Working Paper No. 914

Quality of Match for Statistical Matches Using the American Time Use Survey 2013, the Survey of Consumer Finances 2013, and the Annual Social and Economic Supplement 2014

by

Fernando Rios-Avila
Levy Economics Institute of Bard College

September 2018

The Levy Economics Institute Working Paper Collection presents research in progress by Levy Institute scholars and conference participants. The purpose of the series is to disseminate ideas to and elicit comments from academics and professionals.

Levy Economics Institute of Bard College, founded in 1986, is a nonprofit, nonpartisan, independently funded research organization devoted to public service. Through scholarship and economic research it generates viable, effective public policy responses to important economic problems that profoundly affect the quality of life in the United States and abroad.

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

ISSN 1547-366X

ABSTRACT

This paper describes the quality of the statistical matching between the March 2014 supplement to the Current Population Survey (CPS) and the 2013 American Time Use Survey (ATUS) and Survey of Consumer Finances (SCF), which are used as the basis for the 2013 Levy Institute Measure of Economic Well-Being (LIMEW) estimates for the United States. In the first part of the paper, the alignment of the datasets is examined. In the second, various aspects of the match quality are described. The results indicate that the matches are of high quality, with some indication of bias in specific cases.

KEYWORDS: Statistical Matching; American Time Use Survey; Survey of Consumer Finances; Levy Institute Measure of Economic Well-Being (LIMEW); United States

JEL CLASSIFICATIONS: C14; C40; D31

INTRODUCTION

This paper describes the construction of the synthetic dataset created for use in the estimation of the Levy Institute Measure of Economic Well-Being (LIMEW) for the United States. The LIMEW was developed as an alternative to conventional income measures that provides a more comprehensive measure of economic well-being.[1] Construction of the LIMEW requires a variety of information for households. In addition to the standard demographic and household income information, the estimation process also requires information about household members' time use and information on a household's wealth, assets, and debts. Unfortunately, no single dataset contains all required data for the estimation.

In order to produce LIMEW estimates, a synthetic dataset is created combining information from three datasets, applying a statistical matching process.[2] For the United States, the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) 2014 is used as the base dataset, as it contains good information regarding demographic, social, and economic characteristics, as well as income, work experience, noncash benefits, and migration status of persons 15 years old and over. Time use data comes from the American Time Use Survey (ATUS) 2013, which provides rich data regarding how people divide their time among life's activities, including time spent doing paid and unpaid activities, inside and outside the household, for one person in the household. Wealth data come from the Survey of Consumers Finances (SCF) 2013, which collects detailed information on household finances, income, assets, and liabilities.

This paper is organized as follows. Section one describes the data. Section two assesses the alignment of the information between ASEC and ATUS at the individual level, and the ASEC and the SCF at the household level. Section three briefly describes the methodology and analyzes the matching quality of the statistical matching. Section four concludes.

[1] For details on the background of the LIMEW, see Wolff and Zacharias (2003).
[2] For further details on the methodology, see Kum and Masterson (2010).

1. DATA DESCRIPTION

1.1. Annual Social Economics Supplement (ASEC)

The CPS is a monthly survey administered by the US Bureau of Labor Statistics. It is used to assess the activities of the population and provide statistics related to employment and unemployment in the current labor market. Each household in the CPS is interviewed for four consecutive months, not interviewed for eight, and interviewed again for four additional months. Although the main purpose of the survey is to collect information on the labor market situation, the survey also collects detailed information on demographic characteristics (age, sex, race, and marital status), educational attainment, and family structure.

In March of every year, the previously interviewed households answer additional questions, part of the ASEC supplement formerly known as the Annual Demographic File. In addition to the basic monthly information, this supplement provides additional data on work experience, income, noncash benefits, and migration. In 2014, the ASEC supplement went through a redesign of the income-collection questions. As described in Semega and Welniak (2013), for the ASEC 2014, of the nearly 98,000 addresses in the sample, approximately one-third of the sample was randomly assigned to be eligible to receive the redesigned income questions. The remaining sample (approximately two-thirds) was eligible to receive the set of ASEC income questions used in previous years, referred to here as the "traditional income questions." For the statistical matching purposes, we use the second subsample.

The ASEC 2014 is used as the base dataset (recipient), as it contains rich information regarding demographics and economic status. Because the time use survey (described below) covers individuals 15 years of age and older, younger individuals are discarded from the ASEC sample. This leaves us with a total of 107,369 observations, representing 252,088,834 individuals when weighted. For the household-level analysis, only information regarding the householder[3] is used, leaving 51,466 observations, representing 122,951,925 households when weighted.

[3] The ASEC and the SCF use different definitions in regards to the person of reference in the household. In the ASEC, the householder refers to the person in whose name the housing unit is owned or rented. If the house is owned by a married couple, the householder may be either the husband or wife. For the SCF, the concept of the head of the household is defined as the male in a mixed couple, and the older individual in the case of a same-sex

1.2. American Time Use Survey (ATUS)

The ATUS, a survey sponsored by the Bureau of Labor Statistics and collected by the US Census Bureau, is the first continuous survey on time use in the United States available since 2003. Its main objective is to provide nationally representative estimates of peoples' allocation of time among different activities, collecting information on what they did, where they were, and with whom they were.

The ATUS is administered to a random sample of individuals selected from a set of eligible households that have completed their final month's interviews for the CPS. The ATUS covers all residents who are at least 15 years old and are part of the civilian, noninstitutionalized population in the United States.

The ATUS 2013, which contains a total of 11,345 observations, is used as the donor dataset to obtain information regarding time use, which will be transferred to the ASEC 2014. Since information regarding household income is incomplete, the information was imputed using a univariate imputation process and information from the ASEC 2013. The sample represents a total of 248,718,989 individuals.

1.3. Survey of Consumer Finances (SCF)

The SCF is normally a triennial cross-sectional survey, sponsored by the Board of Governors of the Federal Reserve System in cooperation with the US Department of the Treasury, which collects information on families' balance sheets, pensions, income, and demographic characteristics.[4] The purpose of the survey is to provide detailed information on households' assets and liabilities that can be used for analyzing households' wealth and their use of financial services.

In order to provide reliable information on household wealth distribution, the SCF is based on a dual-frame sample design. On the one hand, a geographically based random sample of respondents is interviewed to obtain a sample that is broadly representative of the population as a whole. On the other hand, a supplemental sample is obtained to include a sample of wealthy

couple. Through the rest of the document, the term "householder" will be used to refer to the person of reference, head of the household, or householder.

[4] Over the 1983–89 and 2007–09 periods, the SCF has collected information in panel data.

families in order to provide accurate information on wealth distribution, as the distribution of nonhome assets and liabilities is highly concentrated. In order to deal with the missing data, most variables with missing values are imputed using a multiple imputation procedure from which five replicates (imputations) for each record are obtained. [5]

The SCF 2013 is used as the donor dataset to obtain information regarding assets, debts, and net worth. For the SCF 2013, a total of 6,015 families/households were interviewed. In order to account for the multiple imputation information, the five replicates are combined and used for the matching procedure. This provides a sample of 30,075 observations, representing 122,530,057 households when weighted.

2. DATA ALIGNMENT AND STATISTICS

2.1. ATUS 2013 – ASEC 2014

In order to create the synthetic dataset and transfer the time use information from the donor to the recipient dataset as closely as possible, five strata variables are used to perform the match within the defined subsamples (cells). These strata variables are *sex, parental status, labor force status, marital status,* and *spouse's labor force status.* The combination of these five strata variables provides a total of 24 cells that are used to perform a within-cell match. Table 1 presents summary statistics that compare the distribution of individuals within the strata variables. Since both datasets were collected within one year of each other, one should expect them to be well aligned.

[5] For information regarding the use and estimation of replicate samples, see Kennickell (2000) and Kennickell, Woodburn, and Woodburn (1999).

Table 1. Summary Statistics, Alignment across Strata Variables

	ASEC	ATUS	diff
Individuals	252,089,444	241,823,036	-0.8%
Sex			
Female	51.5%	51.6%	0.1%
Male	48.5%	48.4%	-0.1%
Parental status			
No	63.8%	64.3%	0.5%
Yes	36.2%	35.7%	-0.5%
Labor force status			
Not employed	42.6%	39.2%	-3.4%
Employed	57.4%	60.8%	3.4%
Spouse			
No	45.0%	43.3%	-1.7%
Yes	55.0%	56.7%	1.7%
Spouse's labor force status			
Spouse not employed	19.8%	19.6%	-0.1%
Spouse employed	35.2%	37.1%	1.9%

Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

As can be observed in table 1, the distribution of the sample with respect to sex and parental status is almost identical for both the ASEC and ATUS, with 48.5 percent of the sample being male, and about 36 percent being parents. The labor force status shows a relatively larger imbalance. The ATUS indicates there is a 3.4 percentage point larger share of employed individuals in the sample compared to the corresponding statistic in the ASEC survey (57.4 percent). The distribution of individuals across marital status presents a less severe imbalance. The statistics show that the share of married individuals is larger (1.7 percentage points) in the ATUS compared to the ASEC. In terms of the spouse's labor force status, the differences in the distribution among married individuals are negligible.

Table 2 presents statistics on additional variables that characterize the observations in both the donor and recipient datasets. The distribution across household income categories shows some imbalance, with the ATUS showing a considerably lower proportion of households in the highest income category, suggesting some undersampling of high-income households. For other demographic characteristics, such as age, race, and educational attainment, the distribution of individuals in both surveys is close. The largest observed differences in this characteristic are seen in the categories of some college (2.3 percentage points) and whites (2.1 percentage

points), with other differences falling below 2 percentage points. Finally, in terms of household structure, the survey's distribution is close in terms of number of children in the household, with slightly larger discrepancies in terms of the number of adult persons in the household, where the ATUS indicates a smaller share of larger households.

As expected, although some differences in the distributions can be observed between both surveys, these differences are small and there are no systematic differences that might seriously affect the quality of the matching process. Based on the strata variables described above, 24 matching cells were created to be used for exact matching between both surveys.

Table 2. Summary Statistics, Alignment across Selected Variables

	ASEC	ATUS	diff
Household income category			
0–14,999	9.4%	11.8%	2.4%
15,000–34,999	18.5%	21.6%	3.1%
35,000–49,999	13.6%	13.7%	0.1%
50,000–74,999	18.3%	18.7%	0.4%
75,000+	40.2%	34.2%	-5.9%
Age category			
15 to 24	17.1%	17.2%	0.1%
25 to 34	16.8%	16.6%	-0.2%
35 to 44	15.8%	15.9%	0.1%
45 to 54	17.0%	17.4%	0.3%
55 to 64	15.7%	15.6%	-0.1%
65 and older	17.7%	17.4%	-0.3%
Race			
White	65.0%	67.1%	2.1%
Black	11.7%	11.7%	0.0%
Other	15.5%	15.3%	-0.2%
Hispanic	7.8%	6.0%	-1.8%
Educational attainment			
Less than high school	16.7%	16.4%	-0.3%
High school	28.2%	28.7%	0.5%
Some college	18.5%	16.2%	-2.3%
College/grad school	36.7%	38.7%	2.0%
Number of children under 18 in household			
0	61.0%	60.7%	-0.3%
1	17.0%	16.9%	-0.1%
2	13.7%	14.0%	0.3%
3	5.7%	5.6%	-0.1%
4	1.9%	2.1%	0.3%
5 or more	0.9%	0.7%	-0.1%
Number of persons in household over 18			
0	0.0%	0.0%	0.0%
1	16.8%	18.7%	2.0%
2	53.4%	55.7%	2.3%
3	18.1%	16.6%	-1.5%
4	8.2%	7.1%	-1.0%
5	2.6%	1.4%	-1.2%
6 or more	1.0%	0.5%	-0.5%

Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

2.2. SCF 2013 – ASEC 2014

Similar to the previous case, in order to create the synthetic dataset that combines the SCF and ASEC information, five strata variables are used to perform the statistical matching. These strata variables are income category, homeownership, family type, and race and age of the householder. In this case, the households/families rather than individuals are used as the unit of observation. The combination of these five strata variables provides a total of 360 cells that are initially used to perform the match. Table 3 presents summary statistics that compare the distribution of observations within the strata variables. Since both datasets were collected within one year of each other, one should expect them to be well aligned.

Table 3. Summary Statistics, Alignment across Strata Variables

	ASEC	SCF	diff
Individuals	122,951,925	122,530,057	-0.34%
Household income category			
<$20k	18.67%	21.32%	2.65%
$20–50k	29.29%	33.38%	4.09%
$50–75k	17.60%	15.77%	-1.83%
$75–100k	11.98%	9.92%	-2.06%
> $100k	22.46%	19.62%	-2.84%
Homeownership			
Renter	35.32%	34.85%	-0.47%
Owner w/mortgage	37.91%	42.92%	5.01%
Owner wo/mortgage	26.77%	22.23%	-4.54%
Family type			
Couple	54.56%	57.15%	2.59%
Single female	27.62%	27.64%	0.02%
Single male	17.82%	15.21%	-2.61%
Race category			
White	67.95%	70.09%	2.14%
Black	12.64%	14.61%	1.97%
Other	6.57%	4.65%	-1.92%
Hispanic	12.84%	10.64%	-2.20%
Age Category			
<35	19.51%	20.76%	1.25%
35–49	25.89%	26.62%	0.73%
50–65	29.70%	29.02%	-0.68%
>65	24.90%	23.59%	-1.31%

Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

As observed in table 3, the distribution of households across income categories shows good balance across both samples, displaying at most a 5 percentage point difference. The SCF has a smaller share of middle-to-high-income households. Based on race and age, the distribution is very well balanced, with a less than 1.5 percentage point difference in the distributions, and a small underrepresentation of Hispanic and other races in the SCF.[6] The largest distributional differences are present across family type and homeownership. The SCF dataset shows a larger share of households within the "couples" categories (2.6 percentage points), while households with single males are underrepresented (2.6 percentage points).[7] Regarding homeownership, both samples present similar shares of renters and homeowners. Within the homeowners category, however, the ASEC underrepresents households with mortgages in about 5 percent of the instances compared to the SCF. Under the assumption the ASEC information is correct, the excess of mortgage debt is redistributed among householders with mortgages. This strategy has the advantage of keeping the total amount of mortgage debt unchanged in the imputed data, although this might imply some overestimation of mortgage debt when comparing households with mortgages in both datasets (see figure 6).

Table 4 presents statistics on additional variables that characterize the observations in both datasets. Information on education and occupation categories corresponds to that of the householder. The surveys are well balanced in terms of the educational attainment of the householder, the number of persons within the household, and the occupational categories.

[6] While the table shows the distribution for four age categories, the strata variable only differentiates between household heads older and younger than 65.

[7] It is possible that the underrepresentation of "couple" households in the ASEC survey compared to the SCF is because the latter uses the definition of a consumer unit, which is compared with the former "household" definition. In the ASEC definition, a household can contain more than one family (couple).

Table 4. Summary Statistics, Alignment across Selected Variables

	ASEC	SCF	diff
Education category			
Less than high school	11.7%	11.0%	-0.7%
High school grad	29.3%	31.3%	2.0%
Some college	27.6%	25.6%	-2.0%
College or higher	31.5%	32.1%	0.7%
Sex of householder			
Female	27.9%	28.4%	0.5%
Male	72.1%	71.6%	-0.5%
Number of persons in household			
1 person	27.5%	25.6%	-2.0%
2 persons	34.1%	33.4%	-0.7%
3 or more	38.4%	41.1%	2.7%
Occupation category			
Occ1: 37–199	26.2%	28.6%	2.4%
Occ2: 203–389	13.3%	11.3%	-2.0%
Occ3: 403–469 & 903–905	8.7%	9.0%	0.4%
Occ4: 503–699	10.2%	10.4%	0.1%
Occ5: 703–889	7.6%	6.7%	-0.9%
Occ6: 473–499	0.6%	0.7%	0.0%
Other	33%	33.4%	0.0%

Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

The distribution of the sex of the householder shows some imbalance across both datasets. In the ASEC, the householder or person of reference is selected randomly in cases of couples. For consistency, we assign the male within the couple to be considered as the householder, a definition closer to the SCF's head of household. While the SCF survey indicates that a large share of householders (72.1 percent) are male, the ASEC shows 71.6 percent of householders are male. The next section describes the quality of the matching.

3. MATCHING QUALITY

Statistical matching is a widely used technique, predominantly in observational studies in the medical literature. This method consists of combining the information from two separate and independent surveys into a single combined dataset from which statistical inferences can be obtained. The methodology enables the combination of the datasets using common information between both surveys, preserving the distributional characteristics of the combined

information.[8] In the following, the match quality between the ASEC 2014 (recipient) and ATUS 2013 (donor), and ASEC 2014 (recipient) and SCF 2013 (donor), correspondingly, are assessed.

3.1. Matching: ATUS and ASEC

In order to obtain a good match, the matching process begins using five strata variables, namely sex, parental status, labor force status, marital status, and spouse's labor force status, to obtain 24 matching cells. Within each of these cells, propensity scores are estimated using logit models. A dummy variable indicating if the observation corresponds to the donor or the recipient survey is used as a dependent variable. A set of demographic variables (i.e., age, educational attainment, race, parental status, marital status, and employment status) and household characteristics (i.e., number of adults, number of children, and household monthly income) are included as independent variables. For subsequent matching rounds, broader matching cells are defined accordingly, keeping the logit specifications consistent across all models, and including the omitted strata variable in the specification. The logit models and propensity scores are estimated using all information within broader cells, but the matching is done only across observations left unmatched from previous rounds.

Turning to the results of the match performance, table 5 shows the distribution of the matched records by matching round. As expected from these types of processes, 93.4 percent of the matches occur on the first round, ensuring the highest level of match quality. At the same time, only 0.03 percent of the weighted sample was left unmatched after eight matching rounds. These unmatched observations should not bias the distributional statistics of the transferred information.

[8] For further details on the matching procedure, see Kum and Masterson (2010).

Table 5. Distribution of Matched Records by Matching Round

Matching round	Records matched	Percent	Cumulative percent
1	235,545,505	93.4	93.4
2	5,419,781	2.2	95.6
3	1,255,582	0.5	96.1
4	3,919,658	1.6	97.6
5	3,543,327	1.4	99.1
6	359,885	0.1	99.2
7	174,574	0.1	99.3
8	1,787,089	0.7	100
9	84,043	0.03	100
Total	252,089,444	100	

Source: Author's calculations based on ASEC 2014 and ATUS 2013 matched data.

Table 6 provides a description of the match quality, comparing some distributional statistics on the weekly hours of household production between the original information (ATUS) and the imputed data (ASEC). Table 6 also presents some statistics on three components of household production.[9] Given the large presence of zero hours allocated to household production in the sample, some ratios and statistics are not available. The percentile ratios are all equivalent with identical Gini coefficients (0.524). The means and medians on the disaggregated components of household production also show a strong equivalence between both surveys, indicating a strong balance in aggregate terms.

[9] Household production can be broadly categorized into three groups or components: care (childcare, education, etc.), procurement (shopping, etc.), and core (cooking, cleaning, laundry, etc.).

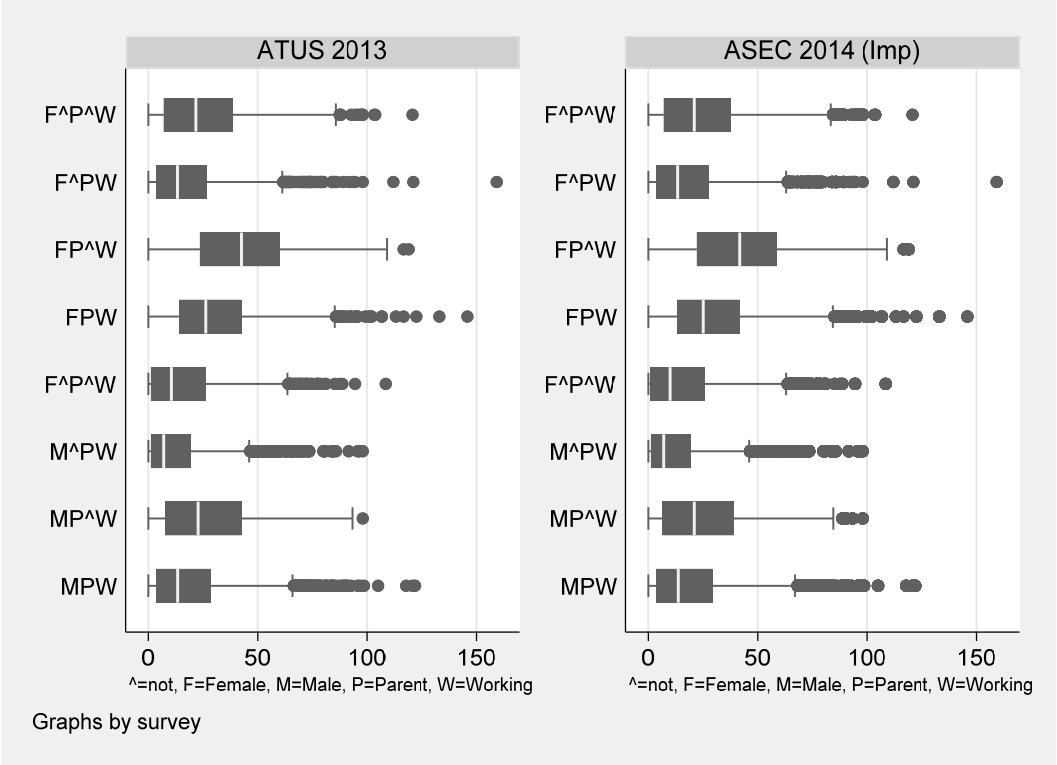
Table 6. Matching Quality: Summary Statistics

	ATUS 2013	ASEC 2014	Ratio ASEC/ATUS
Distributional statistics			
p90/p10	.	.	
p90/p50	3.36	3.36	100%
p50/p10	.	.	
p75/p25	8.29	8.29	100%
p75/p50	2.12	2.12	100%
p50/p25	3.91	3.91	100%
Gini	0.524	0.525	100%
Summary statistics			
Average household production, weekly hours	21.9	21.9	100%
Average care, weekly hours	3.7	3.7	100%
Average procurement, weekly hours	5.3	5.2	100%
Average core, weekly hours	12.9	12.9	100%
Median household production, weekly hours	16.0	16.0	100%
Median care, weekly hours	0	0	
Median procurement, weekly hours	0	0	
Median core, weekly hours	7	7	100%

Note: Household production activities are classified in three classes: care, such as childcare and education; procurement, such as shopping for groceries and clothes; and core, such cooking and cleaning.
Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

Figure 1 presents a visual representation of the distribution of hours allocated to household production using three of the strata variables: sex, parental status, and labor force status. The figure shows that except for some values on the right tail of the distributions—for example, women who are not parents and are not working (F^P^W) and men who are parents and are not working (MP^W)—the overall distributions within the strata variables are analogous, indicating a good match quality.

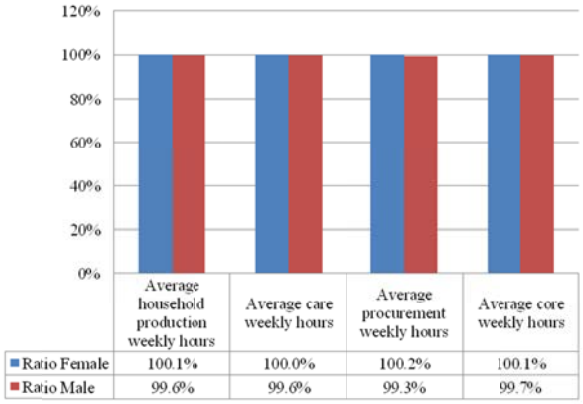
Figure 1. Distritubtion of Hours in Household Production, by Survey



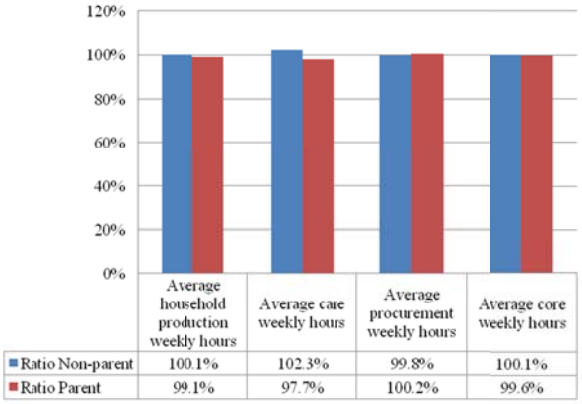
Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

For a detailed review of the performance of the matching, figure 2 shows the ratios of the disaggregated hours allocated to household production (care, procurement, and core) between the imputed data (ASEC) and the donor data (ATUS). Table 7 provides additional information on the mean and median hours of household production per week. The information is shown across the five strata variables used for the matching. With some exceptions, the ratios of mean weekly hours of household production (and subcategories) fall within 5 percent of difference across all strata variables, an indication of good match quality. The largest differences are observed among low-income households and among people with less than a high school education. In both cases the statistics indicate, on average, 11.1 percent and 12 percent more hours respectively allocated to household production. In perspective, while such differences seem large, they might have a small effect on other analyses since the average hours allocated to care among the specific groups are rather few (just about two hours).

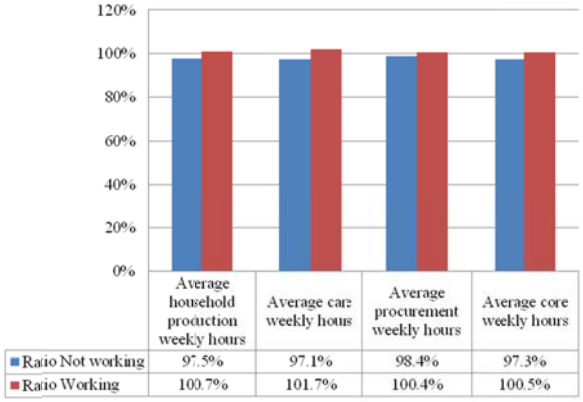
Figure 2. Ratio of Mean Household Production Hours, by Strata Variables



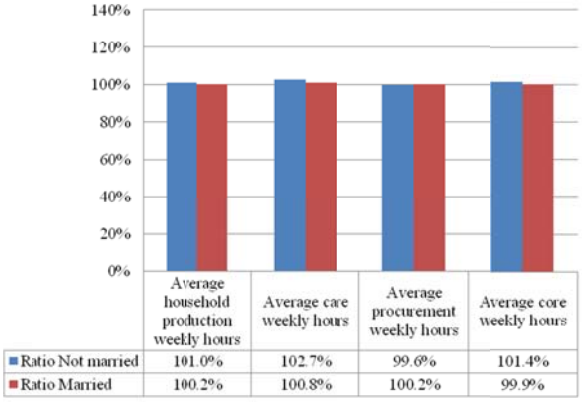
	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio Female	100.1%	100.0%	100.2%	100.1%
Ratio Male	99.6%	99.6%	99.3%	99.7%



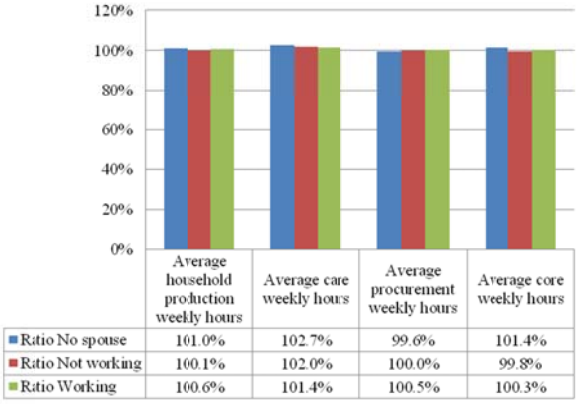
	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio Non-parent	100.1%	102.3%	99.8%	100.1%
Ratio Parent	99.1%	97.7%	100.2%	99.6%



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio Not working	97.5%	97.1%	98.4%	97.3%
Ratio Working	100.7%	101.7%	100.4%	100.5%



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio Not married	101.0%	102.7%	99.6%	101.4%
Ratio Married	100.2%	100.8%	100.2%	99.9%



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio No spouse	101.0%	102.7%	99.6%	101.4%
Ratio Not working	100.1%	102.0%	100.0%	99.8%
Ratio Working	100.6%	101.4%	100.5%	100.3%

Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

Table 7. Average and Median Household Production Weekly Hours, by Selected Variables

	Averages			Median		
	Donor	Recipient	Ratio	Donor	Recipient	Ratio
Hours of household production	21.87	21.86	100.0%	16.0	15.9	99.5%
Care	3.71	3.7	99.7%	0.0	0.0	
Procurement	5.25	5.24	99.8%	0.0	0.0	
Core	12.92	12.92	100.0%	7	6.97	99.6%
Marital status						
Not married	17.06	17.23	101.0%	10.5	10.5	100.0%
Married	25.55	25.6	100.2%	20.42	20.42	100.0%
Parental status						
Nonparent	18.07	18.09	100.1%	12.3	12.3	100.0%
Parent	28.72	28.46	99.1%	23.33	23.33	100.0%
Sex						
Female	26.77	26.8	100.1%	22.2	22.2	100.0%
Male	16.67	16.6	99.6%	10.5	10.5	100.0%
Labor status						
Not working	25.63	24.99	97.5%	21.0	20.4	97.2%
Working	19.45	19.58	100.7%	13.8	14.0	101.7%
Spouse's labor status						
No spouse	17.06	17.23	101.0%	10.5	10.5	100.0%
Not working	22.45	22.47	100.1%	17.6	17.6	100.0%
Working	27.19	27.35	100.6%	21.6	21.6	100.0%
Education						
Less than high school	17.61	19.73	112.0%	10.5	12.8	122.2%
High school	22.7	22.13	97.5%	17.2	16.9	98.7%
Some college	20.56	20.81	101.2%	14.0	14.0	100.0%
College grad	23.62	23.13	97.9%	18.1	17.5	96.8%
Household Income ($)						
0–14,999	21	23.33	111.1%	14.7	17.5	119.0%
15,000–34,999	22.21	22.39	100.8%	17.5	16.7	95.3%
35,000–49,999	20.78	21.57	103.8%	14.0	15.5	110.9%
50,000–74,999	23.03	21.81	94.7%	17.5	16.3	93.3%
75,000+	21.77	21.42	98.4%	15.8	15.2	96.3%
Age group						
15 to 24	12.18	12.62	103.6%	5.8	5.8	100.0%
25 to 34	23.69	23.22	98.0%	17.5	17.5	100.0%
35 to 44	26.64	26.48	99.4%	20.5	20.4	99.5%
45 to 54	22.83	22.95	100.5%	17.5	17.5	100.0%
55 to 64	22.24	21.83	98.2%	17.5	16.9	96.7%
65 and older	24.07	24.01	99.8%	21.0	21.0	100.0%

Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

Table 8 presents additional details on the quality of the match using the cell matching variable. Similar to the results described before, with some exceptions, total household production—in particular procurement and core hours—shows good levels of balance across most of the matching cells (note: procurement and core hours are part of household production). Some of the largest differences are observed for cells 1, 3, 5, 13, and 14, with a difference larger than 20 percent in relative terms in terms of care activities. The imputed sample overestimates the allocation of hours in care activities, but it represents a less-than-one-hour difference. These cells are the ones that had the lowest rate of first-round matching, which could explain these results. In general, it seems that after the statistical match, the imputed sample tends to understate the average hours in household production, but such differences are somewhat small.

Table 8. Ratio and Absolute Differences of Mean Household Production Hours, by Matching Cell

Cell	Sex	Parent status	Labor status	Spouse's status	Average household production weekly hours ratio (abs diff)		Average care weekly hours ratio (abs diff)		Average procurement weekly hours ratio (abs diff)		Average core weekly hours ratio (abs diff)	
C1	W	N	Not working	No	102%	0.39hrs	122%	0.2hrs	101%	0.03hrs	101%	0.16hrs
C2	W	N	Not working	Not working	100%	0.04hrs	101%	0.01hrs	100%	0.02hrs	100%	0.01hrs
C3	W	N	Not working	Working	98%	0.56hrs	80%	0.67hrs	98%	0.18hrs	101%	0.29hrs
C4	W	N	Working	No	102%	0.39hrs	100%	0hrs	103%	0.19hrs	102%	0.2hrs
C5	W	N	Working	Not working	105%	0.9hrs	158%	0.44hrs	101%	0.07hrs	103%	0.38hrs
C6	W	N	Working	Working	100%	0.06hrs	88%	0.12hrs	100%	0.02hrs	101%	0.2hrs
C7	W	Y	Not working	No	98%	0.71hrs	94%	0.63hrs	103%	0.15hrs	99%	0.23hrs
C8	W	Y	Not working	Not working	89%	4.13hrs	91%	0.74hrs	87%	0.95hrs	90%	2.44hrs
C9	W	Y	Not working	Working	99%	0.41hrs	97%	0.52hrs	104%	0.31hrs	99%	0.2hrs
C10	W	Y	Working	No	94%	1.7hrs	91%	0.71hrs	96%	0.24hrs	95%	0.75hrs
C11	W	Y	Working	Not working	99%	0.26hrs	100%	0.01hrs	99%	0.05hrs	98%	0.23hrs
C12	W	Y	Working	Working	100%	0.15hrs	101%	0.09hrs	100%	0.03hrs	99%	0.22hrs
C13	M	N	Not working	No	102%	0.2hrs	136%	0.16hrs	98%	0.07hrs	101%	0.12hrs
C14	M	N	Not working	Not working	94%	1.12hrs	124%	0.24hrs	97%	0.15hrs	91%	1.21hrs
C15	M	N	Not working	Working	100%	0.07hrs	100%	0hrs	101%	0.08hrs	99%	0.15hrs
C16	M	N	Working	No	100%	0.04hrs	97%	0.01hrs	99%	0.03hrs	101%	0.08hrs
C17	M	N	Working	Not working	104%	0.52hrs	98%	0.02hrs	102%	0.1hrs	105%	0.44hrs
C18	M	N	Working	Working	98%	0.3hrs	90%	0.09hrs	100%	0.01hrs	98%	0.22hrs
C19	M	Y	Not working	No	92%	1.74hrs	93%	0.38hrs	86%	0.65hrs	94%	0.71hrs
C20	M	Y	Not working	Not working	88%	3.07hrs	85%	0.89hrs	95%	0.25hrs	87%	1.92hrs
C21	M	Y	Not working	Working	96%	1.23hrs	92%	0.86hrs	97%	0.14hrs	98%	0.25hrs
C22	M	Y	Working	No	99%	0.26hrs	97%	0.11hrs	93%	0.44hrs	102%	0.28hrs
C23	M	Y	Working	Not working	103%	0.56hrs	103%	0.19hrs	103%	0.16hrs	103%	0.21hrs
C24	M	Y	Working	Working	100%	0.07hrs	99%	0.05hrs	101%	0.05hrs	101%	0.08hrs

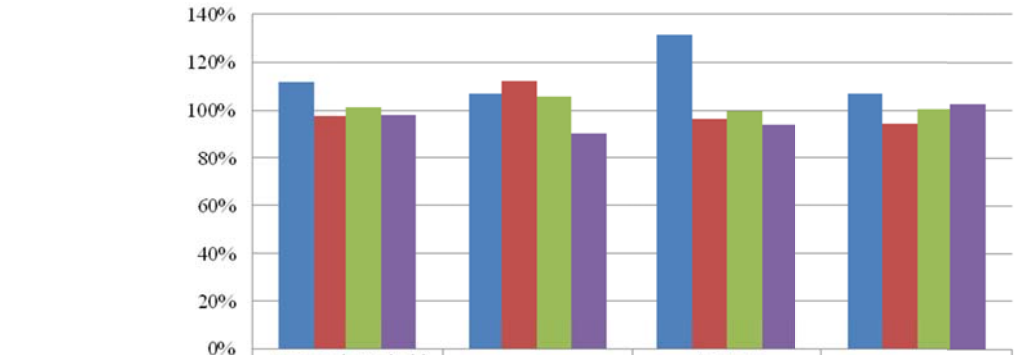
Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

To examine the match quality beyond the framework of the strata variables, figure 3 presents information on ratios for household production and its components across education, household income level, and age group. In addition, table 7 provides the mean and median of total household production for selected variables. In terms of education, people with high school and some college education have good levels of balance between both surveys. People with less than

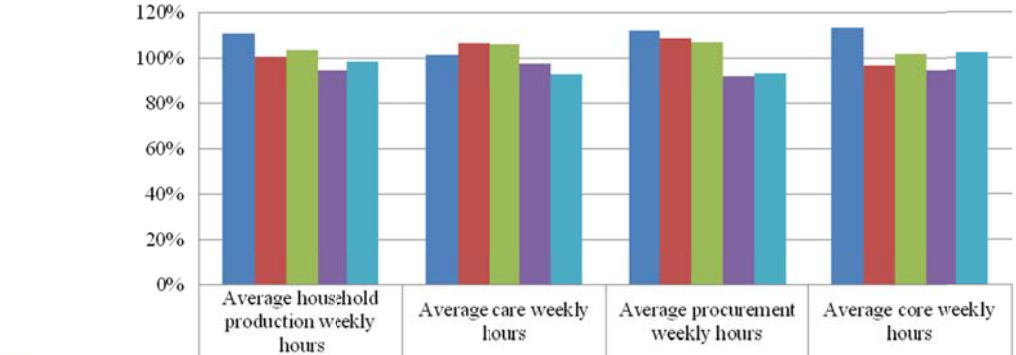
a high school education are imputed with longer hours allocated to household production (2.1 hours more) and all its components. In contrast, there is a small but consistent underestimation of the hours of household production (0.5 hours) for people with at least a college degree. Individuals in the lowest income groups show an underestimation of the hours allocated to household production (2.3 hours), a bias that is particularly large when observing the hours assigned to care and core activities. In contrast, individuals living in the richest households exhibit a somewhat consistent underestimation. Similar gaps are observed when looking at the medians.

In terms of age, the averages and medians indicate the statistical match did a good job imputing hours, as the differences are small for all groups. Looking at care activities, however, the statistical match seems to overestimate the number of hours spent on these activities, especially for people over 55 years of age.

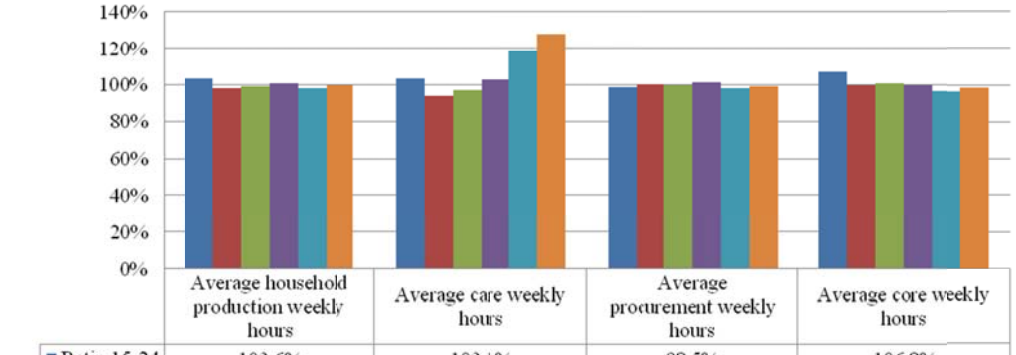
Figure 3. Ratio of Mean Household Production Hours, by Selected Variables



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio Less than highschool	112.0%	107.0%	131.4%	106.8%
Ratio Highschool	97.5%	112.5%	96.6%	94.6%
Ratio Some college	101.2%	105.9%	99.8%	100.5%
Ratio College or higher	97.9%	90.2%	94.1%	102.4%



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio 0-14,999	111.1%	101.4%	112.4%	113.4%
Ratio 15,000-34,999	100.8%	106.8%	108.7%	96.7%
Ratio 35,000-49,999	103.8%	106.4%	107.2%	101.9%
Ratio 50,000-74,999	94.7%	97.9%	92.1%	94.9%
Ratio 75,000+	98.4%	93.1%	93.5%	102.6%



	Average household production weekly hours	Average care weekly hours	Average procurement weekly hours	Average core weekly hours
Ratio 15-24	103.6%	103.1%	98.5%	106.8%
Ratio 25-34	98.0%	94.1%	100.4%	99.5%
Ratio 35-44	99.4%	97.0%	99.8%	100.7%
Ratio 45-54	100.5%	102.8%	101.3%	99.9%
Ratio 55-64	98.2%	118.5%	98.2%	96.4%
Ratio 65+	99.8%	127.9%	99.3%	98.8%

Source: Author's calculations based on ASEC 2014 and ATUS 2013 data.

3.2. Matching: ASEC 2014 and SCF 2013

For the matching process between the ASEC 2014 and SCF 2013, five strata variables, namely income categories, homeownership, family type, and race and age of the householder (head of household), are used to create 360 matching cells. Given the availability of information from both surveys within each cell, and the requirements imposed for consistent estimation of the propensity scores via logit models, we end up with 162 cells in the first round, which represent about 92 percent of the whole sample.[10]

A dummy variable indicating whether the observation corresponds to the donor or the recipient survey is used as the dependent variable. In addition to the strata variables, a set of variables including dummies for zero income, zero wage income, dummies for other sources of income, age (and its square) of the householder, educational attainment, occupation category, and number of people in household are included in the model specification. Standardized indexes for income and wage income are also included. The logit models and propensity scores are estimated using all information within broader cells, but the matching is elaborated only across observations left unmatched from previous rounds. For subsequent matching rounds, broader matching cells are defined accordingly, keeping the logit specifications consistent across all models, and including the omitted strata variable in the specification

Turning to the results of the match performance, table 9 shows the distribution of the matched records by matching round. As expected, a large share of the matches (81.4 percent) occurs on the first round, when the highest level of quality match is ensured. While in the first round the match ratio is lower than in the previous case (ATUS-ASEC), it is still sufficiently large to obtain good matching quality in terms of the strata variables. Only 0.7 percent of the weighted sample is left unmatched after all matching rounds. These unmatched observations are composed of middle-to-high-income renter households, with a mostly nonelderly and predominately Hispanic or white householder. This should not bias the distributional statistics of the transferred information in the aggregate.

[10] For each cell, a minimum of ten observations from both surveys are require to proceed with the estimation of the propensity score. At the same time, in cases where the logit model indicates perfect prediction of outcomes, the respective observations are excluded from the calculation of the propensity scores.

Table 9. Distribution of Matched Records by Matching Round

Matching round	Records matched	Percent	Cumulative percent
1	100,052,472	81.38	81.38
2	4,388,410	3.57	84.94
3	3,307,982	2.69	87.63
4	2,349,275	1.91	89.55
5	665,937	0.54	90.09
6	895,562	0.73	90.82
7	431,845	0.35	91.17
8	1,444,297	1.17	92.34
9	15,546	0.01	92.35
10	5,705,878	4.64	96.99
11	546,977	0.44	97.44
12	1,935,682	1.57	99.01
13	95,399	0.08	99.09
14	23,420	0.02	99.11
15	264,749	0.22	99.33
16	828,494	0.67	100.00
Total	118,682,616		

Source: Author's calculations based on ASEC 2014 and SCF 2013 matched data.

Table 10 provides a better look at the match quality, comparing some distributional statistics on a household's assets and liabilities. Table 10 also presents some statistics on individual asset and debt categories.[11] The upper percentiles and Gini coefficients are equivalent across both samples (0.874). The lower percentiles, however, present a more pronounced difference, with the ASEC presenting lower net worth estimates. This is related to differences in the incidence of homeowners with mortgages shown in table 3. The differences in the percentiles are also replicated when looking at the percentile ratios. The means and medians show a fair level of equivalence between both surveys for the disaggregated components. The largest difference corresponds to asset3 (liquid assets), showing an average difference of 4 percent, or about $2,121.

[11] Assets are classified in gross value of housing (asset1); value of real estate and unincorporated businesses (asset2); liquid assets (checking, saving, cash, etc.) (asset3); total directly held mutual funds (asset4); individual retirement accounts and thrift-type plans (asset5). Similarly, debts are classified in housing debt (debt1) and other debt (debt2).

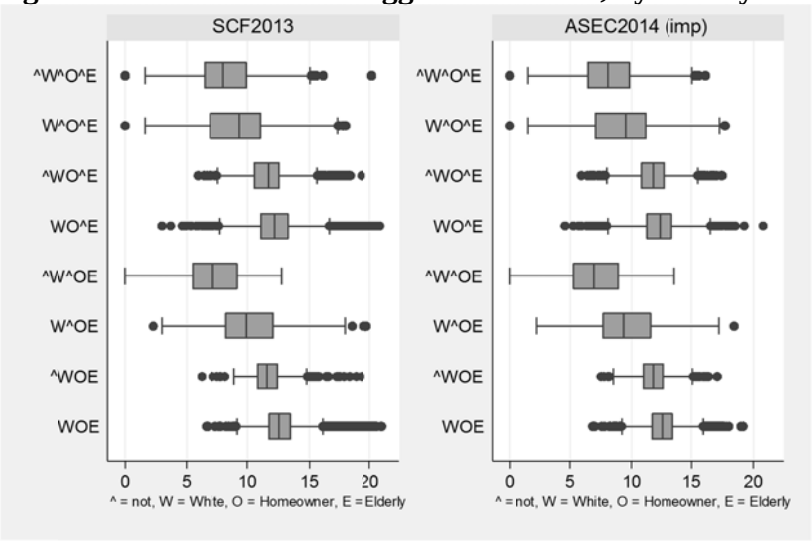
Table 10. Matching Quality: Summary Statistics

	SCF2013	ASEC 2014	Ratio ASEC/SCF
Distributional statistics (net worth)			
p10	($12,950)	($15,300)	118.1%
p25	$300	$100	33.3%
p50	$61,000	$62,200	102.0%
p75	$285,505	$286,500	100.3%
p90	$888,005	$862,800	97.2%
p90/p50	15	14	95.3%
p75/p25	952	2865	301.0%
p75/p50	5	5	98.4%
p50/p25	203	622	305.9%
Gini	0.874	0.875	1002%
Summary statistics			
Average asset1	$171,072	$169,634	99.2%
Average asset2	$163,889	$167,806	102.4%
Average asset3	$45,443	$43,322	95.3%
Average asset4	$106,872	$105,017	98.3%
Average asset5	$90,635	$88,675	97.8%
Average debt1	$67,229	$66,200	98.5%
Average debt2	$15,086	$14,430	95.7%
Average net worth	$495,597	$493,823	99.6%
Median asset1	$90,000	$90,000	100.0%
Median asset2	$0	$0	
Median asset3	$4,830	$4,650	96.3%
Median asset4	$0	$0	
Median asset5	$0	$0	
Median debt1	$0	$0	
Median debt2	$2,800	$2,650	94.6%
Median net worth	$61,000	$62,200	102.0%

Note: Assets are classified in gross value of housing (asset1); value of real estate and unincorporated businesses (asset2); liquid assets (asset3); total mutual funds (asset4); individual retirement accounts and thrift-type plans (assets5). Similarly, debts are classified in housing debt (debt1) and other debt (debt2).
Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

Figure 4 presents a visual representation of the distribution of logged household net worth using three of the strata variables: race, homeownership, and age. The figure shows that for most cases the distribution of the logged net worth is equivalent in both surveys. There are, however, some differences in the distributions regarding extreme values (outliers) among some groups, like households with white elderly homeowners (WOE), nonwhite elderly homeowners (^WOE), or white nonelderly and nonhomeowners (W^O^E). While extreme values might not affect statistics like medians and percentiles, they might create problems when analyzing information at the means for more detailed subgroups.

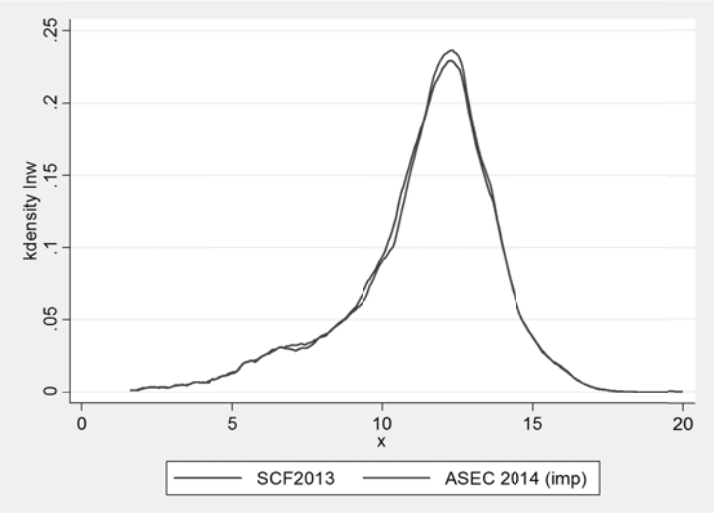
Figure 4. Distritubtion of Logged Net Worth, by Survey



Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

Figure 5 provides an alternative to comparing the distribution of logged net worth between both the donor and the imputed sample. The close superposition between the kernel densities for both suggests that, as indicated before, the moments of the distributions of the imputed and donor samples are highly comparable in the aggregate. A closer look at figure 5, however, still indicates that the presence of outliers might affect the estimation of relevant means for specific groups. Overall, there is a difference of only $1,774 between the mean imputed and donor net worth, and no differences when comparing medians.

Figure 5. Kernel Density of Logged Net Worth, by Survey

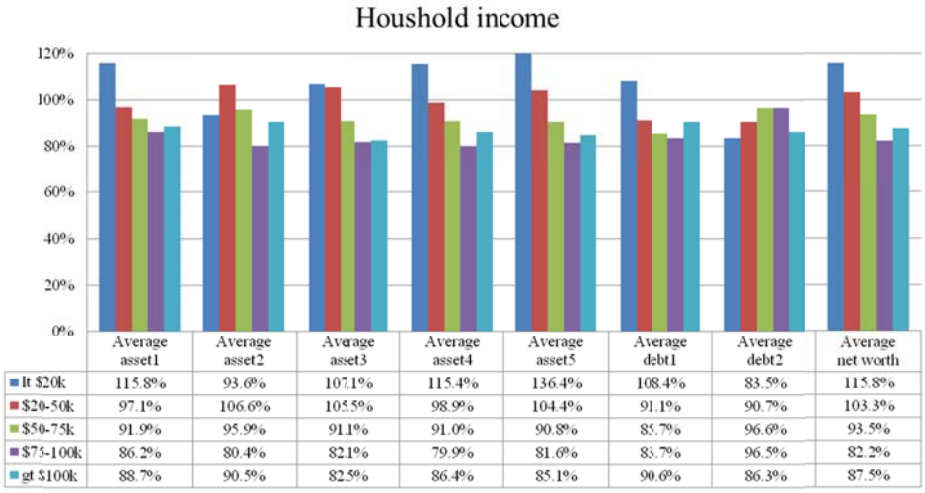


Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

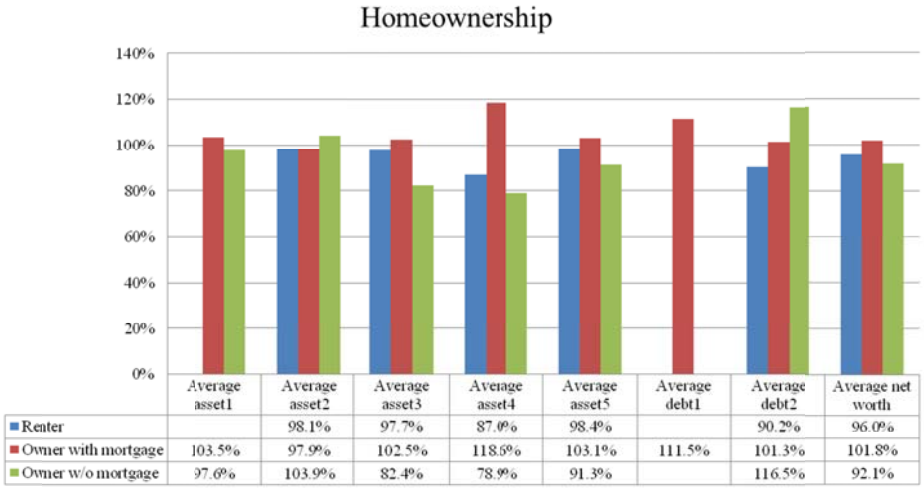
For a detailed review of the performance of the matching, figures 6 and 7 show the ratios of asset and debt values between the imputed data (ASEC) and the donor data (SCF) across the five strata variables used for the matching. Table 11 also presents information on the mean and median gaps of the net worth of the households with respect to the strata characteristics. The first strata variable to be analyzed corresponds to the household income. After the matching, the average values of asset1, asset4, asset5, and net worth are overstated (up to 36 percent) in the recipient dataset among households in the lowest income group. This implies a difference of a little more than $7,411 for asset1 or $11,140 for net worth. In contrast, with a few exceptions, all other assets/debts are understated in the imputed dataset by almost 10 percent on average, with the richest households having the largest bias (14 percent or $227,000 lower net worth). In all cases, debt1 and debt2 are understated for all income groups except the richest, with a bias of less than 15 percent.

With respect to homeownership, the results show a good balance, on average, with net worth differences ranging from $2,500 to $79,500. The groups with the largest imbalances correspond to: homeowners without a mortgage, for which mutual funds (asset4) are understated by almost 22 percent and other debt (debt2) is overstated by 16 percent; and homeowners with a mortgage, for which mortgage debt (debt1) is overstated by about 11 percent and mutual funds (asset4) are overstated by 18 percent. In terms of family type, while households with couples and single women have well-balanced statistics, real estate assets in single-male households are understated by 35 percent (asset2) and mutual funds (asset4) are overstated by 13 percent. In aggregate, net worth is understated by 3 percent ($9,350) (table 11).

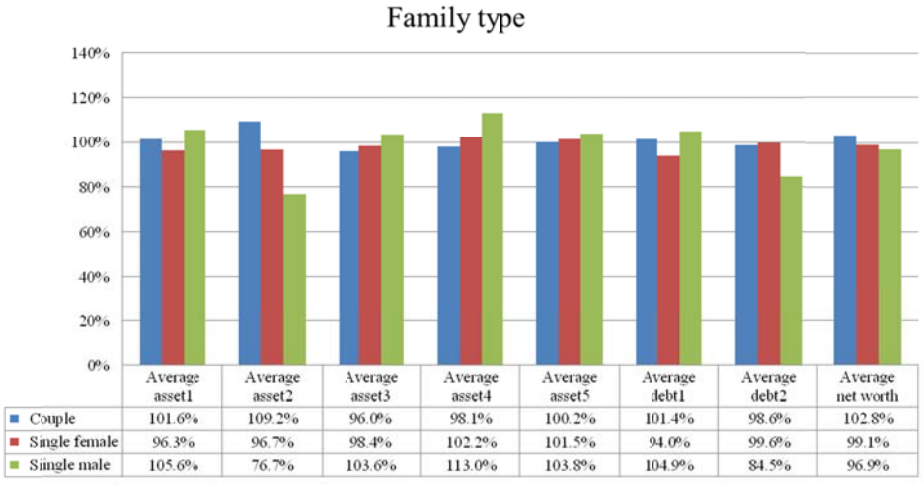
Figure 6. Ratio of Mean Household Assets and Liabilities, by Strata Variables



	Average asset1	Average asset2	Average asset3	Average asset4	Average asset5	Average debt1	Average debt2	Average net worth
lt $20k	115.8%	93.6%	107.1%	115.4%	136.4%	108.4%	83.5%	115.8%
$20-50k	97.1%	106.6%	105.5%	98.9%	104.4%	91.1%	90.7%	103.3%
$50-75k	91.9%	95.9%	91.1%	91.0%	90.8%	85.7%	96.6%	93.5%
$75-100k	86.2%	80.4%	82.1%	79.9%	81.6%	83.7%	96.5%	82.2%
gt $100k	88.7%	90.5%	82.5%	86.4%	85.1%	90.6%	86.3%	87.5%



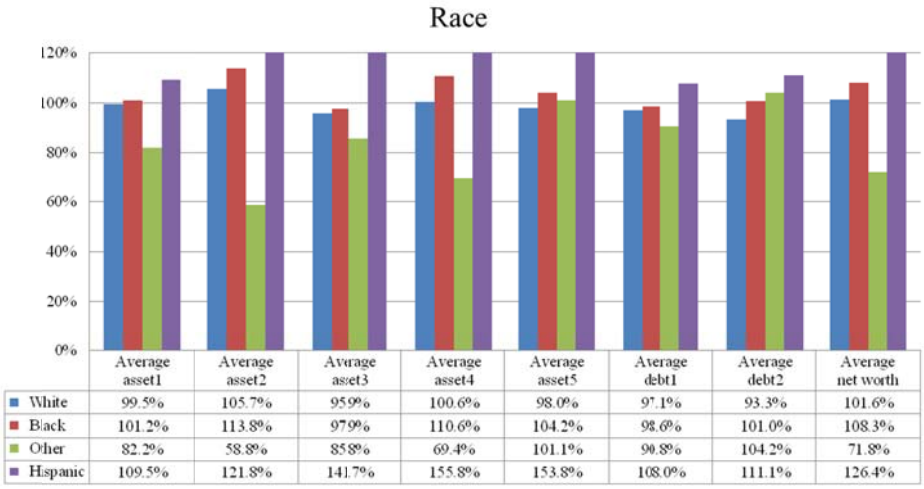
	Average asset1	Average asset2	Average asset3	Average asset4	Average asset5	Average debt1	Average debt2	Average net worth
Renter		98.1%	97.7%	87.0%	98.4%		90.2%	96.0%
Owner with mortgage	103.5%	97.9%	102.5%	118.6%	103.1%	111.5%	101.3%	101.8%
Owner w/o mortgage	97.6%	103.9%	82.4%	78.9%	91.3%		116.5%	92.1%



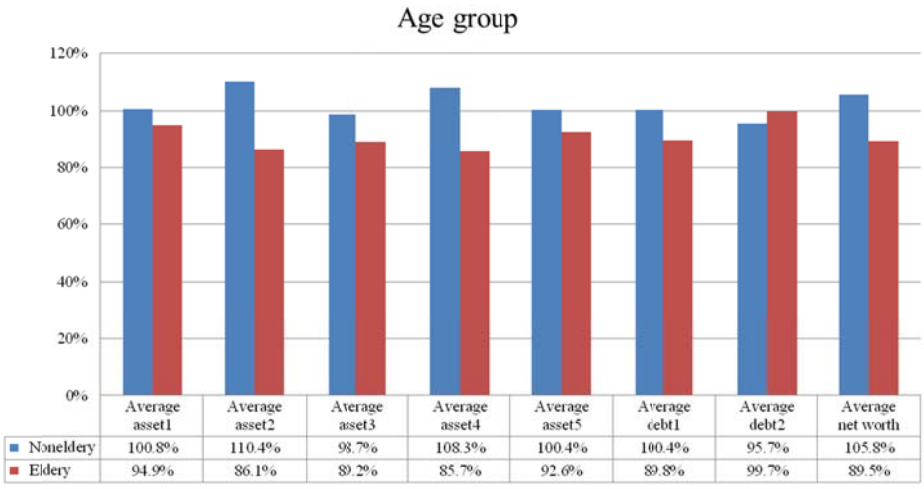
	Average asset1	Average asset2	Average asset3	Average asset4	Average asset5	Average debt1	Average debt2	Average net worth
Couple	101.6%	109.2%	96.0%	98.1%	100.2%	101.4%	98.6%	102.8%
Single female	96.3%	96.7%	98.4%	102.2%	101.5%	94.0%	99.6%	99.1%
Single male	105.6%	76.7%	103.6%	113.0%	103.8%	104.9%	84.5%	96.9%

Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

Figure 7. Ratio of Mean Household Assets and Liabilities, by Strata Variables



	Average asset1	Average asset2	Average asset3	Average asset4	Average asset5	Average debt1	Average debt2	Average net worth
White	99.5%	105.7%	95.9%	100.6%	98.0%	97.1%	93.3%	101.6%
Black	101.2%	113.8%	97.9%	110.6%	104.2%	98.6%	101.0%	108.3%
Other	82.2%	58.8%	85.8%	69.4%	101.1%	90.8%	104.2%	71.8%
Hispanic	109.5%	121.8%	141.7%	155.8%	153.8%	108.0%	111.1%	126.4%



	Average asset1	Average asset2	Average asset3	Average asset4	Average asset5	Average debt1	Average debt2	Average net worth
Noneldery	100.8%	110.4%	98.7%	108.3%	100.4%	100.4%	95.7%	105.8%
Eldery	94.9%	86.1%	89.2%	85.7%	92.6%	89.8%	99.7%	89.5%

Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

When considering race, while the balance statistics show that information corresponding to households with white, black, and Hispanic householders is well balanced, the imputed sample consistently understates the asset/debt holdings from *other* race households by almost 17 percent. In terms of net worth alone, the net assets of "other races" are understated in just over 28 percent of the cases, which implies an almost $151,731 difference. The median gaps show a much smaller absolute difference ($10,000), suggesting that the large differences in the means are driven by outliers. Finally, in terms of age groups, the statistics show that the imputed data is well balanced for most of the asset/debt components except for mortgage debt (debt1). The statistics show that the imputed data understates the debt of elderly households in about 11

percent of the cases. This happens because the share of elderly households with mortgage debt is lower in the ASEC survey compared to the corresponding share in the SCF.[12]

Table 11. Mean and Median Net Worth by Strata Variables

	Averages			Median		
	Donor	**Recipient**		**Donor**	**Recipient**	
Total	$495,597	$493,823	99.6%	$61,000	$62,200	102.0%
Homeownership						
Renter	$63,047	$60,544	96.0%	$60	$40	66.7%
Owner with mortgage	$584,821	$595,236	101.8%	$132,420	$124,782	94.2%
Owner w/o mortgage	$1,001,514	$922,011	92.1%	$245,100	$235,820	96.2%
Income group						
<$20k	$73,136	$81,658	111.7%	$1,000	$2,240	224.0%
$20–50k	$170,974	$138,667	81.1%	$24,930	$24,193	97.0%
$50–75k	$244,142	$239,509	98.1%	$76,240	$67,662	88.7%
$75–100k	$302,437	$315,707	104.4%	$154,520	$116,827	75.6%
> $100k	$1,769,339	$1,593,959	90.1%	$508,345	$379,621	74.7%
Age						
Nonelderly	$393,449	$414,352	105.3%	$31,400	$30,618	97.5%
Elder	$723,441	$733,506	101.4%	$194,651	$174,500	89.6%
Family type						
Couple	$653,716	$726,690	111.2%	$117,735	$125,305	106.4%
Single female	$181,994	$165,414	90.9%	$20,020	$13,865	69.3%
Single male	$252,558	$289,983	114.8%	$23,500	$24,093	102.5%
Ethnicity						
White	$596,808	$650,210	108.9%	$112,500	$114,582	101.9%
Black	$86,188	$85,626	99.3%	$1,470	$1,270	86.4%
Other	$489,367	$387,049	79.1%	$66,005	$55,970	84.8%
Hispanic	$90,887	$122,486	134.8%	$1,850	$2,000	108.1%

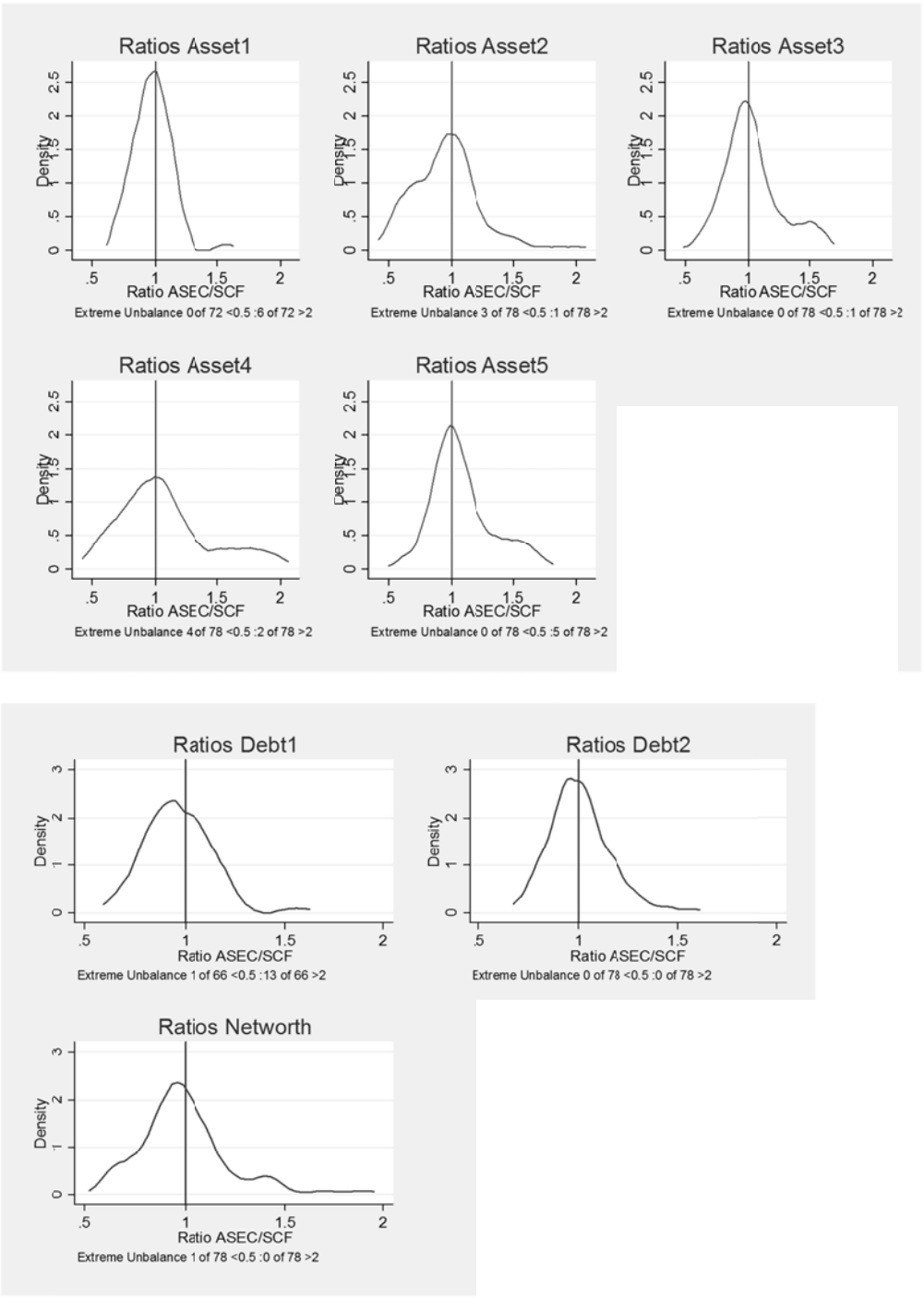
Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

To analyze how the matching performs for more detailed cells, the mean ratios between samples for all assets and debts are calculated for different combinations of the strata variables.[13] Figure 8 plots the densities corresponding to the mean ratios for selected combinations of the strata variables. As can be seen for most of the cases, the distributions of the mean ratios are highly concentrated around one, indicating that, on average, there is good balance between both surveys. As the figure also indicates, for some of the ratios, some large imbalances can be observed (ratios above two). These types of large imbalances for narrower cells are expected, as the SCF also collects information for high-income families, which might appear as large outliers. While for most variables the ratio distributions indicate a good balance, the ones

[12] While ignoring mortgage status as part of the strata variables improves the overall balance of debt1, it also assigns additional debts to households that should have no mortgage debt.
[13] The cell combinations include: race-homeownership, race-age group, race-family type, and race-income group.

corresponding to retirement assets (asset5) suggest that the imputed data tends to overstate it (25 percent).[14]

Figure 8. Kernel Density Ratios of Mean Household Assets and Liabilities



Source: Author's calculations based on ASEC 2014 and SCF 2013 data.

[14] It should be noticed that the level of bias is larger if the information were not to be redistributed.

4. CONCLUSIONS

Overall, the ATUS and ASEC data are well aligned, with the some imbalances with respect to labor force status. The matching quality is good, with some limitations. There is a strong balance across the individual strata variables, showing good balance for aggregate measures (household production) for most of the variables analyzed. The results across the individual matching cells and other variables, however, show less balance.

On the one hand, the imputed information on the hours allocated to care activities shows important (relative) imbalances across many matching cells. The absolute differences, however, are small and should not create a large bias. On the other hand, information across other variables, such as education, household income, and (particularly) age, show important balance problems. The imputed dataset overstates household production of people with less than a high school education, and understates it for those with tertiary education, as well as for people in poor households. Across age, while the aggregate results are balanced, the individual components show large over- and under-estimations for different age groups.

With respect to the SCF and ASEC, the data is also well aligned, with the exception of mortgage holding, with small differences in the proportions of the breakdown by the sex of the householder. The results regarding the quality of the match are mixed. While the overall results show good balance between the imputed and donor surveys, with small underestimations of some items, analyzing the results across the strata variables shows relatively large imbalances (up to 20 percent) for a relatively small subset of strata variables. As we would expect, larger imbalances are observed for narrower groupings. The data shows some underestimation of mortgage debt, probably caused by the differences in the alignment of household property (see table 3). Given that the SCF collects information from high-income households, it is possible that the information transferred from these observations has a strong influence on the cell-specific statistics. These results imply that careful consideration must be taken when making statistical inferences from certain populations. One can make inferences for the aggregate population, but attempting a similar analysis using two or more variables at the same time may carry too much bias to be informative.

REFERENCES

Kennickell, A. B. 2000. “Revisions to the Variance Estimation Procedure for the SCF.” Washington, DC: Board of Governors of the Federal Reserve System.

Kennickell, A. B., and R. L. Woodburn. 1999. “Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth.” *Review of Income and Wealth* 45(2): 193–215.

Kum, H., and T. N. Masterson. 2010. “Statistical matching using propensity scores: Theory and application to the analysis of the distribution of income and wealth.” *Journal of Economic and Social Measurement* 35(3): 177–96.

Semega, J., and E. Welniak, Jr. 2013. “Evaluating the 2013 CPS ASEC Income Redesign Content Test.” US Census Bureau Income Statistics Working Paper. Washington, DC: US Census Bureau.

Wolff, E. N., and A. Zacharias. 2003. “The Levy Institute Measure of Economic Well-Being.” Levy Institute Working Paper 372. Annandale-on-Hudson, NY: Levy Economics Institute of Bard College.