Quality of Match for Statistical Matches for the US

American Time Use Survey 2019, the Survey of Consumer Finances 2019, and the Annual Social and Economic Supplement 2020

This paper describes the quality of the statistical matching between the March 2020 supplement to the Current Population Survey (CPS) and the 2019 American Time Use Survey (ATUS) and Survey of Consumer Finances (SCF), which are used as the basis for the 2019 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.

Matching: ASEC ?meta:asec.year and SCF ?meta:scf.year

For the matching process between the ASEC **?meta:asec.year** and SCF **?meta:scf.year**, we consider 5 variables, namely income categories, homeownership, family type, and race and age of the householder (head of household), as the most important to evaluate the quality of the matching.

For the estimation of the propensity score, in addition to the previously mentioned 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. They are used to create the propensity scores and principal components to elaborate the matching. Standardized indexes for income and wage income are also included.

Turning to the results of the match performance, Figure 1 shows the distribution of the matched records by matching round. As expected, a large share of the matches (51.3 percent) occurs on the first round, when the highest level of quality match is ensured. This is sufficiently large to obtain good matching quality in terms of the strata variables. Only 6.7 percent of the weighted sample is matched in the last round 3 rounds. This should not bias the distributional statistics of the transferred information in the aggregate.
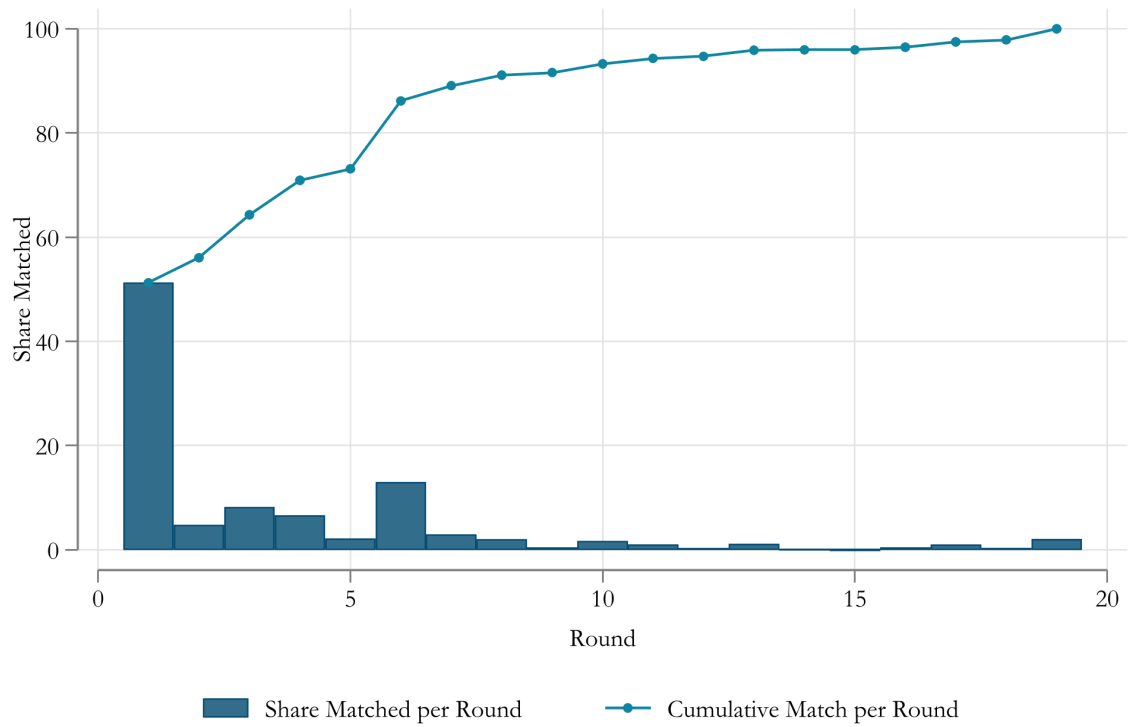
Figure 1: Distribution of Matched Records by Matching Round

Table 1 provides a better look at the match quality, comparing some distributional statistics on a household's assets and liabilities. Table 1 also presents some statistics on total assets and total debt. The upper percentiles and Gini coefficients are equivalent across both samples (**?meta:v9** 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 **?@tbl-t3**. 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 **?meta:v9** 4 percent, or about **?meta:v9** $2,121.

Table 1: Matching Quality: Summary Statistics

	ASEC	SCF	Ratio ASEC/SCF
Distributional Statistics			
Networth			
p10	-10,761	-11,980	89.8
p25	2,000	1,481	135.0
p50	102,000	96,198	106.0
p75	377,000	367,300	102.6
p90	1,150,105	1,126,855	102.1
Gini	0.867	0.872	99.4
Summary Statistics			
Avg. Total Assets	802,317	804,556	99.7
Median Total Assets	205,000	198,505	103.3
Avg. Total Debt	97,781	97,435	100.4
Median Total Debt	25,000	24,000	104.2
Avg. Networth	704,536	707,121	99.6
Median Networth	102,000	96,198	106.0

Figure 2 presents a visual representation of the distribution of logged household net worth using three variables: race, homeownership, and age, using kernel density curves. Because there is a substantial share of the population who has zero or negative networth, we plot both sides of the distribution using a slight modification of the log transformation:

$$log\_networth = log(networth) \qquad if \ networth > 0$$
$$log\_networth = 0 \qquad if \ networth = 0$$
$$log\_networth = -log(-networth) \quad if \ networth < 0$$

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 (WHE), nonwhite elderly homeowners ($\hat{}$WHE), or white nonelderly and nonhomeowners (W^{H}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. However, because the overall distribution is quite similar between donor and recipient sample across subgroups, the matching should suffice for less detailed analysis of the synthetic data.

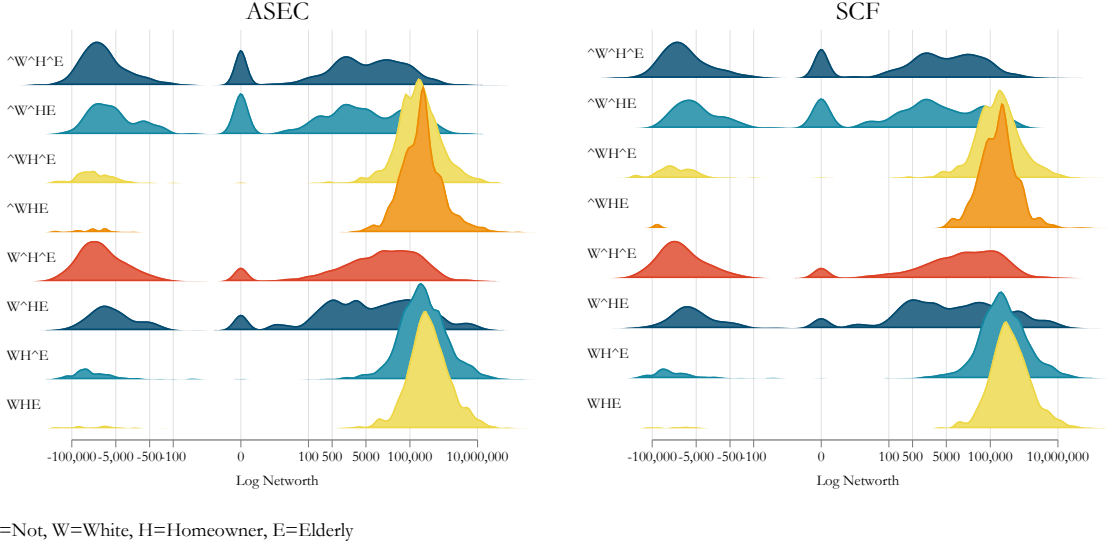$\hat{}$=Not, W=White, H=Homeowner, E=Elderly

Figure 2: Distribution of Logged Net Worth, by Survey

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 five variables used for the matching. Table 2 also presents information on the mean and median gaps of the net worth of the households with respect to the strata characteristics.

Table 2: Average and Median Networth, by Selected Variables

	Mean - ASEC	Ratio to SCF	Median - ASEC	Ratio to SCF
Total				
	704,536	99.6	102,000	106.0
Homeownership				
Renter	79,151	96.3	180	94.7
Owner with Mortgage	922,951	107.9	200,100	102.6
Owner w/o Mortgage	1,139,367	81.7	255,060	86.5
Income Group				
<$20k	166,951	232.7	16,290	1629.0

4

|	Mean - ASEC	Ratio to SCF	Median - ASEC	Ratio to SCF
$20k-$50k	307,902	182.9	56,700	143.0
$50k-$75k	294,215	112.9	58,300	68.3
$75k-$100k	413,844	117.6	98,985	66.0
>$100k	1,593,584	73.7	288,400	56.5
Age				
Non-Eldery	582,776	101.0	66,650	111.5
Elder	1,002,769	95.4	220,301	94.5
Family type				
Couple	1,039,985	99.8	172,400	103.2
Single female	277,320	109.7	39,022	118.2
Single male	315,472	99.0	50,790	120.4
Race				
White	868,515	96.1	156,350	102.2
Black	302,369	250.8	11,200	138.3
Hispanic	262,167	152.4	21,500	143.3
Other	752,871	74.5	99,400	61.3

The first 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 **?meta:v9** $11,140 for net worth. In contrast, with a few exceptions, all other assets/debts are understated in the imputed dataset by almost **?meta:v9** 10 percent on average, with the richest households having the largest bias (**?meta:v9** 14 percent or **?meta:v9** $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 **?meta:v9** $2,500 to **?meta:v9** $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 **?meta:v9** 3 percent (**?meta:v9** $9,350) (Table 2).
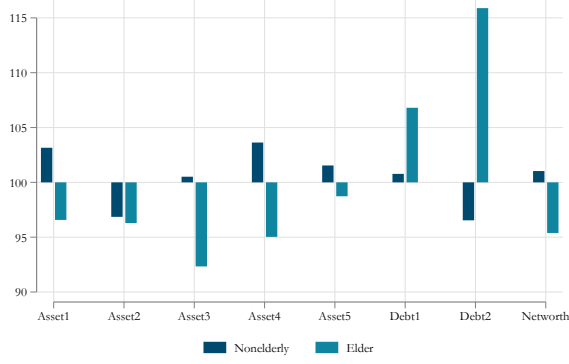
When considering race, while the balance statistics show that information corresponding to households with white, black, and Hispanic householders is well balanced, the imputed sample consistently understates the asset/debt holdings from other race households by almost
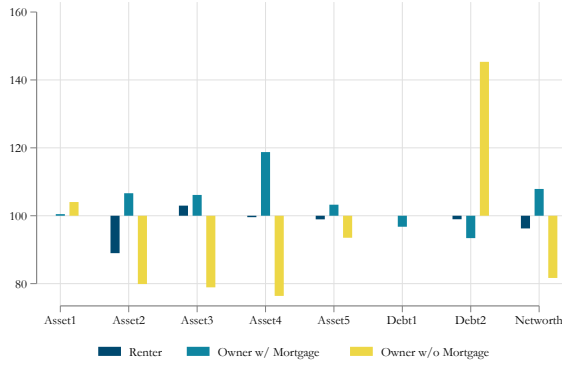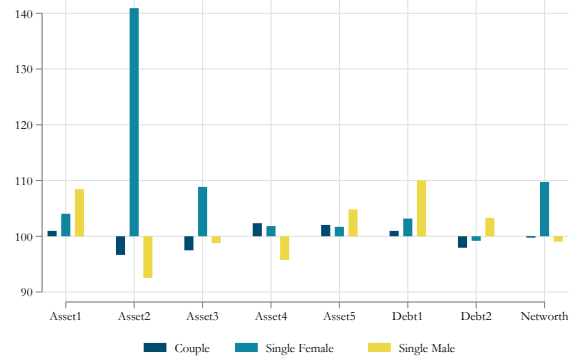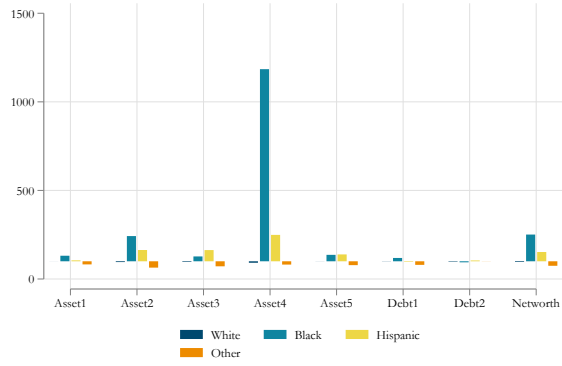
(a) Homeownership
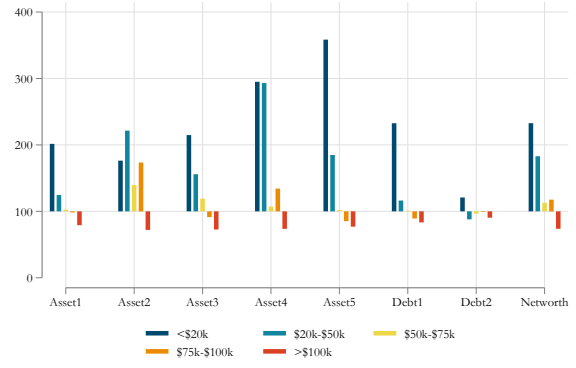
(b) Income group

(c) Age

(d) Family type

(e) Race

Figure 3: Ratio of Mean Household Assets and Liability by Main Variables

?meta:v9 17 percent. In terms of net worth alone, the net assets of "other races" are understated in just over **?meta:v9** 28 percent of the cases, which implies an almost **?meta:v9** $151,731 difference. The median gaps show a much smaller absolute difference (**?meta:v9** $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.

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. @fig2p4 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 corresponding to retirement assets (asset5) suggest that the imputed data tends to overstate it (25 percent).
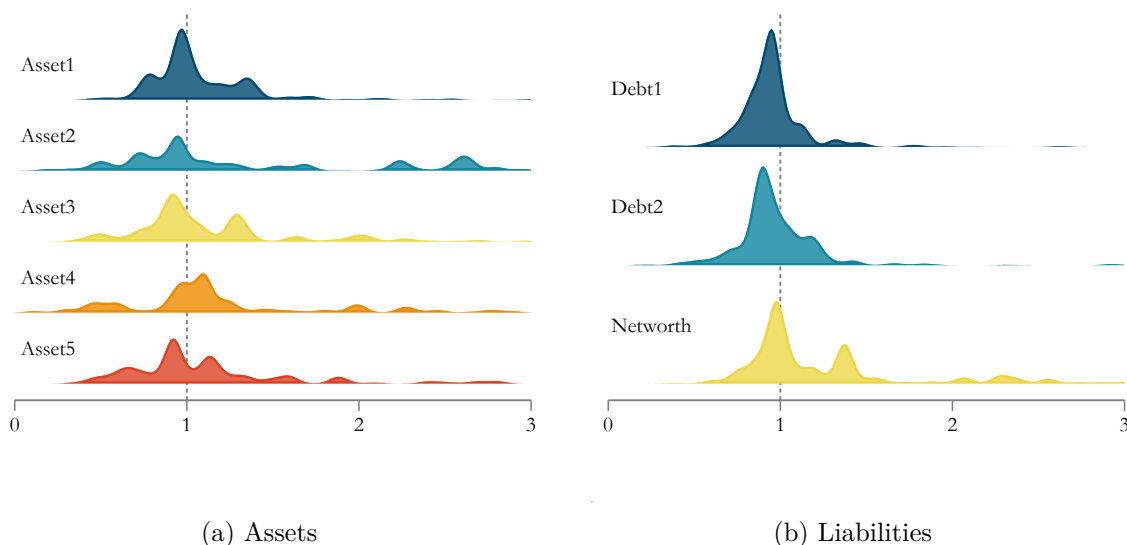

(a) Assets

(b) Liabilities

Figure 4: Kernel Density Ratios of Mean Household Assets and Liabilities

Figure 5: Homeownership