 75.6 percent.

[5] We split the sample into high and low income at the median CU wage (FSALARYM). The median CU wage for the pooled 2013-2014 sample is about $45,000.

Table 1: PVS Match Rates

	Obs.	CE PIK Rate	CE-IRS 1040 Match Rate	CE-IRS W-2 Match Rate	CE-IRS 1099 Match Rate	CE-Any IRS Match Rate
All CE Respondents	36,437	0.766	0.704	0.39	0.586	0.752
Sex						
Female	18,761	0.761	0.700	0.374	0.575	0.747
Male	17,676	0.771	0.709	0.408	0.597	0.757
Race						
White Alone	28,905	0.778	0.719	0.402	0.607	0.765
Black Alone	4,341	0.752	0.662	0.354	0.53	0.732
AIAN Alone	177	0.768	0.655	0.362	0.559	0.734
Asian Alone	2,108	0.656	0.612	0.348	0.476	0.637
Pacific Islander Alone	236	0.665	0.64	0.36	0.492	0.665
Two or more races	670	0.743	0.678	0.261	0.415	0.722
Age						
0-13	6,763	0.675	0.655	0.006	0.073	0.659
14-24	5,353	0.733	0.702	0.369	0.502	0.717
25-44	9,280	0.763	0.721	0.602	0.695	0.75
45-64	9,597	0.824	0.750	0.579	0.77	0.809
65+	5,444	0.814	0.658	0.195	0.794	0.804
Hispanic Origin						
Hispanic	6,628	0.688	0.626	0.295	0.422	0.664
Not Hispanic	29,809	0.783	0.721	0.412	0.622	0.771
Income						
Above Median CU Wages	18,682	0.79	0.768	0.491	0.602	0.784
Below Median CU Wages	17,755	0.741	0.637	0.284	0.569	0.718

Source: CE 2013-2014, IRS 1040s, W-2s and 1099s 2012-2013

Note: All Rates are calculated as # Matches/N, where N is the number of unique individual CE respondents from July 2013-December 2014 in a given category

Table 2 presents MAFID assignment rates (i.e. CE-MAF match rates) and match rates to the 1040 and W-2 data.[6] Overall, 93 percent of sample unit can be assigned a MAFID, while 68 percent can be matched to the universe of 1040s, and about 51 percent can be matched to the W-2 data. However, there are substantial differences across participation in these match rates. Type A non-interviews (non-interviews at valid, non-vacant addresses) and CE participants have similar patterns of match rates. About 95 percent of CE participants can be assigned a MAFID, while 77 percent can be matched to 1040s and 58 percent to W-2s, compared to 94 percent, 76 percent and 57 percent for type A (valid) noninterviews, respectively. Type B and C noninterviews represent attempted interviews at addresses that are either vacant (type B) or invalid (type C). These noninterviews have much lower match rates, as expected – less than 10 percent of type C noninterviews can be matched to the W-2 data, for instance.[7]

Table 2: MAF Match Rates by Nonresponse type

	Obs.	CE-MAF Match Rate	CE-IRS 1040 Match Rate	CE-IRS W-2 Match Rate
All Sample Units	30,917	0.931	0.682	0.508
		Participation		
CE Participant	19,677	0.950	0.774	0.582
Valid Non-interview	5,761	0.942	0.761	0.565
Vacant Non-interview	4,159	0.867	0.307	0.209
Invalid Non-interview	1,320	0.789	0.132	0.098

Source: CE 2013-2014, IRS 1040s and W-2s 2012-2013

3 Date of Birth Analysis

The PVS match rates for the CE are somewhat lower than the PIK assignment found for several other demographic surveys. For instance, the 2013 Current Population Survey Annual Social and Economic Supplement has been processed via PVS has a PIK assignment rate of about 91 percent. It is likely that the relatively low rate of PIK assignment for CE is at least partly due to the coarse PII collected in CE. Unlike other comparable surveys, the CE asks the respondent's age only rather than the full date of birth. Day and month of birth are generally useful fields for linking person records.

[6] A MAFID match is defined as a case where at least one 1040 or W-2 can be matched to the MAFID assigned to a sample unit (some sample units are matched to multiple 1040s or W-2s, as we discuss in Section 5).
[7] Although it may be somewhat surprising that these match rates are nonzero at all, note that the Type C noninterview subtypes with the highest match rates are OUTCOME codes 246, 252 and 341, which all correspond to locations which could be valid addresses for form 1040 purposes, but would not be valid addresses for CE.

To estimate the difference it might make to the PIK rate were CE to collect full date of birth, we performed an additional matching analysis. We used the American Community Survey (ACS), since CARRA staff have access to the full PII collected in the ACS. We compared the PIK assignment rate for the 2013 ACS that was obtained using the full date of birth to an alternative procedure that uses age but not date of birth.

The overall PIK assignment rate for the 2013 ACS is 94 percent. Using the alternative procedure, we obtained a rate of about 84 percent. If collecting full date of birth would make a similar difference for CE, that would bring the CE PIK rate into the neighborhood of the PIK rates of other surveys and censuses that collect date of birth information, including the 2010 Decennial Census (86.4 percent), the CPS ASEC (87 percent in 2016) and the Survey of Income and Program Participation (91 percent in 2014).

4 Imputation and Measurement Error Analysis

We begin with a set of analyses using the individual PIK linkages, which provide useful information about measurement error in CE data and the quality of CE imputations for item non-response. These analyses compare the distributions of wages, self-employment income receipt, retirement income, and social security payroll deductions in the survey and administrative data. In each of these analyses, we present results for the sub-sample of CE respondents who receive a PIK and who appear in both the frame data and cleaned CE internal research files. Some analyses will further subset this data to consider only individuals who can be linked to the administrative records, or who have non-zero income in both the survey and administrative data.

4.1 Wage Receipt

We first validate the CE wage and salary receipt question. The SALARYST variable in the CE research file is a binary variable, coded from the survey question “Did you receive any wages, salary, tips bonuses or commissions?” with a reference period of the previous twelve months. Note that although BLS imputation procedures theoretically may impute values for non-response, there are a minimal imputed cases in our sample for SALARYST. We compare this to an administrative records analogue, a binary variable coded based on whether an individual appears in the universe of Form W-2’s from the previous tax year. The universe of Form W-2’s is not necessarily the same as the universe of all wage earners. The two largest populations of wage earners not covered in the form W-2 universe are workers in the informal sector and independent contractors (or some other non-traditional employment relationship) who may consider their earnings to be wages rather than self-employment income, but who receive earnings reports via Form 1099 and not Form W-2. Work in the informal sector is fundamentally

unmeasurable in the administrative records. The incidence of independent contractor work, however, can partially be examined in the 1099 file to get a sense of misclassification error.[8]

Table 3 presents summary statistics for the two measures of wage receipt, broken down by socio-demographic categories. The second and third columns present the unweighted average of the binary variable for CE wage receipt, and the binary variable for appearing on a W-2 respectively. Since the underlying variables are binary, we interpret these averages as proportions. The final two columns present weighted averages of the two binary variables. Both the CE wage receipt variable and the W-2 wage receipt variable are weighted by the final CE weights (FINLWT21) multiplied by the inverse probability of PIK assignment.[9] This re-weighting is done to address potential systematic differences in PIK assignment across demographic groups.

The proportion of individuals who report CE wages corresponds well with the proportion of individuals who appear on W-2's. Overall, about 50 percent of the linked sample reports W-2 wages and an almost identical proportion report wages in the CE. There is variation across different demographic categories, both in the proportion of individuals who report wages in either data source and in the differences in proportions across data sources. Asians, whites, and Pacific Islanders are more likely to report wages in either data source than are the other race groups, and also have larger differences in the proportion of individual reporting CE vs. W-2 wages. For instance, about 54.6 percent of Asians report CE wages, whereas 51.4 percent have W-2 wage reports. On the other hand, 44.5 percent of blacks report CE wages and 46.5 percent have W-2 wage reports.

The CE survey question and the administrative records have inconsistent time periods, which may lead to discrepancies even if wages are reported correctly in both sources. The W-2 records report wages earned in the previous tax year (i.e. the previous calendar year), while the survey question asks about wages earned in the past twelve months. These two time periods might have a very high or very low degree of overlap, depending on which month a respondent answered the income questions. For instance, a respondent in January 2014 would have a high degree of overlap (11 months), while a respondent in December 2014 would have low overlap (1 month). Despite this concern, there is no clear pattern of an increasing or decreasing gap depending on the degree of overlap between the two income concepts' time periods.

Unsurprisingly, the two age groups roughly corresponding to "prime working age" – 25 to 44 and 45 to 64 – have the highest proportion of wage receipt in either data source. Interestingly, although these middle two age groups have higher proportions of CE wage receipt than W-2 wage receipt, the pattern is reversed for the oldest and youngest age groups – 49.8 percent of individuals aged 14 to 24 report W-2 wages, while only 47.5 percent report CE wages.

[8] CARRA's extract of the information returns master file (IRMF) contains 1099-MISC, which are filed for some contractors, but not 1099-K, a form that has become more common (filed, e.g. on behalf of drivers for Uber and Lyft). CARRA's extract also does not currently have dollar amounts, only the PII on the form.

[9] The probability of PIK assignment is estimated from a probit model of the form $P(PIK|X) = \Phi(X\beta)$, where X is a matrix of the demographic variables presented in the tables: dummy variables for sex, race, age group and Hispanic origin.

Table 3: Proportion of Individuals Who Received Wages by Data Source

	Obs.	Unweighted		Weighted	
		CE Wages	W-2 Wages	CE Wages	W-2 Wages
All individuals	27,914	0.508	0.509	0.502	0.502
Sex					
Female	14,278	0.476	0.491	0.485	0.481
Male	13,636	0.543	0.529	0.548	0.517
Race					
White Alone	22,477	0.517	0.517	0.512	0.512
Black Alone	3,264	0.449	0.471	0.445	0.465
AIAN Alone	136	0.456	0.471	0.453	0.45
Asian Alone	1,382	0.556	0.531	0.546	0.514
Pacific Islander Alone	157	0.573	0.541	0.568	0.52
Two or more races	498	0.357	0.351	0.327	0.317
Age					
14-24	3,923	0.477	0.503	0.475	0.498
25-44	7,078	0.815	0.79	0.814	0.79
45-64	7,911	0.715	0.702	0.716	0.701
65+	4,434	0.2	0.239	0.201	0.24
Hispanic Origin					
Hispanic	4,561	0.45	0.428	0.435	0.413
Not Hispanic	23,353	0.52	0.525	0.515	0.52
Income					
Above Median CU Wages	14,760	0.652	0.622	0.637	0.606
Below Median CU Wages	13,154	0.347	0.383	0.35	0.385
Second Interview Month					
1	1,917	0.514	0.505	0.506	0.499
2	2,120	0.51	0.502	0.503	0.495
3	2,087	0.515	0.516	0.511	0.509
4	2,124	0.534	0.517	0.526	0.506
5	2,072	0.49	0.497	0.482	0.491
6	2,145	0.502	0.51	0.491	0.498
7	2,152	0.495	0.51	0.49	0.505
8	2,147	0.501	0.509	0.496	0.505
9	2,090	0.517	0.507	0.516	0.503
10	3,094	0.516	0.518	0.514	0.512
11	3,099	0.494	0.502	0.48	0.484
12	2,867	0.513	0.517	0.508	0.515

Source: CE 2013-2014, IRS W-2s 2012-2013
Note: The "Weighted" columns use weights equal to the CE sample weights multiplied by the inverse probability of being assigned a PIK.

Note that Table 3 simply compares the distributions of the two wage receipt concepts. Comparing the averages of these two distributions is instructive, but it should be noted that the distributions can differ for two reasons: an individual may report CE wages despite not appearing in the universe of W-2's, or an individual may appear in the universe of W-2 despite reporting no wages on the CE. To examine the relative sizes of these two sources of inconsistency, Table 4 presents a cross-tabulation of wage receipt for the two data sources in Panel A. Over 85 percent of the linked sample has concordant responses between the two sources – 44.3 percent report both CE and W2 wages, while 42.5 percent report neither CE nor W2 wages. The two types of inconsistencies are roughly balanced in size: 6.7 percent of linked respondents report W-2 but not CE wages, while 6.6 percent report CE but no W-2 wages.

Table 4: Wage Receipt Agreement by Data Source (%)

Panel A: W-2 Wages

	W-2 Wages	No W-2 Wages
CE Wages	44.28	6.56
No CE Wages	6.67	42.5

Panel B: W-2 or 1099-MISC Wages

	W-2 or 1099-MISC Wages	No W-2 or 1099-MISC Wages
CE Wages	46.29	4.55
No CE Wages	9.41	39.75

Source: CE 2013-2014, IRS W-2s and 1099-MISC 2012-2013

As noted above, comparing reported CE wage receipt with receipt of a Form W-2 potentially excludes independent contractors or other workers who may not be present in the W-2 universe. To examine the degree to which this could be explain the proportion of CE respondents who report CE wages, but are not in the W-2s, Panel B of Table 4 reports an additional agreement exercise. This lower panel shows the degree of agreement between CE wage receipt and either having a form W-2 or having a form 1099-MISC, an information return filed on behalf of some but not all independent contractors. When considering this additional administrative records

source of wage receipt, the proportion of individuals with CE wages but no wages in the W-2 or 1099-MISC decreases to 4.55 percent. However, the proportion of individuals who report no CE wages but received a W-2 or 1099-MISC increases to 9.41 percent.

4.2 Reported Wage Amounts

Next, we turn our analysis to a comparison of the amount of wages reported on the CE with the amount of wages reported in the Form W-2 linked sample. We begin by comparing the central tendency of the two wages distributions using summary statistics (average and median), and then compare the distributions beyond the mean through the use of data visualization (density plots).[10] For brevity, we focus our discussion of the summary statistics on average wages, but additional tables summarizing median wages across income sources are available in the appendix.[11] We will separately analyze the measurement error properties of the two data sources—by comparing the non-imputed CE responses to the W-2 Wage data—as well as analyzing CE imputation procedures by comparing CE imputed and non-imputed responses to the W-2 wage data.

We begin with an analysis of the non-imputed responses to the CE question asking "How much did you earn before taxes?" which is asked to all individuals who responded affirmatively to the wage receipt question analyzed previously. We compare this total (variable SALARYX) to the total of all unique W-2 wages for each linked individual in the sample. As with previous analyses, there are slight differences between the two wage concepts. CE wages refer to all wages before taxes; we define W-2 wages as the sum of all taxable wages and deferred compensation. The major wedge between the two wage concepts is due to pre-tax payroll deductions (e.g. for 401k retirement contributions). W-2 wages will be smaller than pre-tax wages by the amount of any pre-tax deductions.[12]

Table 5 presents comparisons of average wages in the two data sources, broken out by socio-demographic categories. These summary statistics are calculated for all individuals with PIKs who either report wages on the CE or can be linked to the universe of form W-2s. As with the previous comparison tables, the first two columns report unweighted averages of CE and W-2 wages respectively, while the second two columns report weighted averages, with weights equal to the CE final weights multiplied by the inverse probability of linkage. Our discussion will focus on the weighted statistics, but most patterns are consistent with the unweighted statistics.

[10] To preserve confidentiality of the data, all medians are calculated as interpolated medians, so that each statistic is supported by a sufficiently large number of observations.
[11] See Appendix Tables A1-A3.
[12] To give a rough sense of the size of this wedge, consider that the average insurance premium paid by employee was about $1200 in 2014 (AHRQ, 2014), and that the average participant in a defined contribution benefit plan contributed about 6 percent of pre-tax earnings to a 401k or equivalent plan. For someone earning the unweighted average wage in Table 5 ($46,149), these two deductions would amount to about $4000.

Table 5: Mean Individual Wages by Data Source
Universe: Individuals with Non-imputed CE Wage Responses or W-2 Wages

		Unweighted		Weighted		
	Obs	Mean Wages, CE	Mean Wages, W-2	Mean Wages, CE	Mean Wages, W-2	Net Difference
All individuals	27,914	$46,149	$44,372	44,240	42,463	1,777
Sex						
Female	14,278	37,331	34,586	36,116	33,573	2,543
Male	13,636	54,641	53,896	52,178	51,332	846
Race						
White Alone	22,477	47,539	45,847	45,511	43,789	1,722
Black Alone	3,264	33,169	30,322	32,291	29,631	2,660
AIAN Alone	136	33,046	35,256	52,540	55,983	-3,443
Asian Alone	1,382	52,687	55,381	29,589	30,250	-661
Pacific Islander Alone	157	34,991	36,644	35,851	38,333	-2,482
Two or more races	498	33,536	30,523	32,084	29,712	2,372
Age						
14-24	3,923	13,847	10,814	13,562	10,602	2,960
25-44	7,078	47,882	45,183	46,303	43,495	2,808
45-64	7,911	56,102	57,553	54,273	55,981	-1,708
65+	4,434	39,442	34,905	37,444	33,719	3,725
Hispanic Origin						
Hispanic	4,561	31,450	29,989	30,852	29,481	1,371
Not Hispanic	23,353	48,759	46,663	46,604	44,558	2,046
Income						
Above Median CU Wages	14,760	59,556	57,322	57,441	54,985	2,456
Below Median CU Wages	13,154	17,943	20,798	17,710	20,347	-2,637
Second Interview Month						
1	1,917	47,461	44,611	44,772	41,769	3,003
2	2,120	43,966	44,899	42,362	43,518	-1,156
3	2,087	43,302	44,318	41,833	42,700	-867
4	2,124	49,258	44,877	47,377	43,272	4,105
5	2,072	46,417	43,514	44,490	41,821	2,669
6	2,145	51,515	47,582	50,499	45,831	4,668
7	2,152	42,254	43,590	40,501	41,626	-1,125
8	2,147	42,870	42,397	41,951	40,812	1,139
9	2,090	45,535	45,829	44,346	45,184	-838
10	3,094	46,369	41,677	44,468	39,915	4,553
11	3,099	47,988	45,117	46,010	43,004	3,006
12	2,867	46,108	44,857	42,558	41,590	968

Source: CE 2013-2014, IRS W-2s 2012-2013

CE wages are, on average, slightly higher than W-2 wages–for the full linked sample, average CE wages are 44,240, compared to an average W-2 wage of 42,463; this gap is relatively small compared to average wages, representing a discrepancy of 3.3 percent of CE wages. There is some variation in the difference in wages across data sources between different socio-demographic groups. Whites and Asians earn higher wages, on average, than do blacks and other races. Whites, Blacks and multi-race individuals report higher CE wages than W-2 wages while the exception of Asians, Native Americans and Pacific Islanders have higher W-2 wages than CE wages on average.

Looking across age groups, there is the hump-shaped pattern in wages with wages rising through the prime working age groups, and then falling for the elderly. The 45-64 age group report lower CE wages than W-2 wages, while the other cohorts report higher CE wages relative to W-2 wages. The largest errors, defined as the difference between CE and W-2 wages, occur for the oldest and youngest cohorts, both in absolute and relative terms. The difference between CE and IRS wages for individuals aged 14-24 is about $2,960 on average, which is more than 20 percent of the average CE wage for this group (the difference for the oldest cohort is over $3,700, but this is only about 11 percent of their average wages). Once again, there is no consistent pattern in the difference between CE and W-2 wages across second interview months. This suggests that individuals respond to the wage question as if it referred to the previous tax year and not to the past twelve months.[13] While leveraging administrative tax data seems to be consistent with current survey responses, these responses may in fact not be working as intended, as BLS requires a twelve-month income figure to overlap with the expenditure reporting period.

The results in Table 5 present summary statistics of the unconditional distributions of CE and W-2 wages. The differences in average wages observed in Table 5 can occur for two reasons: individuals who report income from both sources may report different values, or individuals may report income on one source but not the other (and the wages they do report are systematically different from those for individuals who report both sources of wages). We explore each of these possibilities in Tables 6 and 7.

Table 6 repeats the structure of Table 5, reporting average wages from both the CE survey responses and IRS W-2 records, but restricts the sample to only individuals who report both CE and W-2 wages. For the sample of individuals with either CE or W-2 wages (Table 5), average weighted CE wages are about $1,777 higher than weighted W-2 wages. On the other hand, average weighted CE wages are about $624 higher than weighted W-2 wages for the sample of individuals who report both CE and W-2 wages (Table 6).

[13] Since nominal wages tend to rise over time, we might expect a systematic wedge between "past twelve months" and previous tax year income in later months. Consider two individuals responding in January and December 2014. If both saw increasing wages from 2013-2014 and provided their twelve month true wages, we would expect a higher discrepancy for the individual responding in December (who would have earned 11 months of 2014 salary and 1 month of 2013 salary) than in January (who would have earned 1 month of 2014 salary and 11 months of 2013 salary).

Table 6: Comparing Mean Individual Wages by Data Source
Universe: Nonimputed CE Responses, All Individuals With Both CE and W-2 Wages

		Unweighted		Weighted		
	Obs	Mean Wages, CE	Mean Wages, W-2	Mean Wages, CE	Mean Wages, W-2	Net Difference
All individuals	9,124	$48,328	$47,942	46,397	45,773	624
Sex						
Female	4,504	39,564	38,859	38,383	37,609	774
Male	4,620	56,872	56,797	54,340	53,866	474
Race						
White Alone	7,647	49,739	49,296	47,675	46,992	683
Black Alone	833	35,031	33,675	54,648	56,091	-1,443
AIAN Alone	41	36,110	41,809	34,117	32,859	1,258
Asian Alone	442	54,266	56,389	35,236	33,582	1,654
Pacific Islander Alone	62	36,598	36,373	37,234	38,972	-1,738
Two or more races	99	37,115	35,465	33,296	35,192	-1,896
Age						
14-24	1,070	15,624	12,036	15,278	11,687	3,591
25-44	3,841	49,696	47,351	48,065	45,514	2,551
45-64	3,711	57,201	59,405	55,379	57,574	-2,195
65+	502	41,977	44,249	39,838	41,864	-2,026
Hispanic Origin						
Hispanic	1,259	33,819	33,603	33,269	33,024	245
Not Hispanic	7,865	50,651	50,237	48,488	47,805	683
Income						
Above Median CU Wages	6,359	61,073	59,590	58,980	57,178	1,802
Below Median CU Wages	2,765	19,019	21,152	18,857	20,814	-1,957
Second Interview Month						
1	667	48,015	48,962	45,338	45,349	-11
2	744	45,837	45,532	44,771	44,125	646
3	722	45,611	46,129	44,111	44,639	-528
4	704	51,693	50,178	49,698	49,227	471
5	683	48,177	47,785	46,221	45,932	289
6	721	53,005	53,849	51,847	51,857	-10
7	682	44,048	45,187	42,211	42,629	-418
8	684	45,468	44,878	44,133	43,201	932
9	660	48,490	48,662	47,455	48,672	-1,217
10	994	48,816	45,711	46,876	43,933	2,943
11	940	50,644	50,691	48,949	47,592	1,357
12	923	48,865	47,754	44,865	43,554	1,311

Source: CE 2013-2014, IRS W-2s 2012-2013

Since the differences between average CE wages and average W-2 wages are smaller for the sample of individuals who report both CE and W-2 wages, it is unlikely that misreporting by these individuals is the dominant source of the CE-W-2 differences in the unconditional averages in Table 5. Table 7 examines another possibility: that individuals who report CE wages but not W-2 wages (or vice-versa) are systematically different from the rest of the sample. The first two columns of Table 7 report average W-2 wages and the proportion of individuals who report no CE wages, but can be linked to the universe of form W-2s. Likewise, the final two columns report average wages and the proportion of individuals who report CE wages but cannot be linked to the universe of W-2s. Note that the proportions in Table 7 differ slightly from Table 4, because Table 7 only includes non-imputed wage amounts, while Table 4 reports percentages who reported wage receipt.

Approximately 20 percent of individuals with PIKs can be linked to the universe of W-2s but report no CE wages, while about 5 percent of individuals report CE wages but cannot be linked to the universe of W-2s. Both of these groups have lower wages, on average, than the overall sample reported in Table 5 or in the sample of dual-reporters in Table 6. Individuals who report CE wages but not W-2 wages have lower wages, on average, than individuals who report W-2 wages but not CE wages. This is consistent with the differences between CE and W-2 wages reported in Table 5 being largely a result of item non-response by individuals who have lower than average wages.[14]

Tables 5-7 summarize differences between reported CE and W-2 wages at the mean of the wage distribution. While there appear to be some substantial differences between the two data sources, the average difference could arise from any number of patterns across the distribution. To further examine how reported CE wages and W-2 wages compare across the entire wage distribution, we examine the entire distributions graphically. Figures 1 and 2 show two visualizations which compare these two distributions. In Figure 1, kernel density plots for the CE and W-2 wage distributions are overlaid. For a smooth distribution (e.g. wages) kernel density plots provide a more informative visualization than do histograms. Figure 2 depicts the average difference between CE and W-2 wages by percentiles of the W-2 wage distribution, among individuals reporting wages from both sources.

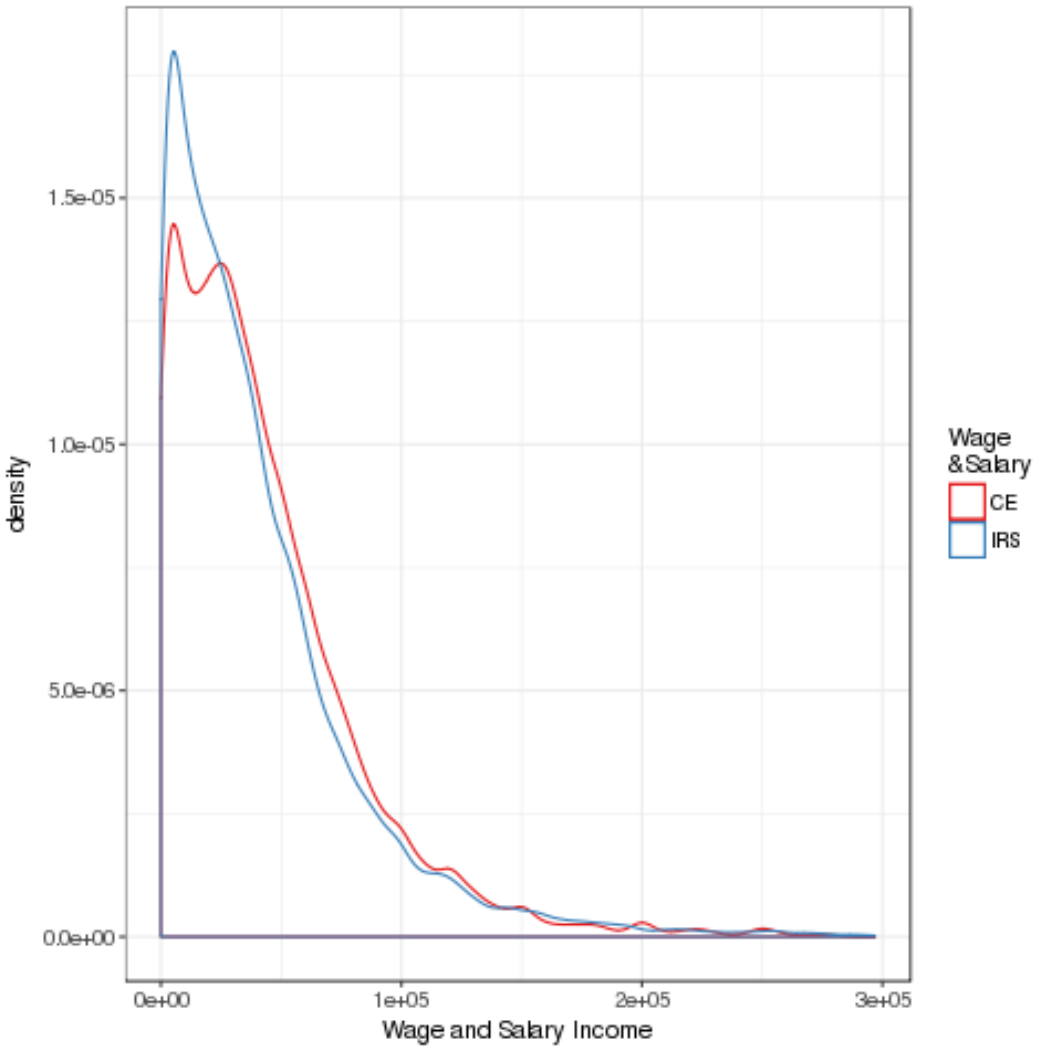
[14] Note that this would not directly affect the breakouts shown for the "Above/Below Median CU Wage" categories in Table 5, since these categories are based on survey wages

Table 7: Mean Wages for Respondents who Reported Wages in only one Data Source
Universe: Nonimputed CE Wages or W-2 Wages

	Mean W-2 Wages, Observations with no CE Wages	Proportion with W-2 but no CE Wages	Mean CE Wages, Observations with no W-2 Wages	Proportion with CE but no W-2 Wages
All individuals	$37,975	0.18	$30,304	0.04
Sex				
Female	26,911	0.18	20,227	0.04
Male	48,710	0.19	39,186	0.05
Race				
White Alone	39,213	0.18	31,038	0.05
Black Alone	26,339	0.21	21,051	0.04
AIAN Alone	23,576	0.17	7,920	0.04
Asian Alone	53,855	0.21	42,854	0.05
Pacific Islander Alone	37,375	0.15	18,383	0.04
Two or more races	24,086	0.15	19,363	0.05
Age				
14-24	9,370	0.23	7,615	0.08
25-44	40,417	0.25	30,670	0.06
45-64	53,816	0.23	46,660	0.05
65+	26,514	0.13	27,874	0.02
Hispanic Origin				
Hispanic	23,432	0.15	21,702	0.07
Not Hispanic	40,271	0.19	33,076	0.04
Income				
Above Median CU Wages	52,200	0.19	45,287	0.05
Below Median CU Wages	20,368	0.17	12,808	0.04
Second Interview Month				
1	34,970	0.16	43,490	0.05
2	35,082	0.15	27,374	0.05
3	36,597	0.17	25,080	0.05
4	40,065	0.19	26,223	0.05
5	43,432	0.17	31,859	0.05
6	40,635	0.17	25,755	0.04
7	35,405	0.19	32,453	0.04
8	35,083	0.19	33,626	0.05
9	35,436	0.19	40,087	0.06
10	40,965	0.2	27,153	0.04
11	38,228	0.2	25,105	0.04
12	41,143	0.19	28,722	0.04

Source: CE 2013-2014, IRS W-2s 2012-2013

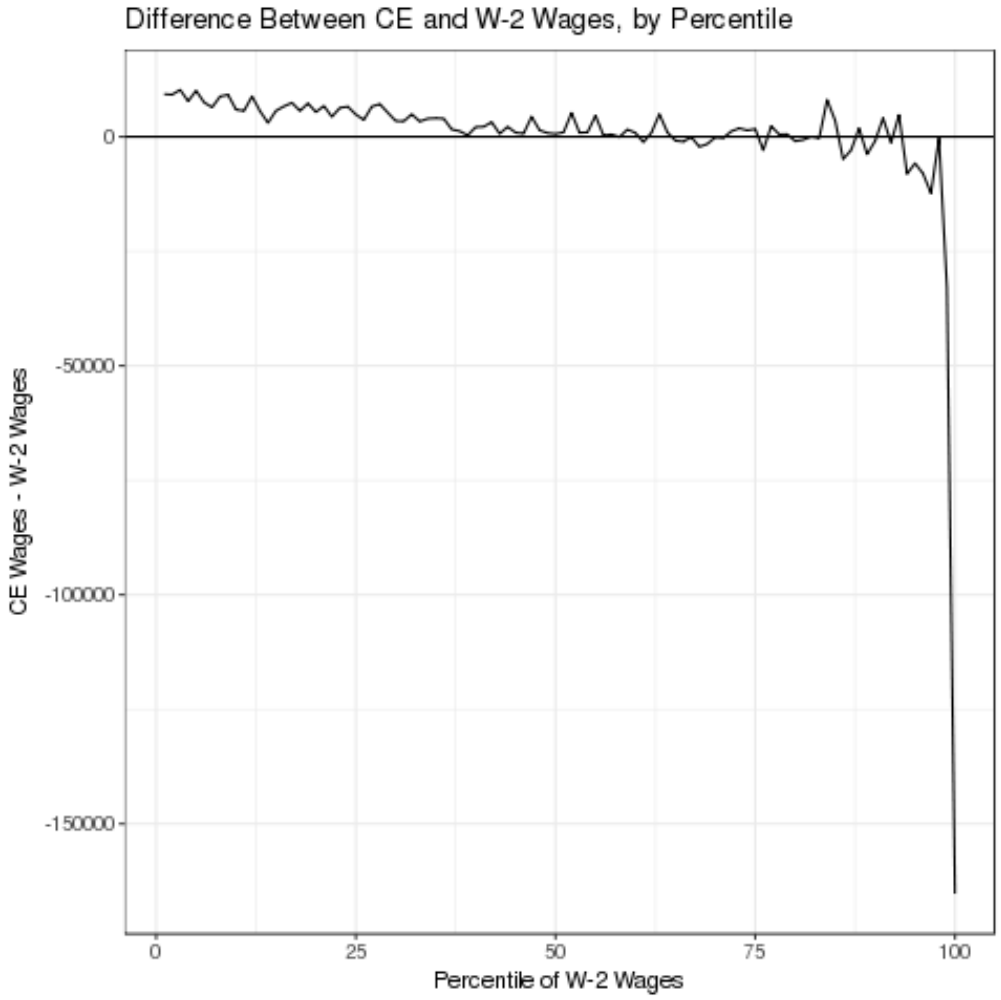
Figure 1: Kernel Density Plot of CE vs. W-2 Wages



Source: CE 2013-2014, IRS W-2s 2012-2013

Two patterns are immediately apparent from these visualizations. First, the kernel density clearly show excess mass at low wage amounts for the W-2 wage distribution relative to the CE wages: this is consistent with there being a large number of individuals who have W-2 but do not report CE wages and who have relatively low W-2 wages. Second, among people who report wages from both sources, there is some evidence of over-reporting CE wages at the bottom of the (W-2) wage distribution, and substantial under-reporting at the very top of the wage distribution.

Figure 2: Average Difference between CE and W-2 Wages by Percentile of the W-2 Wage Distribution



Source: CE 2013-2014, IRS W-2s 2012-2013

4.3 Imputed Wage Amounts

The previous analysis comparing reported CE wage amounts to wage amounts from the linked W-2 sample showed that, although the two sources have similar distributions, there is evidence of disparities, particularly at the bottom of the wage distribution. Some of these disparities could be related to item nonresponse. We next compare the two wage distributions, including imputed wage values, to examine how item non-response and imputation process might affect the previous results.

For individuals who report receiving wages but do not report a wage amount, BLS allocates a value using a multiple imputation method similar to other demographic surveys such as the Current Population Survey. This imputation process uses other information reported by the respondent (such as age, sex and whether they report a wage bracket) to impute another similar response for the item nonresponse. For wages, the variable SALARYXM contains both the raw responses in SALARYX and imputed responses for individuals who did not report a wage amount.

Table 8 reports average CE wages (including imputed and non-imputed responses) compared with average W-2 wages. As with previous tables, the second and third columns are unweighted averages, while the final two columns report weighted averages. Note that Columns 3 and 5 are identical to Columns 3 and 5 from Table 5 since both tables report unconditional averages of the CE and W-2 wage distributions.

Overall, the distribution of CE wages including imputed values is closer to the distribution of W-2 wages, at least on average. This suggests that the imputation process is reducing the discrepancy between survey and administrative records, at least in the middle of the wage distribution. There is some heterogeneity across some demographic groups, however. The gap between CE and W-2 wages is larger for men when including imputed values relative to non-imputed wages and all race groups have smaller gaps between CE and W-2 wages when including imputed values except for Native Americans and Pacific Islanders. Individuals ages 25-44 have gaps between CE and W-2 wages that are much smaller when including imputed values; on the other hand, the youngest cohort has close to no change, and the cohort aged 45-64 actually has a larger gap on average when including imputations.

As with the analysis of the reported CE wage values, one source of difference between the CE and W-2 wage distributions may arise from individuals who report CE wages but cannot be linked to the universe of W-2s, or vice versa. To investigate this possibility, Table 9 reports average wages for individuals who report CE wages but not W-2 wages and vice versa, as well as the proportion of individuals who report wages in one but not the other dataset. The structure of Table 9 mirrors that of Table 7 above.

Table 8: Mean Individual Wages by Data Source
Universe: All Responses with CE wages (imputed and non-imputed responses) or W-2 Wages

		Unweighted		Weighted		
	Obs.	Mean Wages, CE	Mean Wages, W-2	Mean Wages, CE	Mean Wages, W-2	Net Difference
All individuals	27,914	$43,458	$44,372	41,753	42,463	-710
Sex						
Female	14,278	36,562	34,586	35,545	33,573	1,972
Male	13,636	49,858	53,896	47,627	51,332	-3,705
Race						
White Alone	22,477	45,043	45,847	43,268	43,789	-521
Black Alone	3,264	30,461	30,322	29,721	29,631	90
AIAN Alone	136	36,359	35,256	52,479	55,983	-3,504
Asian Alone	1,382	52,194	55,381	30,035	30,250	-215
Pacific Islander Alone	157	31,006	36,644	31,910	38,333	-6,423
Two or more races	498	32,419	30,523	31,162	29,712	1,450
Age						
14-24	3,923	13,725	10,814	13,288	10,602	2,686
25-44	7,078	45,214	45,183	43,911	43,495	416
45-64	7,911	53,704	57,553	52,299	55,981	-3,682
65+	4,434	36,862	34,905	35,429	33,719	1,710
Hispanic Origin						
Hispanic	4,561	29,648	29,989	29,179	29,481	-302
Not Hispanic	23,353	46,118	46,663	44,235	44,558	-323
Income						
Above Median CU Wages	14,760	56,459	57,322	54,503	54,985	-482
Below Median CU Wages	13,154	18,018	20,798	17,709	20,347	-2,638
Second Interview Month						
1	1,917	44,254	44,611	42,152	41,769	383
2	2,120	41,561	44,899	40,854	43,518	-2,664
3	2,087	42,583	44,318	40,327	42,700	-2,373
4	2,124	44,299	44,877	42,382	43,272	-890
5	2,072	45,521	43,514	43,857	41,821	2,036
6	2,145	47,726	47,582	46,849	45,831	1,018
7	2,152	40,201	43,590	38,340	41,626	-3,286
8	2,147	40,994	42,397	40,281	40,812	-531
9	2,090	43,368	45,829	42,483	45,184	-2,701
10	3,094	43,526	41,677	41,252	39,915	1,337
11	3,099	43,870	45,117	42,217	43,004	-787
12	2,867	43,576	44,857	40,782	41,590	-808

Source: CE 2013-2014, IRS W-2s 2012-2013

Table 9: Mean Wages for Respondents who Reported Wages in only one Data Source
Universe: All Responses with CE wages (imputed and non-imputed responses) or W-2 wages

	Mean W-2 Wages, Observations with no CE Wages	Proportion with W-2 but no CE Wages	Mean CE Wages, Observations with no W-2 Wages	Proportion with CE but no W-2 Wages
All individuals	$16,897	0.05	$37,199	0.19
Sex				
Female	13,254	0.06	34,474	0.17
Male	21,171	0.05	39,578	0.21
Race				
White Alone	18,228	0.05	39,034	0.19
Black Alone	10,214	0.06	24,100	0.2
AIAN Alone	22,276	0.06	34,112	0.19
Asian Alone	19,278	0.05	47,177	0.18
Pacific Islander Alone	7,175	0.04	27,676	0.22
Two or more races	6,878	0.06	30,441	0.35
Age				
14-24	11,487	0.09	8,651	0.13
25-44	16,061	0.05	30,464	0.1
45-64	27,682	0.06	42,925	0.1
65+	15,281	0.08	28,118	0.07
Hispanic Origin				
Hispanic	12,636	0.05	25,504	0.26
Not Hispanic	17,736	0.05	40,637	0.18
Income				
Above Median CU Wages	22,958	0.03	51,336	0.21
Below Median CU Wages	14,417	0.08	16,010	0.16
Second Interview Month				
1	11,487	0.04	40,345	0.19
2	19,025	0.05	35,327	0.21
3	14,565	0.05	36,494	0.19
4	14,272	0.05	34,701	0.2
5	12,991	0.05	34,737	0.18
6	16,043	0.05	39,948	0.2
7	30,075	0.05	33,976	0.19
8	17,971	0.05	44,531	0.21
9	11,134	0.05	42,615	0.21
10	16,300	0.06	36,554	0.17
11	19,304	0.06	34,906	0.18
12	16,884	0.07	34,888	0.17

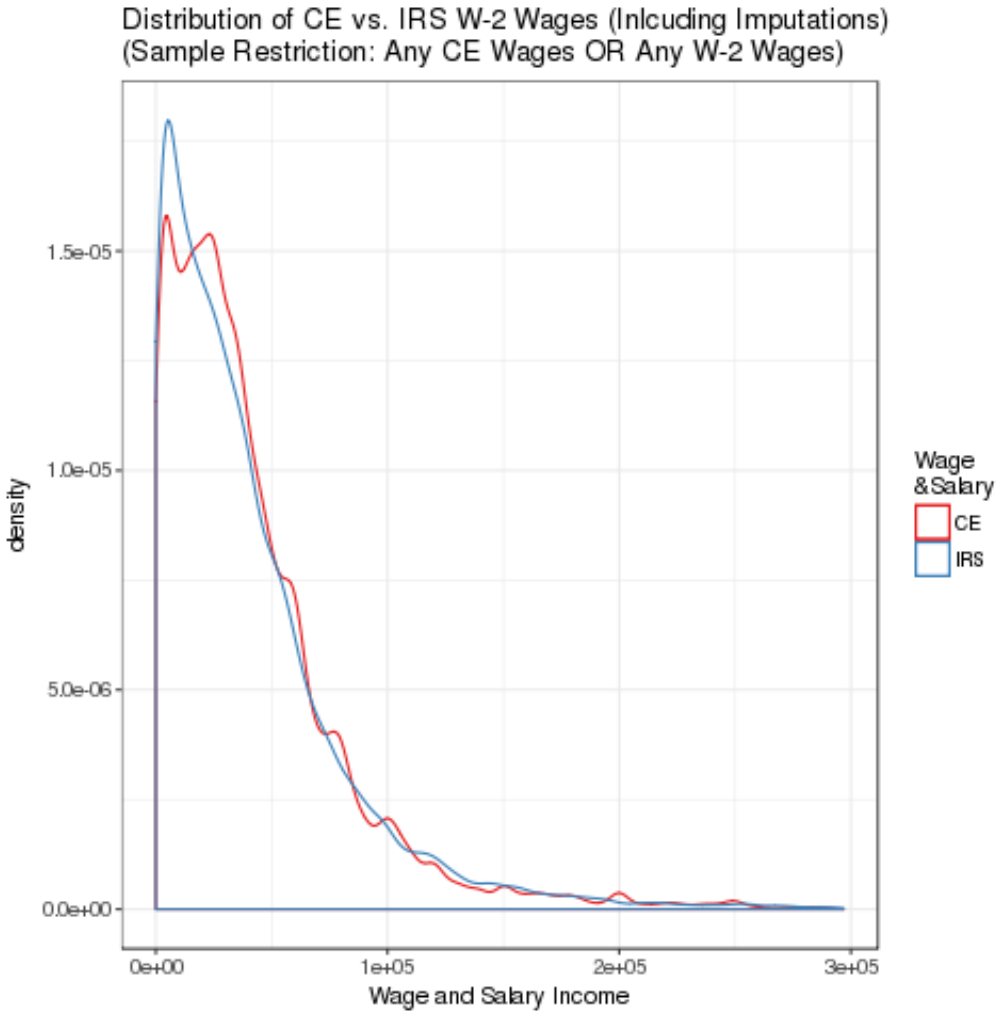
Source: CE 2013-2014, IRS W-2s 2012-2013

There are two notable features of this set of averages. First, the proportion of individuals with W-2 but no CE wages is substantially smaller when including imputed CE wages (Column 2). About 18 percent of individuals had W-2 wages but did not report CE wages (See Table 7), while about five percent of individuals have W-2 wages but neither report CE wages nor receive an imputed wage amount (see Table 9). However, there are many more individuals with CE wages but no W-2 wages when including imputations—about 19 percent of individuals either report CE wages or receive an imputed value but do not have W-2 wages (see Table 9), compared to about 4 percent who report CE wages but do not have W-2 wages (see Table 7).

Individuals who have W-2 but not CE wages when including imputations are substantially poorer than the set of individuals who report no CE wages but W-2 wages (not including imputations) ($16,897 vs. $37,975); individuals who have CE wages but not W-2 wages are slightly richer on average when including imputations, however ($37,199 vs. $30,304). Among age groups in Table 9, the youngest cohort is both most likely to have W-2 but not CE wages (and vice versa), and the poorest. Including imputations results in a higher average wage for this age group for the sample who have W-2 but not CE wages, and a higher average wage for individuals who have CE wages but not W-2 wages.

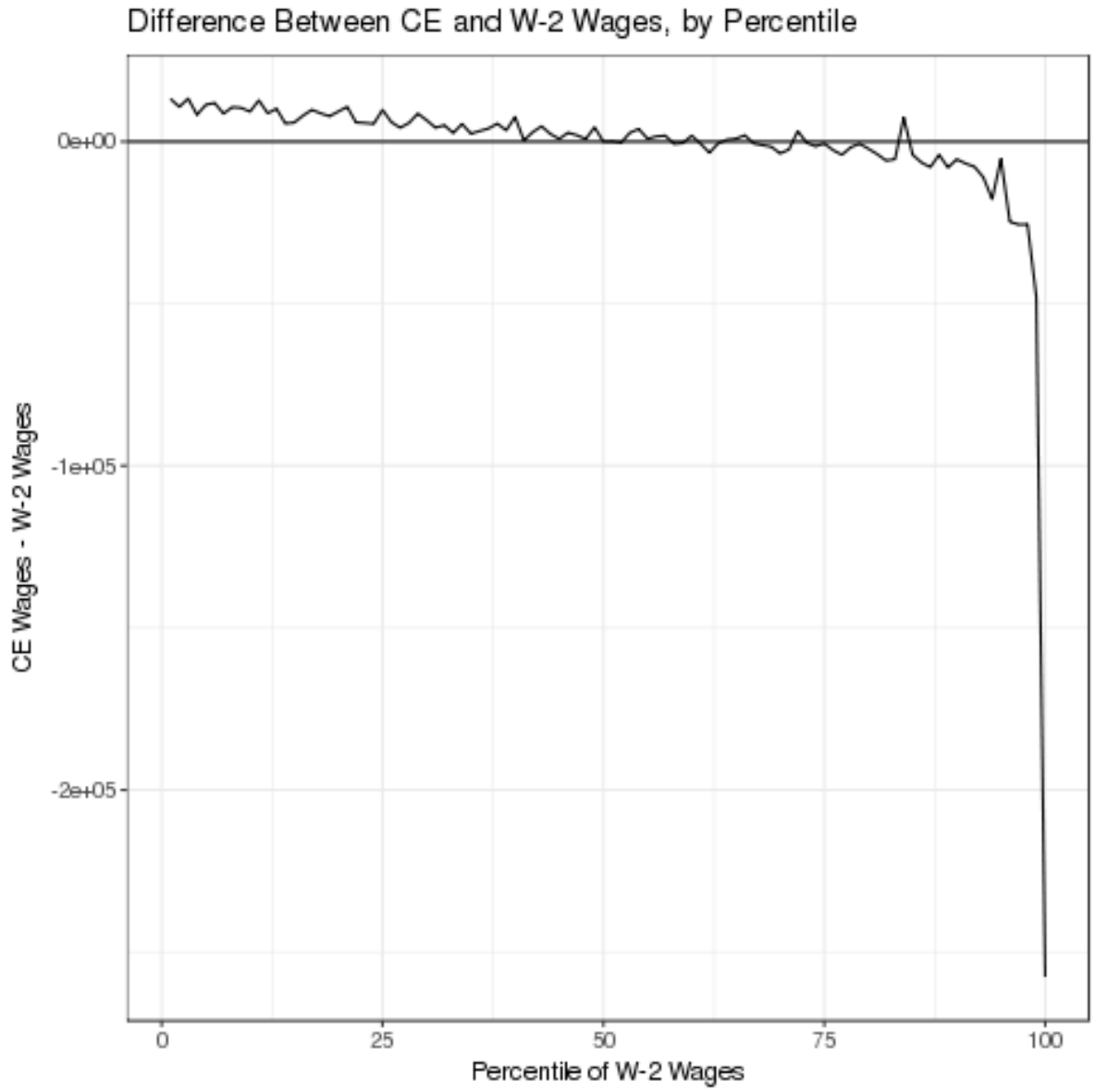
Looking beyond the mean, even when including imputations, there is still a consistent pattern of substantial differences between the two wage distributions at the bottom of the wage distribution, and some evidence of under-reporting at the very top of the wage distribution. Figures 3 and 4 produce two visualizations to illustrate this, analogous to Figures 1 and 2 above. Figure 3 presents a kernel density plot of the CE and W-2 wage distributions when including imputations, which shows excess mass for the W-2 wage distribution at very low wages. In Figure 4, a plot of the average difference between CE and W-2 wages by percentile of the W-2 wage distribution shows substantial over-reporting of wages on the CE at the bottom of the distribution, and very large under-reporting for the top one percent of the wage distribution. Note that the discrepancies between survey and administrative records continue to be present when including imputed survey wage responses. This suggests that while the imputation process may reduce discrepancies in the middle of the wage distribution, it is not doing do at the tails. However, insofar as discrepancies in the tails are due to misreporting, the imputation process would not necessarily be expected to reduce these discrepancies.

Figure 3: Kernel Density Plot of CE vs. W-2 Wages, Including Imputed and Non-imputed CE wages



Source: CE 2013-2014, IRS W-2s 2012-2013

Figure 4: Average Difference between CE and W-2 Wages by Percentile of the W-2 Wage Distribution, Including Imputed and Non-imputed CE wages



Source: CE 2013-2014, IRS W-2s 2012-2013

4.4 Bracketed Wage Amounts

The CE wage questions consist of a three-part process. Respondents are first asked whether they receive any wages. Those who indicate that they do receive wages are then asked how much they earned in the past twelve months. Finally, respondents who do not know the exact amount are asked to select one of eleven brackets in which their wages fall. We have analyzed the first two

of these wage questions in the previous two subsections.[15] We complete our wage analysis by comparing these bracketed CE wage responses to the W-2 records.

We assign individuals to CE wage brackets as follows. If an individual reported a wage bracket (this is reported in the CE variable SALARYB), we assign that wage bracket. If an individual did not report a wage bracket, but did report a (non-imputed) wage amount in SALARYX, we assign the bracket in which this wage amount falls. To compare this hybrid bracketed series to the W-2 wage data, we calculate in which CE bracket an individual's W-2 wages would fall for all individuals who have both a CE bracketed wage amount and W-2 wages. Table 10 reports agreement between these two bracketed wage series.

Table 10: Agreement Between CE and IRS Wage Income By CE Bracket

		W-2 Brackets ($)										
		[0-5K)	[5K-10K)	[10K-15K)	[15K-20K)	[20K-30K)	[30K-40K)	[40K-50K)	[50K-70K)	[70K-90K)	[90K-120K)	120K+
CE Brackets ($)	[0-5K)		173	79	46	66	38	27	15	*	*	*
	[5K-10K)	232		132	64	60	11	10	14	*	*	*
	[10K-15K)	134	169		116	80	44	19	16	*	*	*
	[15K-20K)	82	126	166		181	58	36	24	13	*	*
	[20K-30K)	97	108	163	261		237	87	80	32	11	12
	[30K-40K)	71	51	89	99	403		183	109	28	29	15
	[40K-50K)	31	24	34	42	141	328		204	48	26	22
	[50K-70K)	24	24	29	36	80	154	351		157	79	56
	[70K-90K)	13	*	13	*	17	23	41	273		98	42
	[90K-120K)	*	*	*	*	11	11	11	42	143		114
	120K+	*	*	*	*	*	*	*	20	30	100	

Source: CE 2013-2014, IRS W-2s 2012-2013
Note: * denotes a cell which has been suppressed to preserve confidentiality.

[15] Note that respondents who report a bracketed wage amount are assigned an imputation within this bracket, so the results in section 4.3 indirectly analyzed these bracketed responses.

For all CE wage brackets, the modal case is one of agreement, in which the CE wage brackets and W-2 wage brackets align. For most interior brackets (i.e., not including the top and bottom brackets) there are slightly fewer underestimates than over-estimates–the number of individuals with CE wages in a lower bracket is smaller than the number of individuals with CE wages in a higher bracket, for any given W-2 wage bracket. This is consistent with the previous results, which suggests over-reporting of CE wages in the bottom half of the wage distribution.

Table 11 provides an alternate way of comparing the two bracketed wage amounts. As in Table 10, individuals are ordered by their CE wages. Within these brackets, we report the average CE wages, and the average W-2 wages, both unweighted (first two columns) and weighted by CE final weights multiplied by the inverse probability of linkage (final two columns), as above. Note that this shows a different gap than in Figures 1 and 2 because this shows averages within a bracket defined by survey wages. When fixing CE wage brackets, W-2 wages are higher, on average, than CE wages for the first six brackets (and marginally higher in the next-to-highest bracket). In fact, the average W-2 wages are above the upper bracket bound for the first four brackets. This reflects differences between the two data sources within each CE bracket in the upper tails: Table A4, in the Appendix, reports similar results for median wages within each bracket, and shows no evidence of median matched W-2 wages outside of the CE brackets.

Table 11: Mean Income by Data Source, by CE Income Bracket

		Unweighted		Weighted		
		Mean CE Wages	Mean W-2 Wage	Mean CE Wages	Mean W-2 Wage	Net Difference
CE Brackets	[0-5K)	$1,992	$10,511	$1,989	$10,085	-8,096
	[5K-10K)	7,162	11,548	7,128	10,948	-3,820
	[10K-15K)	12,179	15,262	12,154	15,053	-2,899
	[15K-20K)	17,101	20,531	17,082	20,352	-3,270
	[20K-30K)	24,439	26,337	24,457	25,969	-1,512
	[30K-40K)	34,175	35,436	34,188	34,841	-653
	[40K-50K)	43,956	43,711	43,950	43,522	428
	[50K-70K)	57,974	57,907	57,962	57,689	273
	[70K-90K)	77,891	75,760	77,909	75,308	2,601
	[90K-120K)	101,219	101,456	101,113	100,451	662
	120K+	197,471	190,140	196,537	185,603	10,934

Source: CE 2013-2014, IRS W-2s 2012-2013

4.5 Social Security Deductions

We now turn to several other CE questions which can be directly compared to IRS administrative records. The first of these concerns Social Security (and Medicare) payroll deductions. The exact CE question is "Was there any money deducted from your pay for Social Security, including Medicare?"

Individuals who earn wages must pay Federal Insurance Contribution Act (FICA) taxes unless they are one of several exempted categories, including some full time students, and some types of employees who are covered under an employer-provided retirement plan (and hence are excluded from the Social Security (SS) system). Most non-exempt wage earners have their FICA taxes deducted from their paychecks. Form W-2, which is filed for almost all wage earners, reports the amount of FICA-eligible wages, as well as the amount of FICA taxes deducted from an individual's wages. The W-2 extracts which CARRA receives from the IRS contain only the amount of FICA-eligible wages reported on a W-2 form. Thus, we will define an individual as having W-2 SS deductions if they can be linked to a W-2 form with nonzero FICA eligible wages.[16]

Table 12 compares the proportion of individuals who report having any SS or Medicare payroll deductions in the CE to the proportion who have SS payroll deductions in the W-2s, broken out by various sociodemographic categories. As with previous tables, the first two columns report unweighted proportions, while the final two columns report proportions weighted by CE final weights and inverse probability weights. The denominator in both CE and W-2 proportions is the total number of CE respondents.

Overall, a substantially larger proportion of individuals have SS payroll deductions in the W-2 data than report having SS payroll deductions on the CE. The difference in proportions between the two data sources is about 10 percentage points for the total CE sample, but some demographic groups have substantially larger or smaller differences between CE and W-2 payroll deduction proportions. Black and AIAN respondents both have larger gaps (about 14 and 14.4 percentage points, respectively) than do other racial groups, and respondents from below median income CUs have larger gaps than do higher income individuals.

[16] Note that the CE question asks about both SS and Medicare deductions, but some individuals in public pension plans may be exempt from SS payroll taxes (hence have no FICA wages), but still will be required to pay Medicare payroll taxes.

Table 12: Any SS Payroll Deductions Reported by Data Source

		Unweighted		Weighted		
		SS Deductions, CE	SS Deductions, W-2	SS Deductions, CE	SS Deductions, W-2	Net Difference
All individuals	27,914	0.382	0.483	0.378	0.478	-0.100
			Sex			
Female	14,278	0.359	0.463	0.357	0.462	-0.105
Male	13,636	0.407	0.504	0.400	0.494	-0.094
			Race			
White Alone	22,477	0.394	0.49	0.392	0.486	-0.094
Black Alone	3,264	0.31	0.451	0.307	0.447	-0.140
AIAN Alone	136	0.301	0.449	0.288	0.432	-0.144
Asian Alone	1,382	0.412	0.502	0.399	0.485	-0.086
Pacific Islander Alone	157	0.433	0.529	0.416	0.505	-0.089
Two or more races	498	0.251	0.331	0.225	0.301	-0.076
			Age			
14-24	3,923	0.335	0.485	0.333	0.481	-0.148
25-44	7,078	0.623	0.756	0.624	0.759	-0.135
45-64	7,911	0.549	0.66	0.551	0.661	-0.110
65+	4,434	0.135	0.219	0.134	0.218	-0.084
			Hispanic Origin			
Hispanic	4,561	0.323	0.408	0.311	0.396	-0.085
Not Hispanic	23,353	0.394	0.498	0.391	0.494	-0.103
			Income			
Above Median CU Wages	14,760	0.507	0.59	0.495	0.577	-0.082
Below Median CU Wages	13,154	0.242	0.363	0.247	0.366	-0.119
			Second Interview Month			
1	1,917	0.386	0.476	0.387	0.473	-0.086
2	2,120	0.395	0.48	0.393	0.474	-0.081
3	2,087	0.387	0.493	0.384	0.488	-0.104
4	2,124	0.381	0.483	0.364	0.474	-0.110
5	2,072	0.36	0.477	0.352	0.473	-0.121
6	2,145	0.379	0.478	0.373	0.469	-0.096
7	2,152	0.376	0.49	0.371	0.487	-0.116
8	2,147	0.375	0.484	0.373	0.480	-0.107
9	2,090	0.388	0.483	0.385	0.483	-0.098
10	3,094	0.399	0.49	0.398	0.487	-0.089
11	3,099	0.373	0.476	0.364	0.458	-0.094
12	2,867	0.385	0.488	0.384	0.487	-0.103

Source: CE 2013-2014, IRS W-2s 2012-2013

4.6 Self-Employment Income Receipt

We next examine how self-employment income receipt varies between CE responses and another source of administrative records - the universe of Form 1040 tax returns. Although CE questions ask both about receipt of self-employment income and the amount of self-employment income received, the Form 1040 data sent to CARRA by the IRS does not contain the amount of self-employment income received by a tax unit. CARRA's Form 1040 data extract does contain flags for whether schedule C (reporting wages from self-employment or a sole proprietorship) was filed by a tax unit. We will therefore focus our analysis of the proportion of CE respondents who are in CUs which report any self-employment income compared to the proportion of individuals who can be linked to a Form 1040 that included a Schedule C filing.[17]

Table 13 shows that the differences in the proportions of individuals with some self-employment income are quite small. Overall, about 16 percent of individuals are in CUs which have CE self-employment income, compared to about 15 percent who appear on Form 1040 with self-employment income. There are a few substantial differences when stratifying by race (blacks are more likely to have 1040 self-employment income than CE self-employment income, unlike other race groups) and by age (the youngest cohort is substantially less likely to be on a Form 1040 with self-employment income than to be in a CU with self-employment income). The stark results for the youngest cohort may be a result of the differing income receiving units in the way we define self-employment receipt. College students who live away from home for most of the year may be considered part of their parents' CU for CE purposes but may file a tax return independently of their parents for 1040 purposes. Since college students are less likely to file a Schedule C than their parents, family-level self-employment income would be present in the CE but not in the college student's 1040.

[17] Note that while Schedule C represents the most common type of self-employment income declared on Form 1040, there are other types of self-employment income; the IRS 1040 proportions are thus likely under-estimates of all forms of self-employment.

Table 13: Any Self-employment or Sole Proprietorship Income Reported by Data Source

	Obs.	Unweighted		Weighted		
		Any SE Income, CE	Any SE Income, IRS 1040	Any SE Income, CE	Any SE Income, IRS 1040	Net Difference
Full Sample	27,914	0.157	0.154	0.156	0.150	0.006
Sex						
Female	14,278	0.147	0.153	0.146	0.151	-0.005
Male	13,636	0.168	0.155	0.166	0.149	0.017
Race						
White Alone	22,477	0.167	0.158	0.166	0.155	0.011
Black Alone	3,264	0.094	0.135	0.097	0.131	-0.034
Asian Alone	1,382	0.167	0.156	0.156	0.152	0.004
AIAN Alone	136	0.154	0.14	0.191	0.122	0.069
Pacific Islander Alone	157	0.089	0.102	0.099	0.117	-0.018
Two or more races	498	0.131	0.09	0.141	0.085	0.056
Age						
14-24	3,923	0.166	0.026	0.170	0.054	0.116
25-44	7,078	0.149	0.233	0.148	0.232	-0.084
45-64	7,911	0.19	0.23	0.189	0.233	-0.044
65+	4,434	0.093	0.138	0.093	0.138	-0.045
Hispanic Origin						
Hispanic	4,561	0.135	0.134	0.132	0.128	0.004
Not Hispanic	23,353	0.161	0.158	0.161	0.154	0.007
Income						
Above Median CU Wages	14,760	0.178	0.168	0.175	0.164	0.011
Below Median CU Wages	13,154	0.134	0.138	0.134	0.134	0.000
Second Interview Month						
1	1,917	0.169	0.142	0.163	0.135	0.028
2	2,120	0.151	0.145	0.151	0.142	0.009
3	2,087	0.132	0.149	0.128	0.149	-0.021
4	2,124	0.172	0.174	0.179	0.163	0.016
5	2,072	0.178	0.156	0.179	0.153	0.026
6	2,145	0.176	0.14	0.183	0.136	0.047
7	2,152	0.155	0.162	0.161	0.160	0.001
8	2,147	0.148	0.142	0.144	0.134	0.010
9	2,090	0.185	0.162	0.176	0.157	0.019
10	3,094	0.164	0.165	0.160	0.162	-0.002
11	3,099	0.138	0.148	0.135	0.148	-0.013
12	2,867	0.131	0.157	0.133	0.151	-0.018

Source: CE 2013-2014, IRS 1040s 2012-2013

4.7 Retirement Income

Income from employer-sponsored retirement plans (both defined benefit and defined contribution) and Individual Retirement Accounts (IRAs) are an important income source for older Americans. Income from these sources is accurately captured in administrative tax records (due to regulations on pension funds and the tax-favored status of IRAs), but recent research has suggested that surveys may not accurately capture either the prevalence or amount of retirement income received by older Americans (Bee and Mitchell, 2016, 2017).

We add to this recent literature by comparing CE responses to questions about retirement income with administrative records on retirement fund distributions reported on IRS Form 1099-R. The CE contains two questions about retirement income, asked for the CU as a whole, first asking "did you or any member of your household receive any retirement, survivor, or disability pensions?", and then a follow up asking for the amount of this income for CU's who indicate retirement income receipt.[18] IRS Form 1099-R is sent to all recipients of retirement income and to the IRS by all retirement plans which made distributions in a given tax year. The extracts received by CARRA from the IRS contain fields detailing the amount of the distribution, and whether the distribution was from an employer-sponsored retirement plan or an individual retirement account. Since the CE questions are asked at the CU level, but the 1099-R are issued to individuals, we conduct our analysis at the household level and categorize households by the demographic characteristics of the respondent.

Table 14 summarizes a comparison between the CE and administrative records for receipt of any retirement income, broken out by select demographic categories. Because of the small implicit cell sizes for smaller race groups, we collapse AIAN, Pacific Islanders and two or more races into an "Other Race" category, and due to the nature of the income source, we omit the overall average, instead focusing on receipt by the full retirement age population aged 65 and older. The second column of Table 14 reports the proportion of households who report receiving retirement income on the CE. The final three columns report the proportion of households who receive retirement income in the universe of Form 1099-Rs – Column 3 reports the proportion with any

[18] Note that the instructions to field representatives clarify that retirement income should be defined as "Regular income from annuities and IRA or KEOGH retirement plans", which is a narrower definition of retirement income than the definition captured by the 1099-R universe (which is the universe of all disbursements, whether regular or not.) It is debatable whether retirement income should only be counted only if it is regular given almost no one who has an IRA or 401k uses them in this manner. The CPS ASEC was redesigned in 2014 to add questions about retirement account withdrawals reflecting the changing retirement landscape. Further, as shown in Bee and Mitchell (2017), the underreporting of retirement income is not limited to IRAs and 401ks, but rather defined benefit income is also missing. Even if one wants to stick with the regular income concept the CPS ASEC has about the same retirement income rate of receipt for *persons* 65+ (37 percent) that the CE has for anyone in the household among householders 65+. If one looks at household level receipt in the CPS, receipt is closer to 47 percent. In other words, relative to the 2013 (pre-redesign) CPS ASEC, the CE is still very much underreporting retirement income receipt. It is possible that the respondent, when reporting retirement income, is only reporting his or her own receipt and not counting spouses retirement income. See Bee and Mitchell (2017) for further details on comparing 1099-R and CPS ASEC retirement income receipt.

1099-R income, Column 4 1099-R pension (employer-sponsored) income and Column 5 1099-R IRA income.

Across all age, sex and racial/ethnic respondent demographic groups, the proportion of households reporting retirement income on the CE is dramatically smaller than the proportion with retirement income in the administrative records. Among the over 65 population, about 39 percent of CU householders report retirement income for someone in the household, while 73 percent received a retirement distribution on the 1099-R (of these, 64 percent receive employer-sponsored income, while 34 percent receive an IRA distribution). Male and female respondent households have similar proportions of CE retirement income reporting, while male householders have a slightly higher proportion of 1099-R retirement income. Whites and non-Hispanics have higher gaps between CE and 1099-R retirement income receipt than do non-whites and Hispanics.

Table 14: Any Retirement Income for Householders by Source

	Obs.	Any CE Retirement	Any 1099-R	Any 1099-R Pension	Any 1099-R IRA
			Age of Respondent		
25-44	2207	0.014	0.114	0.085	0.038
45-54	1314	0.046	0.178	0.137	0.061
55-64	1308	0.173	0.370	0.307	0.123
65-74	975	0.369	0.687	0.605	0.268
75-84	511	0.411	0.784	0.701	0.444
85+	190	0.404	0.763	0.671	0.397
All Respondents 65+	1676	0.386	0.725	0.642	0.337
			Sex		
Female	910	0.388	0.704	0.630	0.328
Male	766	0.383	0.751	0.656	0.347
			Race		
White Alone	1408	0.403	0.746	0.655	0.375
Black Alone	173	0.316	0.657	0.609	0.107
Asian Alone	53	0.210	0.453	0.414	(D)
Other Race	44	(D)	(D)	0.476	(D)
			Hispanic Origin		
Hispanic	119	0.236	0.483	0.457	(D)
Not Hispanic	1557	0.396	0.742	0.655	0.355

Note: Other Race includes AIAN, Pacific Islander and Two or More Races
Source: CE 2013-2014, IRS 1099-R 2012-2013

Table 15 follows up the analysis of retirement income receipt with an analysis of the degree to which retirement income amounts reported on the CE are comparable to retirement income distributions from the 1099-R's. Especially for pension distributions, mean 1099-R retirement income amounts are substantially larger than the mean reported CE amounts. IRA distributions are smaller on average, than employer-sponsored pension distributions. However, as Table 14 suggests, few individuals over 65 have IRA, but not employer-sponsored pension income. Overall, average retirement income for the over 65 population is about $9,500 in the CE, but the average retirement income in the 1099-R is nearly double this, at around $18,500.

Table 15: Mean Retirement Income Amounts by Data Source

	Obs.	Any CE Retirement	Any 1099-R	Any 1099-R Pension	Any 1099-R IRA
			Age of Respondent		
25-44	2207	$3137	$3218	$2797	$3471
45-54	1314	7317	9123	8650	7147
55-64	1308	15544	17534	16004	9663
65-74	975	10457	19822	17776	8310
75-84	511	9022	18919	15843	5382
85+	190	7439	13656	9738	4455
All Respondents 65+	1676	9559	18475	16204	6352
			Sex		
Female	910	6940	15008	12460	4826
Male	766	14146	22467	20011	8998
			Race		
White Alone	1408	9587	19113	16547	6554
Black Alone	173	8346	14589	14788	2152
Asian Alone	53	16963	8287	9666	(D)
Other Race	44	(D)	(D)	5952	(D)
			Hispanic Origin		
Hispanic	119	6897	8384	8329	(D)
Not Hispanic	1557	9719	18960	16492	6400

Note: Other Race includes AIAN, Pacific Islander and Two or More Races
Source: CE 2013-2014, IRS 1040s 2012-2013

5 Survey Non-response Bias Analysis

Thus far, the analysis of CE income-related responses (wages, payroll deductions, self-employment and retirement income) has proceeded by linking individual respondents to administrative records by the use of unique anonymous identifiers (PIKs). This analysis provides insight into the measurement error properties of CE responses, and can inform ongoing efforts to

improve the imputation process used to address item non-response. This PIK-based analysis, however, is silent about survey non-response bias.

Our final set of analyses in this section will focus on survey non-response bias in two ways. First, by linking responding and non-responding households to administrative tax records by address and comparing the characteristics of responding and non-responding households, we will describe the qualities and extent of survey non-response bias. Second, we will use the same address matching process to provide a validation of an important component of the weighting process used by BLS to address bias due to survey non-response.

5.1 Comparing WSI and AGI Income

The first analysis of survey non-response bias will compare characteristics of responding and non-responding households from linked administrative records. Using survey frame information, most responding and non-responding households can be assigned a MAFID (there was about a 93 percent MAF match rate). We use these assigned MAFIDs to link these sample units to two sets of administrative records: the universe of Form 1040 tax returns (used above in the self-employment analysis) and the universe of Form W-2 wage returns (used above in the wage analysis).

This address-based linkage presents certain challenges that are less pressing concerns for the previous person-based matching. The most pressing of these challenges are the related problems of definitional differences between addresses and tax units, and potential duplicate matches for individual addresses. Because multiple tax units may reside at the same address, or may file taxes from the same address (in case of households moving), there are many more addresses with multiple matches than was the case in the previous person based match. Unfortunately, the relatively sparse demographic information available for non-responding households and on the administrative tax data makes separating cases where there are multiple matches difficult. Our approach to aggregating these cases is as follows: first, we drop any CE sample units that are linked to 10 or more unique Form 1040s, and second, we average the incomes listed on all linked 1040s for each CE sample unit. In multiple match cases, we are thus comparing average income of all the tax units at an address level. Note that this approach to aggregation will understate incomes of households with more than one tax unit (as in a household with two unmarried adults).

We will compare the distribution of two income concepts between responding and non-responding households: total wage and salary income (WSI), and adjusted gross income (AGI). WSI is the total amount of wage and salary income reported by a tax unit on a form 1040. AGI is a broader definition of income, including all gross income (wages, interest, dividends, taxable gains, etc.) but excluding “above-the-line” deductions. Table 16 summarizes a comparison of responding and non-responding households in terms of their total WSI. The second and third columns compare medians (respondents in Column 1, non-respondents in Column 2), while the final two columns compare means (respondents in Column 4, non-respondents in Column 5).

Using either the median or mean as a measure of central tendency, the typical non-responding sample unit is richer than the typical responding sample unit. The gap between responding and non-responding sample units is larger for means than for medians, which indicates larger differences in the right tail of the wage distribution. The largest gaps between responding and non-responding sample units are for Whites (the average White non-responder's WSI was nearly $55,000, compared to about $46,000 for White responders) and for Married filing jointly tax units (non-responders' average WSI was $78,000, compared to $66,000 for responders). Appendix Table A5 provides a similar set of analyses comparing aggregated address linked W-2 wages between responding and non-responding sample units, with largely similar results.

Table 16: Household Wage and Salary Income by CE Response Status
Universe: Responding and Non-responding SUs Linked to 1040s

		Median WSI		Mean WSI	
	Obs.	CE Respondent Household	CE Non-Respondent Household	CE Respondent Household	CE Non-Respondent Household
All households	21,169	$28,964	$32,102	$45,035	$52,383
Race of Householder					
White Alone	10,340	30,110	34,565	45,662	54,909
Black Alone	1,470	21,601	23,230	30,395	34,551
AIAN Alone	35	36,817	36,544	52,072	28,041
Asian Alone	555	39,822	37,712	59,724	62,571
Pacific Islander Alone	220	23,597	26,229	35,868	45,604
Two or more races	235	26,643	28,270	39,765	39,880
Unknown	8,314	29,110	31,946	45,816	53,588
IRS 1040 Filing Status					
Head of Household	2,962	21,824	23,616	28,239	30,200
Married Filing Jointly	8,984	47,595	54,874	65,553	78,437
Married Filing Separately	293	27,952	31,910	32,616	38,210
Single	8,930	22,509	24,830	30,705	34,527

Source: CE 2013-2014 and IRS 1040s, 2012-2013

Table 17 turns to a comparison of the average AGI between responding and non-responding sample units. The table structure mirrors Table 16, with a comparison of medians in the first two columns, and comparison of means in the second two columns. Consistent with the previous WSI results, non-responding sample units have substantially higher AGI than responding households, both when looking at medians and means. The differences between responding and non-responding sample unit median AGI are much smaller than the differences between mean AGI. This again suggests that there is substantial differences in the tail of the distribution. The differences between responding and non-responding sample unit mean AGI are much larger than are the difference between mean WSI across these groups. Whites and sample units with a householder of unknown race have large gaps between respondents and non-respondents, as do Pacific Islanders, and Married Filing Jointly tax units.

Table 17: Household Adjusted Gross Income (AGI) by CE Response Status

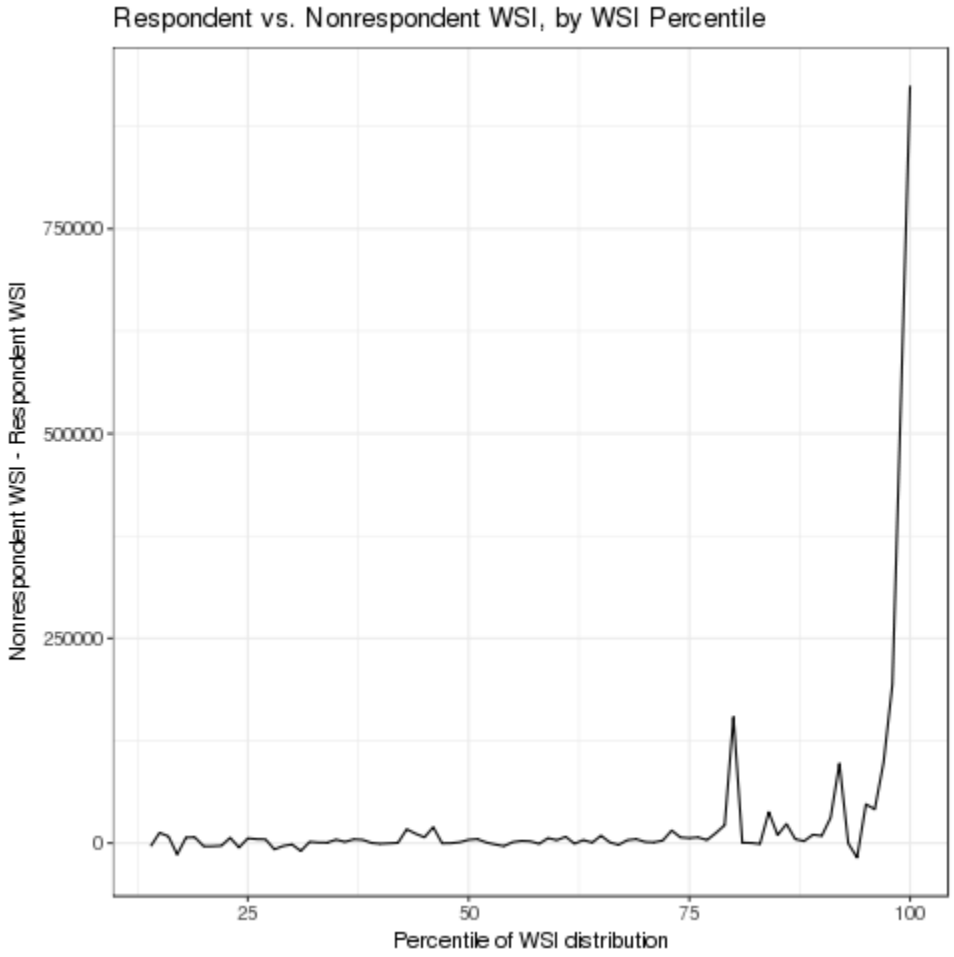
	Obs.	Median AGI CE Respondents	Median AGI CE Non-Respondents	Mean AGI CE Respondents	Mean AGI CE Non-Respondents
All households	21,169	$40,088	$43,918	$63,344	$89,263
Race of Householder					
White Alone	10,340	43,476	48,427	66,707	91,014
Black Alone	1,470	24,137	26,706	34,932	40,084
AIAN Alone	35	47,360	38,394	59,179	38,310
Asian Alone	555	48,830	43,459	71,507	84,439
Pacific Islander Alone	220	28,026	30,592	49,341	113,689
Two or more races	235	33,375	36,108	48,530	49,432
Unknown	8,314	39,478	44,713	63,918	107,091
IRS 1040 Filing Status					
Head of Household	2,962	24,126	26,672	32,104	36,046
Married Filing Jointly	8,984	68,277	74,559	96,189	131,873
Married Filing Separately	293	32,933	37,258	47,723	43,182
Single	8,930	30,855	32,720	41,712	66,175

Source: CE 2013-2014 and IRS 1040s, 2012-2013

For both WSI and the broader AGI income concept, the difference in average income between responding and non-responding sample units is substantially larger than the difference in median

income between these sample unit types. This pattern is suggestive of patterns of differential non-response at the top of the income distribution. To further analyze this, we produce visualizations, similar to Figure 2, which summarize the average difference between non-respondent and respondent income by percentile of the income distribution. Figure 5 depicts this for the WSI income concept, while Figure 6 depicts this for the broader AGI income concept.

Figure 5: Difference in Wage and Salary Income by WSI Percentile



The pattern of income differences between non-responding and responding sample units in Figure 5 points clearly towards differences in the tail of the income distribution being the driver of overall average differences between non-responding and responding households. For the bottom 75 percent of the income distribution average income differences between non-responding and responding sample units are small (less than $5000). In the top quartile, and especially at the top two percentiles of the income distribution, however, substantial deviations emerge. Non-responding sample units in the top 1 percent of the WSI income distribution earn nearly $1 million more than responding sample units in the top 1 percent.

Figure 6: Difference in Adjusted Gross Income, by AGI percentile

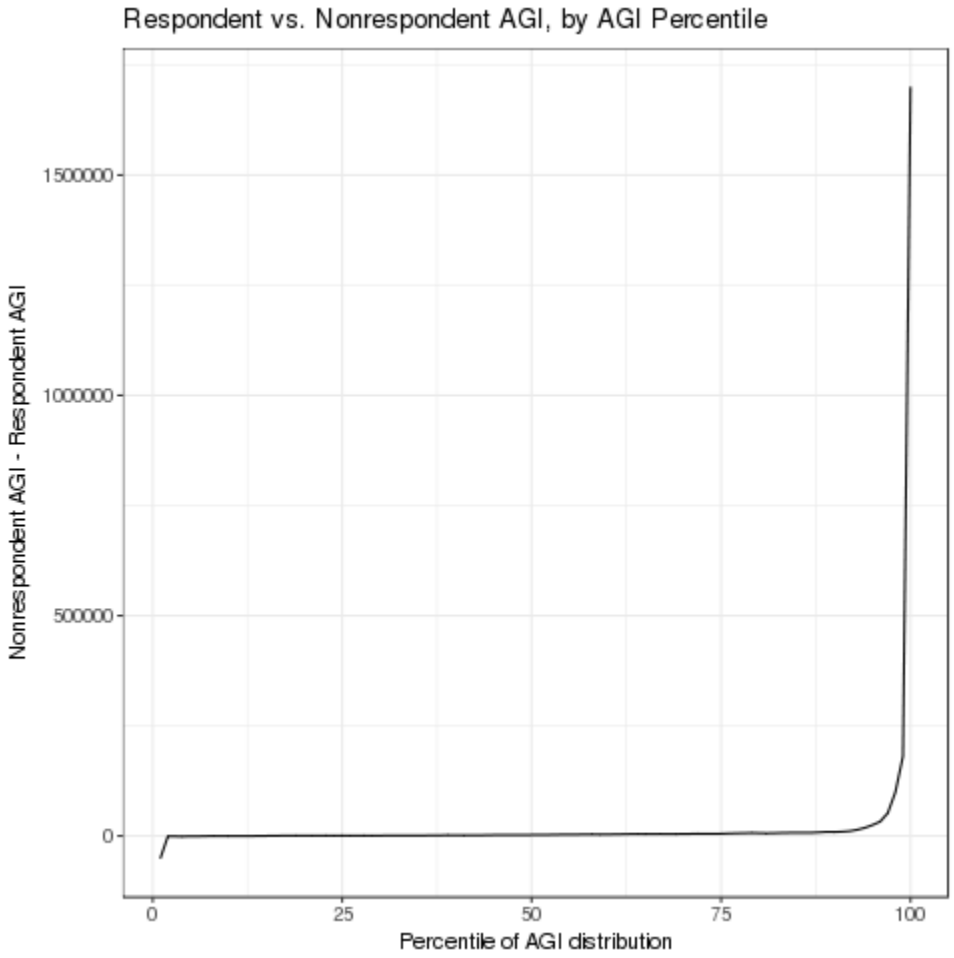
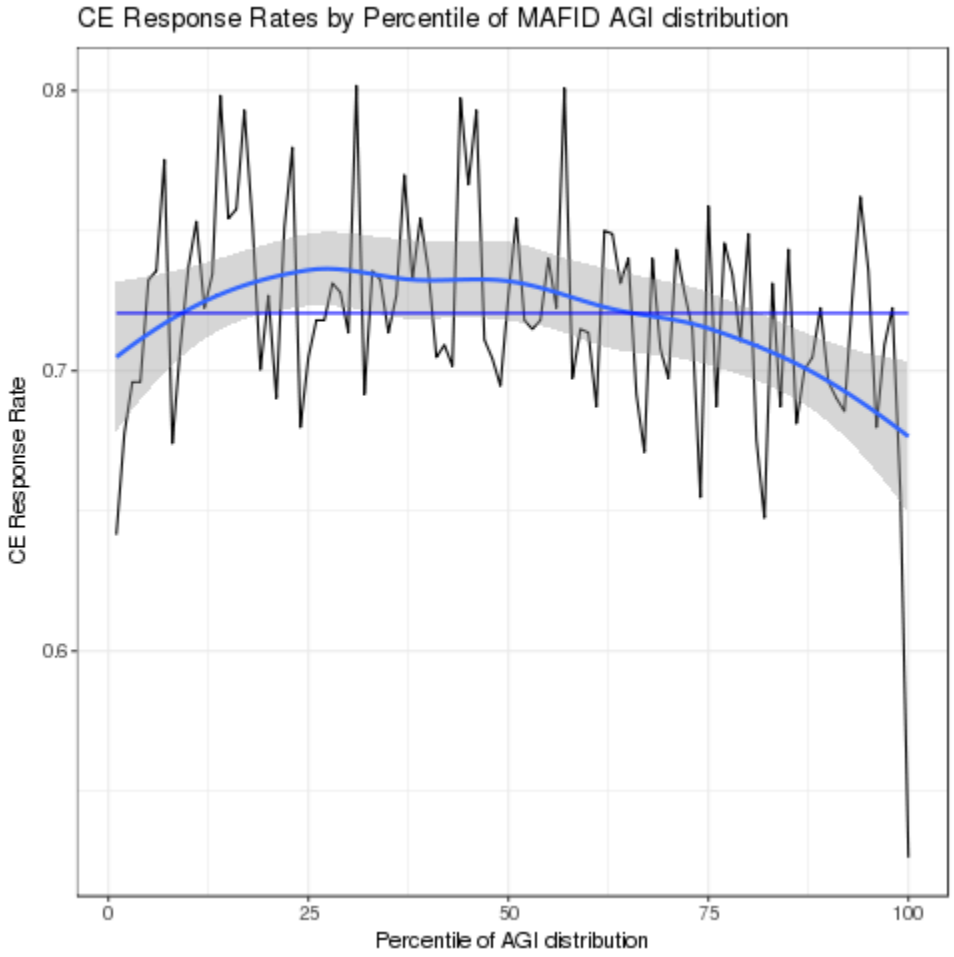


Figure 6 repeats this exercise for the broader AGI income concept. Here the results are even more stark than in the WSI case: from the second through 90th percentiles, differences between non-responding and responding sample units are negligible. At the top percentile of the AGI distribution, again we observe a difference between non-responding and responding sample units over 1.5 times as large as the WSI case. Note that AGI includes business income or losses for schedule C or S firms, and so can be negative for tax units who experience large business losses in a given tax year. Thus the seemingly anomalous result that non-responding sample units in the bottom 1 percent of the AGI distribution have lower incomes on average than responding sample units is driven by these business losses. Both responding and non-responding sample units have negative AGI in the bottom percentile, but non-responding sample units have even more negative AGI (-101,098.05 vs. -49,631.79).

Since sample units who own businesses large enough to sustain tens of thousands of dollars of business losses are likely more similar to sample units at the top of the income distribution, the bottom and the top of the income distributions both illuminate potential non-response patterns related to advantaged households. As a final analysis, we present visualizations of the response

rate to the CE by percentile of the AGI distribution in Figure 7. Response rates are somewhat variable between the second and 90th percentiles, but average about 72 percent. At the very bottom and very top of the income distribution, however, response rates are substantially lower—the top 1 percent of the AGI distribution has a response rate of around 52 percent, while the bottom 1 percent of the AGI distribution has response rates of 64 percent.

Figure 7: Sample Unit Response Rate by Percentile of the AGI distribution



These results together paint a picture of differential survey non-response by more advantaged households. Non-responding households have higher incomes on average, measured either in terms of wages and salary or in terms of a broader AGI definition of income. This difference in average incomes is in fact the result of differences at the very top of the income distribution—sample units in the top 1 percent of the AGI distribution, and to a lesser extent, sample units with larger business losses are substantially less likely to respond to the CE. Further, these non-responding households are substantially richer even than other responding households in the top 1 percent of the income distribution.

5.2 Analysis of Zip Code AGI Classification

Linking respondents and non-respondents to the universe of tax returns provides direct evidence for potential survey non-response bias along income lines. This confirms existing conventional wisdom that higher income households are less likely to respond to surveys. BLS employs a number of weighting processes to address this nonresponse bias. One major part of the weighting process involves categorizing responding and non-responding sample units into "low," "middle" and "high" income categories on the basis of zip-code level information on average AGI provided by the IRS. Since we have information on the actual individual incomes of sample units, we can provide a validation of this categorization, and assess the degree to which categorizing sample units by zip code AGI is properly categorizing sample units by individual AGI.

For this part of the weighting process, BLS matches both respondents and Type A non-interviews to the most recently available annual average zip code AGI data available. For each interview month, BLS sorts sample units by zip code AGI and categorizes all CUs and Type A non-interview sample units into "low", "middle" and "high" AGI categories. These categories are assigned so that the bottom 10 percent of sample units in a Census region in a given month are in the "low" category, the middle 80 percent are in the "middle" category, and the top 10 percent are in the "high" category.

BLS provided these categorizations (in the variable WTIRSINC) for respondents only. We use the address matching process used above to assign an individual AGI value to these responding sample units (thus using a process which would be feasible for both respondents and non-respondents). Using these address-matched sample unit AGI values, we construct a set of categorizations analogous to the zip code categories in WTIRSINC, assigning the bottom 10 percent of sample units to a "low TU AGI" category, the middle 80 percent of sample units to a "middle TU AGI" category, and the "top 10 percent" to a high TU AGI category. Table 18 summarizes the agreement between the WTIRSINC categories and the TU AGI categories for the full address linked sample, and broken out by Census region, with the columns representing IRS-derived grouping and the rows reflecting the CE-derived groups.

It appears that there is not much agreement between these categories, particularly for the "low" and "high" zip code AGI categories. Only about 18 percent of sample units categorized as low zip code AGI are in the bottom 10 percent of tax unit AGI, while only 32 percent of sample units categorized as high zip code AGI are in the top 10 percent of tax unit AGI. There is more agreement in the middle category. About 82 percent of sample units categorized as middle zip code AGI are in the middle 80 percent of the tax unit AGI distribution. There is regional variation in the degree of agreement between these two categorizations. While 21 percent of low zip code AGI sample units in the Northeast region are in the bottom 10 percent of the TU AGI distribution, only 13 percent of low zip code AGI sample units in the South region are in the bottom 10 percent of the TU AGI distribution. The Northeast region also has the highest degree of agreement for the high zip code AGI category (36 percent), while the West region has the lowest level of agreement for the high zip code AGI category (30 percent).

Table 18: Agreement Between CE Zip Code AGI Categories and Form 1040 AGI

	Bottom 10% TU AGI	Middle 80% TU AGI	Top 10% TU AGI
Full Sample			
Bottom 10% Zip Code AGI	0.177	0.811	0.012
Middle 80% Zip Code AGI	0.097	0.821	0.082
Top 10% Zip Code AGI	0.059	0.623	0.318
Northeast Region			
Bottom 10% Zip Code AGI	0.212	0.765	0.023
Middle 80% Zip Code AGI	0.095	0.827	0.078
Top 10% Zip Code AGI	0.036	0.607	0.357
Midwest Region			
Bottom 10% Zip Code AGI	0.2	0.8	0
Middle 80% Zip Code AGI	0.095	0.823	0.082
Top 10% Zip Code AGI	0.063	0.619	0.317
South Region			
Bottom 10% Zip Code AGI	0.131	0.842	0.027
Middle 80% Zip Code AGI	0.1	0.819	0.081
Top 10% Zip Code AGI	0.073	0.613	0.313
West Region			
Bottom 10% Zip Code AGI	0.161	0.833	0.005
Middle 80% Zip Code AGI	0.099	0.816	0.085
Top 10% Zip Code AGI	0.054	0.642	0.304

Source: CE 2013-2014 and IRS 1040s, 2012-2013

Table 19: Summary Statistics of AGI, by CE Zip Code AGI Categories

	Average TU AGI	Average TU AGI, First Quartile	Average TU AGI, Second Quartile	Average TU AGI, Third Quartile	Average TU AGI, Fourth Quartile
			Full Sample		
Bottom 10% Zip Code AGI	$33,937	$8,648	$17,784	$29,128	$80,122
Middle 80% Zip Code AGI	56,207	9,453	29,502	52,901	132,949
Top 10% Zip Code AGI	129,492	19,937	59,427	108,002	330,007
			Northeast Region		
Bottom 10% Zip Code AGI	47,789	6,882	17,171	30,459	136,645
Middle 80% Zip Code AGI	60,584	11,640	33,137	59,365	138,195
Top 10% Zip Code AGI	148,806	29,579	71,259	124,245	370,142
			Midwest Region		
Bottom 10% Zip Code AGI	29,526	7,389	15,803	29,802	64,756
Middle 80% Zip Code AGI	54,850	5,874	30,496	52,910	130,119
Top 10% Zip Code AGI	121,261	19,211	52,856	101,580	308,967
			South Region		
Bottom 10% Zip Code AGI	31,154.51	9,632	17,839	28,758	68,389
Middle 80% Zip Code AGI	50,845.44	11,284	26,252	47,252	118,588
Top 10% Zip Code AGI	123,043.70	17,002	53,137	100,718	321,317
			West Region		
Bottom 10% Zip Code AGI	31,435	10,070	19,605	29,008	66,440
Middle 80% Zip Code AGI	62,236	8,775	31,605	56,611	151,831
Top 10% Zip Code AGI	133,346	21,972	65,075	114,451	331,887

Source: CE 2013-2014 and IRS 1040s, 2012-2013

Of course, some degree of disagreement between these two categorizations would be expected, as long as the income distributions within zip codes are not perfectly equal. We investigate this further by examining the distribution of tax unit AGI within each zip code AGI category. Table 19 summarizes this distribution of TU AGI for the full linked sample, and also breaks out the summary statistics by Census region. The first column reports the mean TU AGI for all tax units within a zip code AGI category. Columns 2 through 5 report the average TU AGI within a quartile of the AGI distribution within a zip code AGI category. For instance, the first row entry of Column 2 represents the average AGI for tax units who are in the bottom 25 percent of the AGI distribution within the "Bottom 10% Zip Code AGI" category.

On average, the zip code categories do correspond to increasing tax unit AGI: the overall national average tax unit AGI in the low zip code AGI category is about $34,000, about $56,000 in the middle zip code AGI category and $129,000 in the high AGI category. Nonetheless, there is substantial inequality within each zip code AGI category. Indeed, for the full sample, and within each Census region, the top quartile of tax unit AGI in the low zip code AGI category have higher incomes than the average AGI in the middle zip code AGI category. The bottom quartile of tax unit AGI in the high zip code AGI category have lower incomes than the average tax unit in the low zip code AGI category.[19]

This widespread of incomes within the BLS-assigned zip code AGI categories suggests that the weighting process may be proceeding under incorrect assumptions about which sample units are in fact low or high income. Although the average incomes within each category are consistent with the low, middle and high income assignments, there are a substantial number of individuals who are categorized as high income by WTIRSINC in fact have quite low tax unit incomes. Additionally, many sample units who are categorized as low income by WTIRSINC are in fact at the upper end of the middle income category in terms of tax unit AGI. This misclassification suggests that, if possible, adopting a weighting process that categorizes respondents and non-respondents by address-linked tax unit AGI may yield conceptually superior results.

6 Conclusion

In order to investigate the possibility of using administrative records to improve the Consumer Expenditure Survey, we link IRS tax records on wages, payroll deductions, self-employment income, retirement income and broad measures of tax unit income to responses and non-responses in the CE survey. These linked data have made possible new insights into the measurement error properties of CE income data, have illuminated possible ways that the imputation process could be improved, and have allowed for an examination of both survey non-response bias and the weighting process used to address it.

[19] Low income individuals in a high AGI zip code may in fact be high-wealth individuals who are retired. However, householders in the bottom quartile of high AGI zip codes are in fact younger on average than individuals in the top 3 quartiles of these zip codes, so it is more likely that the AGI of these individuals represents their true resources.

About 77 percent of CE respondents can be assigned a PIK, the unique person identifier used to link survey and administrative records, while about 93 percent of sample units can be assign MAFIDs, the unique household identifier used to link survey and administrative records by address. We are able to link about 77 percent of individuals to 1040s by MAFID, and about 70 percent by PIK; we can link 39 percent of individuals to the universe of W-2s by PIK. This implies a tax return filing rate of about 90 to 95 percent, and an employment to population ratio of about 52 percent, both numbers that align with aggregate statistics.

Our analysis of the individuals who can be assigned PIKs suggests that, overall, average wages in the CE and average wages in the universe of W-2's have only very small differences. Nonetheless, looking throughout the wage distribution for individuals with both CE and W-2 wages, there is evidence of over-reporting of wages (CE wages are higher than W-2 wages) at the bottom of the wage distribution, and under-reporting of wages (CE wage are lower than W-2 wages) at the top of the distribution. There are a relatively small number of individuals who report no CE wages, but who nonetheless earned wages in the W-2 universe; the imputation process does not always assign an imputed value to these individuals. Under-reporting and incorrect wage item non-response seem to be concentrated among individuals in the youngest cohort, aged 14-24.

We also compare CE responses to administrative records for questions about payroll deductions, self-employment income and retirement income. We find that W-2 wage records suggest a slightly higher SS payroll deduction rate than the CE responses (suggesting under-reporting), but very similar rates of self-employment income receipt across the two data sources. Retirement income appears to be dramatically under-reported on the CE. Approximately twice as many individuals receiving retirement income in the universe of 1099-R than report retirement income on the CE, although this may in part represent a conceptual difference between the CE and IRS definitions of retirement income.

We find evidence of systematic survey non-response across the income distribution by linking responding and non-responding households to administrative records by address. Non-responding sample units are much more likely to be richer than responding households, measured either by wage and salary income or by broader AGI. Furthermore, nonresponse rates are higher at the top of the income distribution, and high income non-responding sample units are substantially richer even than high income responding sample units. This indicates that very high income individuals are systematically less likely to respond to the CE. We additionally use the address-linked data to analyze one part of the non-response weighting process, and find that the categorization of sample units by zip code AGI does not line up cleanly with the individual incomes of these sample units. In particular, poor sample units in high-income zip codes and richer sample units in low-income zip codes are likely misclassified under this categorization.

These results suggest potential ways forward to improve income statistics produced from the CE. In particular, both the PIK and MAFID based analysis suggest areas where current imputation and weighting procedures may be producing less than optimal outcomes. Incorporating the linkage of administrative records into these production processes has the potential to improve the accuracy and quality of statistics produced from the CE.

References

Bee, C. Adam, and Joshua Mitchell. 2017. *Do Older Americans Have More Income Than We Think?* US Census Bureau.

Bee, C. Adam, and Joshua Mitchell. 2016. *The Hidden Resources of Women Working Longer: Evidence from Linked Survey-Administrative Data.* NBER Working Paper No. 22970.

Luque, Adela, and Renuka Bhaskar. 2014. *2010 American Community Survey Match Study*. U.S. Census Bureau.

O'Hara, Amy, Sonya Rastogi, James Noon, Ellen A. Zapata, Cindy Espinoza, Leah B. Marshall, Teresa A. Schellhamer, and J. David Brown. 2012. *2010 Census Match Study.* US Census Bureau.

Wagner, Deborah, and Mary Layne. 2014. *The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software.* US Census Bureau.

Appendix: Additional Tables and Figures

Table A1: Comparing Median Individual Wages, CE vs. W-2 (Nonimputed CE Responses only)

	Obs.	Unweighted		Weighted	
		Median Wages, CE	Median Wages, W-2	Median Wages, CE	Median Wages, W-2
All individuals	27,914	$34,820	$29,796	33,000	28,719
			Sex		
Female	14,278	27,998	24,646	27,000	23,980
Male	13,636	41,020	36,238	40,000	35,096
			Race		
White Alone	22,477	34,999	30,870	34,460	29,844
Black Alone	3,264	25,997	22,004	25,000	21,146
AIAN Alone	136	23,950	21,740	37,261	35,319
Asian Alone	1,382	37,985	36,104	23,000	21,446
Pacific Islander Alone	157	31,250	28,622	32,000	31,460
Two or more races	498	23,628	20,637	20,000	20,206
			Age		
14-24	3,923	9,000	7,369	8,400	7,274
25-44	7,078	38,001	33,562	36,000	32,446
45-64	7,911	41,998	40,458	41,061	39,788
65+	4,434	19,978	15,673	20,000	15,490
			Hispanic Origin		
Hispanic	4,561	25,002	22,997	24,960	22,606
Not Hispanic	23,353	36,001	31,232	35,000	30,000
			Income		
Above Median CU Wages	14,760	50,001	42,499	48,000	40,929
Below Median CU Wages	13,154	16,000	15,754	15,900	15,524
			Second Interview Month		
1	1,917	32,039	29,175	30,000	27,788
2	2,120	33,486	30,832	32,789	30,567
3	2,087	31,267	28,584	30,824	28,447
4	2,124	35,011	29,578	34,857	28,517
5	2,072	31,963	27,900	30,000	26,257
6	2,145	35,000	31,293	35,000	30,841
7	2,152	31,964	28,987	30,000	27,755
8	2,147	35,042	31,123	35,000	30,104
9	2,090	33,614	29,866	32,000	28,666
10	3,094	35,032	30,404	34,987	28,846
11	3,099	35,968	29,941	35,000	28,782
12	2,867	34,978	29,887	32,589	28,376

Source: CE 2013-2014, IRS W-2s 2012-2013

Table A2: Comparing Median Individual Wages, CE vs. W-2 (Imputed and Nonimputed CE Responses)

	Obs.	Unweighted		Weighted	
		Median Wages, CE	Median Wages, W-2	Median Wages, CE	Median Wages, W-2
All individuals	27,914	$31,490	$29,796	30,000	28,719
			Sex		
Female	14,278	26,000	24,646	25,212	23,980
Male	13,636	36,000	36,238	35,000	35,096
			Race		
White Alone	22,477	32,904	30,870	25,212	23,980
Black Alone	3,264	24,063	22,004	35,000	35,096
AIAN Alone	136	23,950	21,740	25,212	23,980
Asian Alone	1,382	37,998	36,104	35,000	35,096
Pacific Islander Alone	157	28,046	28,622	25,212	23,980
Two or more races	498	23,084	20,637	35,000	35,096
			Age		
14-24	3,923	8,572	7,369	8,025	7,274
25-44	7,078	35,000	33,562	34,649	32,446
45-64	7,911	40,000	40,458	40,000	39,788
65+	4,434	22,022	15,673	21,263	15,490
			Hispanic Origin		
Hispanic	4,561	23,000	22,997	23,000	22,606
Not Hispanic	23,353	34,000	31,232	32,366	30,000
			Income		
Above Median CU Wages	14,760	45,000	42,499	44,806	40,929
Below Median CU Wages	13,154	16,420	15,754	16,000	15,524
			Second Interview Month		
1	1,917	30,003	29,175	28,000	27,788
2	2,120	30,804	30,832	30,000	30,567
3	2,087	29,985	28,584	30,000	28,447
4	2,124	29,998	29,578	28,568	28,517
5	2,072	30,017	27,900	27,984	26,257
6	2,145	31,989	31,293	31,796	30,841
7	2,152	30,001	28,987	28,000	27,755
8	2,147	32,447	31,123	32,000	30,104
9	2,090	31,995	29,866	30,000	28,666
10	3,094	33,620	30,404	32,000	28,846
11	3,099	32,005	29,941	31,538	28,782
12	2,867	31,999	29,887	30,000	28,376

Source: CE 2013-2014, IRS W-2s 2012-2013

Table A3: Comparing Median Individual Wages, CE vs. W-2 Nonimputed CE Responses, all individuals with both CE and W-2 Wages

	Obs.	Unweighted		Weighted	
		Median Wages, CE	Median Wages, W-2	Median Wages, CE	Median Wages, W-2
All individuals	9,124	$36,000	$34,620	$36,000	$34,081
Sex					
Female	4,504	30,000	29,107	30,000	28,790
Male	4,620	44,329	40,969	42,928	40,402
Race					
White Alone	7,647	38,000	35,595	36,690	35,014
Black Alone	833	28,012	26,463	28,012	25,678
AIAN Alone	41	32,333	25,469	32,333	24,338
Asian Alone	442	39,954	38,842	39,954	38,261
Pacific Islander Alone	62	32,500	31,320	36,616	31,110
Two or more races	99	29,833	27,081	26,667	26,396
Age					
14-24	1,070	11,932	8,608	11,125	8,622
25-44	3,841	40,003	36,099	38,002	35,931
45-64	3,711	45,002	43,885	44,001	43,535
65+	502	22,534	23,799	21,917	23,841
Hispanic Origin					
Hispanic	1,259	27,012	26,666	27,012	26,384
Not Hispanic	7,865	38,996	36,272	38,000	35,854
Income					
Above Median CU Wages	6,359	50,001	46,381	50,001	45,736
Below Median CU Wages	2,765	17,999	18,091	17,999	17,932
Second Interview Month					
1	667	35,005	33,667	35,005	33,141
2	744	36,947	35,178	37,008	34,606
3	722	34,012	32,166	34,012	31,614
4	704	40,004	36,622	38,463	36,317
5	683	33,903	30,500	32,852	29,882
6	721	38,019	36,112	38,019	35,708
7	682	34,556	33,548	33,170	32,304
8	684	38,027	35,400	38,027	34,954
9	660	37,111	34,019	36,064	33,776
10	994	37,992	35,550	36,123	34,886
11	940	39,790	35,531	39,527	34,944
12	923	35,974	35,732	34,972	35,120

Source: CE 2013-2014, IRS W-2s 2012-2013

Table A4: Median CE vs. W-2 Income, by CE Income Bracket

		Unweighted		Weighted	
		Mean CE Wages	Mean W-2 Wage	Mean CE Wages	Mean W-2 Wage
CE Brackets ($)	[0-5K)	$1,854	$4,385	$1,854	$4,248
	[5K-10K)	7,179	7,910	7,094	7,784
	[10K-15K)	12,002	12,697	12,000	12,541
	[15K-20K)	17,069	16,949	17,046	16,903
	[20K-30K)	24,573	23,779	24,573	23,711
	[30K-40K)	34,512	32,207	34,537	31,980
	[40K-50K)	44,088	41,051	44,099	41,050
	[50K-70K)	58,000	54,395	58,000	54,461
	[70K-90K)	78,000	73,497	78,000	73,686
	[90K-120K)	100,000	97,159	100,000	96,488
	120K+	164,915	146,513	165,000	145,692

Source: CE 2013-2014, IRS W-2s 2012-2013

Table A5: Comparing Total MAFID W-2 Wages Responding vs. Non-responding Households

	Median W-2 Wages		Mean W-2 Wages	
	CE Respondent Households	CE Non-Respondent Households	CE Respondent Households	CE Non-Respondent Households
All households	$50,279	$51,594	$73,991	$96,274
Race of Householder				
White Alone	52,912	54,445	76,840	86,486
Black Alone	32,470	38,029	47,925	52,784
AIAN Alone	52,662	32,407	64,635	32,255
Asian Alone	71,114	61,915	104,063	99,944
Pacific Islander Alone	32,394	40,314	76,165	70,974
Two or more races	37,370	50,095	62,298	66,937
Unknown	49,630	49,726	73,407	145,655

Source: CE 2013-2014, IRS W-2s 2012-2013

Table A6: Comparing Mean CE and IRS W-2 Wages, Individuals with Imputed CE Wages Only

	Obs.	Unweighted		Weighted	
		Mean Wages, CE	Mean Wages, W-2	Mean Wages, CE	Mean Wages, W-2
All individuals	7,630	$40,016	$46,767	38,709	44,943
Sex					
Male	4,055	43,915	58,701	42,256	56,342
Female	3,575	35,594	33,425	34,799	32,671
Race					
White Alone	5,937	41,705	48,070	40,388	45,963
Black Alone	1,012	27,768	33,267	27,219	32,202
AIAN Alone	36	36,116	24,270	29,064	28,210
Asian Alone	403	51,822	63,560	52,453	67,780
Pacific Islander Alone	45	26,533	48,034	28,382	46,943
Two or More Races	197	31,701	35,308	30,656	37,564
Age					
14-24	777	13,658	11,602	13,065	11,465
25-44	1,724	39,042	46,106	38,275	44,750
45-64	1,713	48,449	62,196	48,047	61,086
65+	380	32,657	46,836	31,648	45,920
Hispanic Origin					
Not Hispanic	6,284	42,696	49,267	41,238	47,273
Hispanic	1,346	27,507	29,306	27,494	29,308
Income					
Above Median CU Wages	4,883	52,136	57,469	50,597	55,338
Below Median CU Wages	2,747	18,472	25,619	18,016	24,700
Second Interview Month					
1	495	40,887	43,911	39,345	41,910
2	536	38,227	55,327	39,006	53,450
3	547	41,494	52,364	38,159	48,948
4	608	37,810	43,089	35,929	40,091
5	524	43,674	44,247	42,650	42,210
6	607	42,678	42,962	42,383	42,541
7	644	37,618	44,653	35,874	43,905
8	649	39,529	45,453	39,169	44,535
9	610	40,356	51,945	40,021	49,456
10	827	39,451	42,883	37,046	39,882
11	852	39,009	44,748	38,040	44,100
12	731	40,711	52,714	38,736	50,150

Source: CE 2013-2014, IRS W-2s 2012-2013