



Using Survey Data for Tax Analysis

December 2021

Overview

For tax analysis, the Congressional Budget Office uses information from a sample of tax returns supplemented with survey data.

Tax returns have detailed information about income from taxable sources but limited information on demographic characteristics. The Current Population Survey (CPS) has broad demographic information but limited information on income and expenditures. The two sources lead to different population, income, and tax liability estimates.

In this document, CBO uses the earned income tax credit (EITC) and the child tax credit (CTC) to illustrate the types of policies for which CPS-based estimates may be used to approximate those from tax returns. Simulated EITC benefits based on the CPS are substantially lower than totals from the tax data, but CPS-based estimates of the CTC are much closer to totals from the tax data.

Estimates using merged data from the two sources are more similar to results using only the tax data than results from using the CPS alone. However, the merged data overstate the share of benefits to unmarried male taxpayers relative to what either underlying data set shows.

Data Sources That CBO Uses and Their Strengths and Weaknesses

CBO Uses Various Data Sources for Its Analyses

CBO uses data from a variety of sources to produce its 10-year baseline projections and other analyses. Information from administrative and survey-based data sources are used to help assess the state of the economy, projected spending and revenues, and the effects of alternative policies.

No single data source contains all the necessary information, so CBO frequently needs to merge information from multiple sources. One complication of that approach is that alternative sources of data can sometimes appear inconsistent. For example, they may use different units of observation or they may define the same concepts differently.

To reconcile those differences, CBO considers which data sources are best suited to its analysis and uses its judgment, based on all available information, to adjust the data when necessary.



CBO Uses Data From Individual Income Tax Returns and CPS Data for Its Tax Analyses

To analyze the tax system and the distribution of household income, CBO uses a sample of tax returns from the Internal Revenue Service's (IRS's) Statistics of Income (SOI). Those core data are supplemented with household survey data from the Census Bureau's March CPS (the Annual Social and Economic Supplement).

Individual income tax returns have detailed information about income from taxable sources, but they have limited information on demographic characteristics and are not required from people who receive very low income and thus do not file taxes.¹ The March CPS is nationally representative and has a broad set of demographic information; its information on income and expenditures necessary to estimate taxes is more limited, though.

¹ Some people who would not otherwise be required to file a tax return choose to do so to claim a refund. For example, people who did not receive the full economic impact payments for which they were eligible could claim the remainder by filing a tax return; that is likely to increase the number of filers in 2021 and 2022.

CBO Merges the CPS Data With the Tax Return Data

CBO constructs tax-filing units in the CPS on the basis of relationship and income information. Those units are statistically matched to records in the SOI by using a set of common demographic characteristics and by ranking the data by income. The resulting data set represents the noninstitutionalized civilian population in the United States.

- Tax-filing units constructed from the CPS that are matched to the SOI represent *tax filers*. For those records, tax returns provide information on taxable income and the CPS provides information on transfer income.
- Unmatched units represent *tax nonfilers*. The CPS is the source of data for all income for nonfilers.

The CPS includes data on race and ethnicity, but the merged data may not preserve relationships between those variables and outcomes of interest (such as tax liability after credits and income).

Using Tax Return Data for Tax Analysis Has Strengths

SOI data come from income tax returns that are filed by individuals who have income above a threshold or who claim a refund of withheld taxes. The information collected is used to determine tax liability.

- A tax-filing unit consists of an individual, his or her spouse, and any dependents.
- Income information consists of annual income from taxable sources (such as wages and business income). Some types of taxable income (capital gains or partnership income, for instance) are not reported in the CPS.
- Demographic information includes marital status. Age and sex can be linked through a filer's Social Security number.

The weighted sample of tax returns is representative of all filers. Sampling rates vary by income group; the highest rates are for filers with very low income (large losses) or very high income.

Using Tax Return Data for Tax Analysis Has Weaknesses

The information collected in tax returns may not be well-suited to certain types of analyses:

- Distributional analyses by other demographic characteristics, such as race and ethnicity; or
- Analyses of policies that affect nonfilers, such as economic stimulus payments.

Using CPS Data for Tax Analysis Has Strengths

The CPS collects extensive information about demographic characteristics and also provides much of the information necessary to estimate taxes.

- It groups individuals into families and households.
- It provides information on current family and household characteristics as well as income and employment during the previous year.
- It includes a broad set of demographic information for each individual, such as age, sex, race, ethnicity, and educational attainment.
- It includes people with low income who are not required to file taxes.

Using CPS Data for Tax Analysis Has Weaknesses

The CPS may not have all the information necessary to estimate taxes.

- Not all sources of taxable income are reported (for example, capital gains and partnership income). Because a disproportionate share of income from those sources goes to higher-income taxpayers, accurately measuring income for high-income taxpayers using CPS data alone may be difficult.
- It lacks information on expenditures that can be deducted from income (such as charitable contributions). Higher-income taxpayers receive the largest benefits from those and other itemized deductions.
- Analysts must construct the tax-filing units, and not all eligibility requirements for head-of-household filing status and dependents are reported. Family structure and the presence of dependents affect tax liability and eligibility for various credits.

Survey-Based Methods Are Better-Suited for Analyzing Certain Types of Policies Than Other Types

For analyzing some types of tax policies, survey-based methods can approximate tax-based estimates. That approximation can be illustrated by comparing estimates of the EITC and CTC.

- SOI data can provide a target benchmark for CPS estimates.
- Credit amounts are based on broad measures of income and demographic information that are reported in the CPS. (Estimating other tax expenditures that apply to specific sources of income or expenditures would require additional imputations to the CPS.)
- For both tax credits, the distribution of benefits by sex clearly differs.

For each credit, eligibility and benefit amounts are determined differently; that affects how similar estimates based on the CPS are to estimates from the SOI.

- The EITC is narrowly targeted—only low-income workers are eligible, and workers with children receive the largest amounts.
- The CTC is more broad-based and less dependent on family structure. In 2021, the CTC is fully refundable and phases out for people with relatively high income.

Survey-Based Estimates May Differ From Tax-Based Estimates

Survey estimates are based on how researchers form tax-filing units. Those estimates can differ from SOI data for several reasons:

- People may misreport family relationships in the CPS.
- The data are from different points in time, so members of a CPS household and their marital status in March may differ from those who resided together in the previous year and their marital status at the end of the previous year.
- The tax-filing units created by analysts (on the basis of the IRS's rules) from CPS data could differ from the spouse and dependents listed by a taxpayer on his or her tax return.
- Taxpayers may incorrectly claim credits. The SOI contains returns that have not been audited, and some payments may go to ineligible taxpayers.
- Reported income may differ.

CBO Used Data for 2018 for This Analysis

For this analysis, CBO used the March 2019 CPS and a sample of tax returns from 2018.

- Tax-filing units were constructed in the CPS by grouping individuals on the basis of their family relationships, age, and income.
- Tax variables were constructed using CBO's microsimulation tax model.
- Taxes were simulated using 2018 law and the 2021 expansions of the child tax credit, EITC, and child and dependent care credit.

How the CPS Compares With Tax Data in Terms of Characteristics of the Population

CPS and Income Tax Return Samples Are Drawn From Different Populations

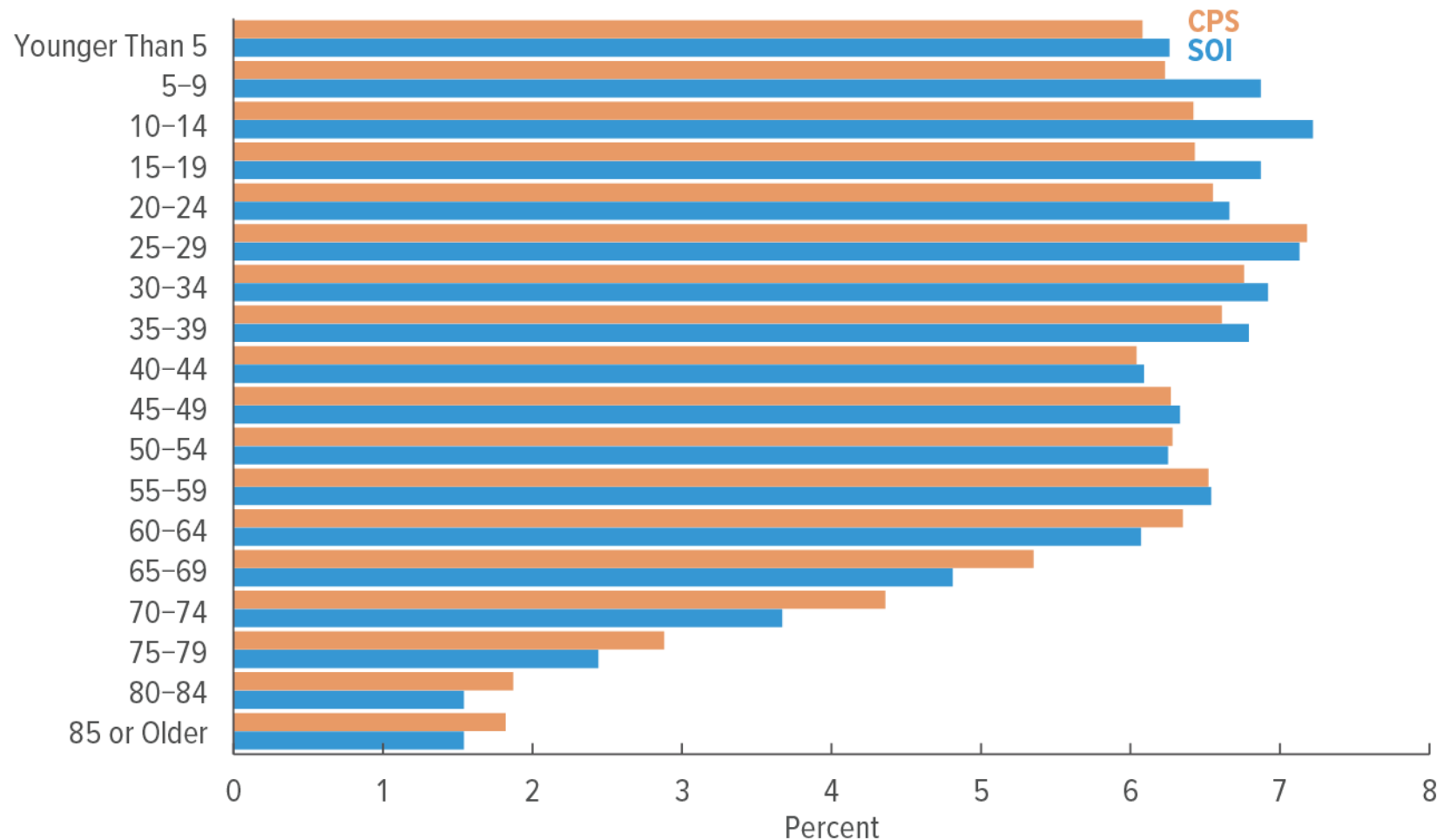
The CPS samples come from the civilian noninstitutionalized U.S. population. People with income above the filing threshold are required to file an income tax return. People who have income taxes withheld or who can claim a refund can also file.

The CPS includes people who do not file income tax returns and people who were born or became U.S. residents during the year.

The sample of income tax returns in the SOI includes returns from U.S. citizens living abroad, people who died during the year, and people living in institutions or on military bases.

The Population in the CPS Tends to Be Older Than the Population in the SOI

Distribution of Individuals, by Age Group



In 2018, data for 294 million people were reported in the SOI; in the CPS, data were reported for 324 million people. The SOI has data from a larger share of people under age 25 and a smaller share of people age 60 or older than the CPS does. In part, that is because many elderly people have little income other than Social Security benefits and need not file tax returns.

A Smaller Share of Tax-Filing Units in the CPS Are Unmarried Taxpayers With Dependents Compared With the SOI

Distribution of Tax-Filing Units, by Family Structure

	CPS		SOI	
	Number (Millions)	Share of Total (Percent)	Number (Millions)	Share of Total (Percent)
Married Couples With Dependents	30	18	27	19
Married Couples Without Dependents	34	21	28	19
Unmarried Women With Dependents	14	9	17	12
Unmarried Women Without Dependents	40	24	30	21
Unmarried Men With Dependents	6	4	8	6
Unmarried Men Without Dependents	41	25	32	22
Other ^a	0	0	3	2
All	165	100	144	100

Differences in the composition of tax-filing units reflect differences in underlying populations and differences in how CBO constructs those units on the basis of household relationships (compared with how taxpayers file and claim dependents).

a. Includes people who are married and filing separately as well as qualifying widows and widowers.

The CPS Has Fewer Dependents Than the SOI

Distribution of Dependents, by Taxpayer's Marital Status

	CPS		SOI	
	Number (Millions)	Share of Total (Percent)	Number (Millions)	Share of Total (Percent)
Married Couples	59	65	53	57
Unmarried Women	23	26	26	28
Unmarried Men	8	9	12	13
Other ^a	0	0	1	1
All	91	100	92	100

Dependents in the SOI are more likely to be part of an unmarried individual's tax-filing unit and less likely to be in a joint filer's tax-filing unit than are dependents in the CPS.

a. Includes people who are married and filing separately as well as qualifying widows and widowers.

How the CPS Compares With Tax Data in Terms of People's Income

Income Reported in the CPS Differs From Income Reported in the SOI for Several Reasons

The SOI only has income from taxable sources. The CPS does not include all taxable forms of income (such as capital gains), but it includes nontaxable income (such as means-tested transfers).

Self-reported income in the CPS is likely to include taxable and nontaxable components. For example, pretax contributions to health and retirement plans are not taxable and are excluded from wages in the SOI but may be included in the CPS's measure of wages.

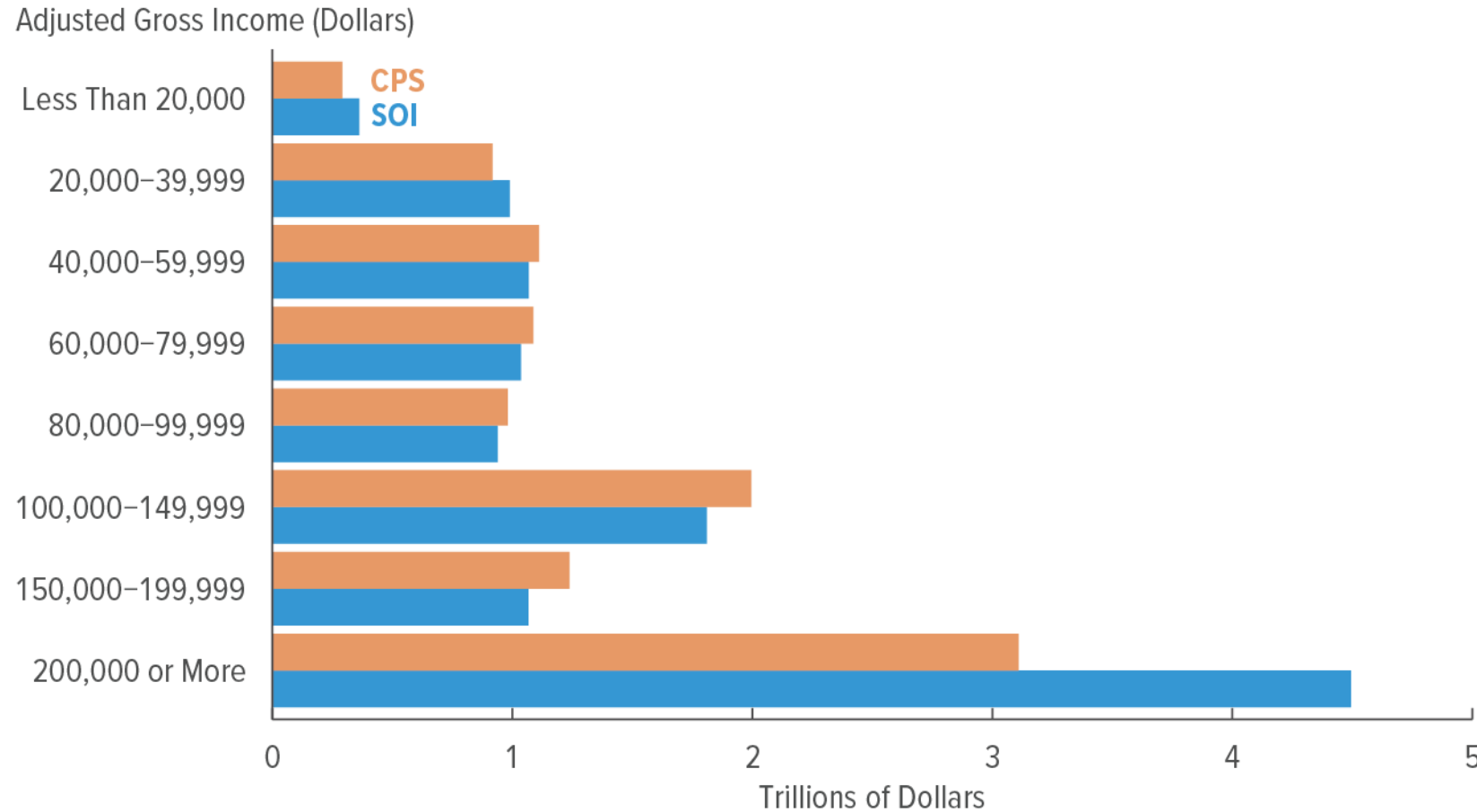
Filers have an incentive to underreport income on a tax return, particularly for items without third-party reporting.

Comparisons of self-reported income in the CPS and administratively reported income have shown significant reporting errors in the CPS.

In 2018, adjusted gross income (AGI) was \$833 billion (or 7 percent) lower in the CPS than was reported on tax returns.

Total Adjusted Gross Income Is Lower in the CPS Than in the SOI

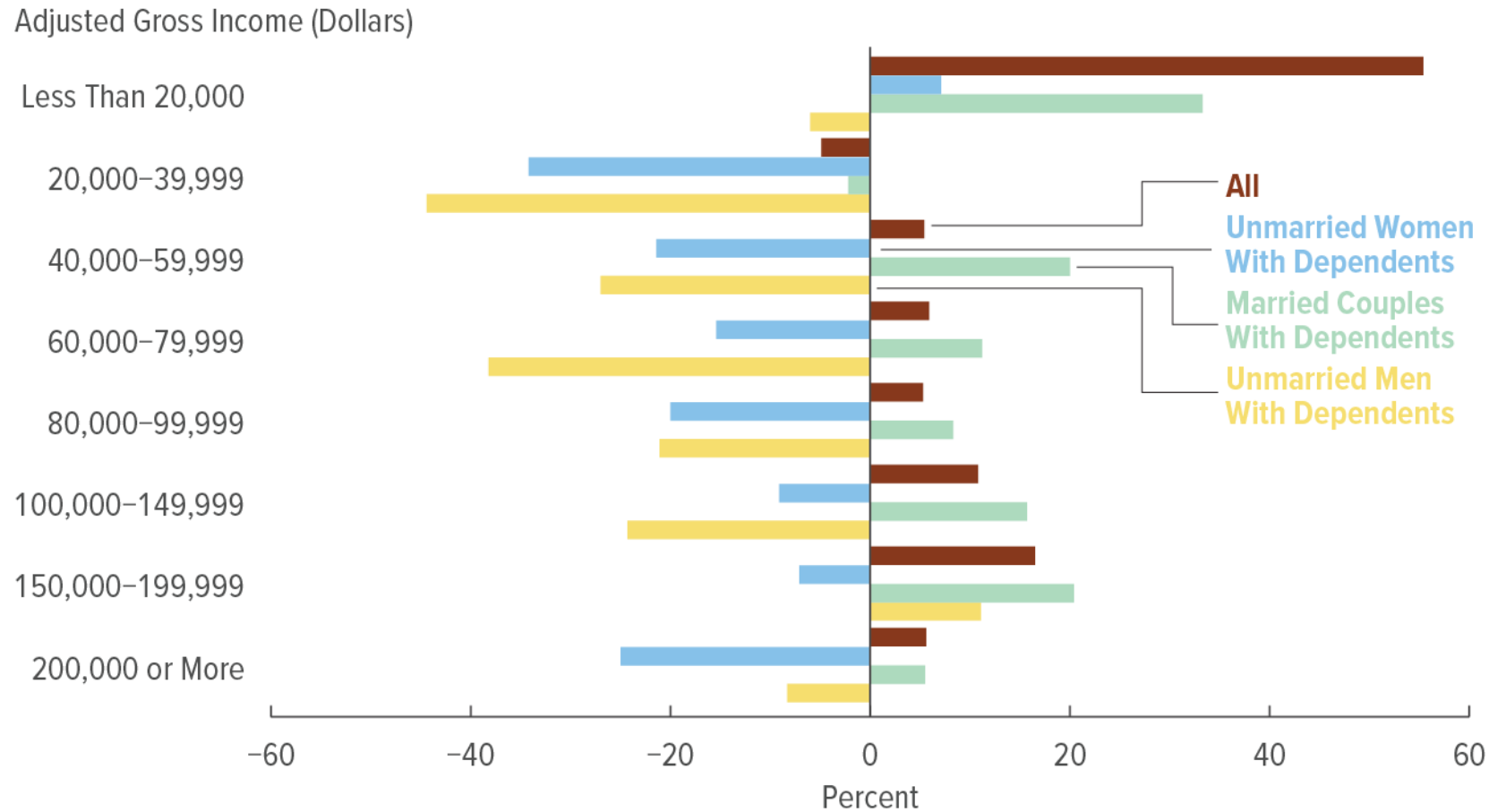
Total Adjusted Gross Income, by AGI Group



The discrepancy between the CPS and SOI is largest for taxpayers with very high income. Some income sources that are significant for high-income taxpayers are not reported in the CPS (such as capital gains and partnership income) or are underreported relative to tax data (such as dividends). The CPS has more tax-filing units in the highest income group than the SOI has, but its total AGI is lower.

Distributions of Tax-Filing Units by Income Group Are Similar in the CPS and SOI

Difference in the Number of Tax-Filing Units in the CPS and SOI, by AGI Group

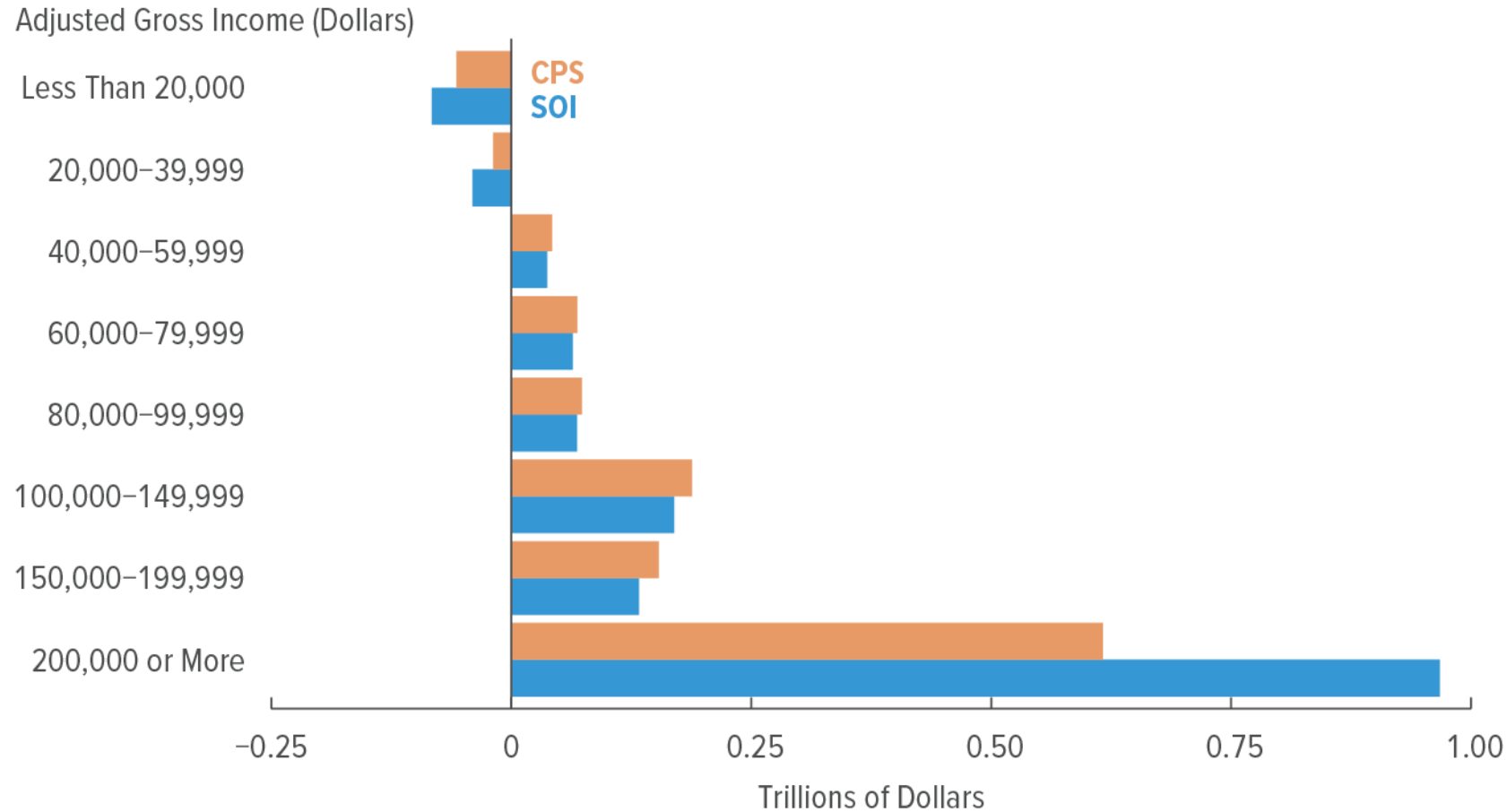


Overall, the CPS generally has more tax-filing units in each income group than the SOI does, and the largest percentage difference is among tax-filing units with an AGI of less than \$20,000. However, the CPS generally has fewer unmarried taxpayers with dependents than the SOI, particularly in groups with lower income.

How the CPS Compares With Tax Data in Terms of Tax Liabilities

Total Tax Liability After Credits Is Smaller in the CPS Than the SOI

Total Tax Liability After Credits, by AGI Group

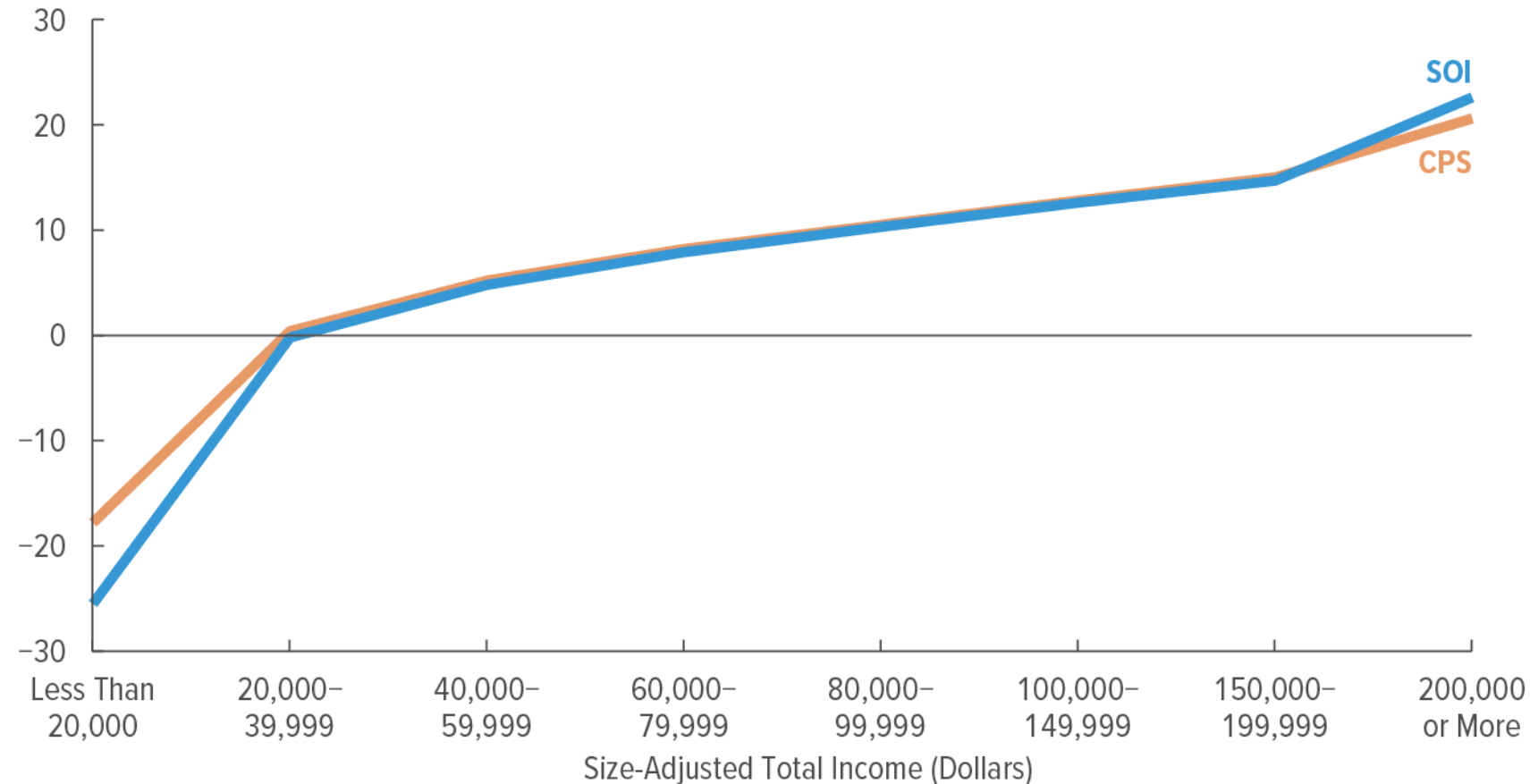


In 2018, total tax liability after credits was \$249 billion (or 19 percent) less in the CPS than in the SOI. The largest differences were at very low and very high incomes.

Average Tax Rates Are Higher for Lower-Income Taxpayers and Lower for Higher-Income Taxpayers in the CPS Than in the SOI

Average Tax Rate

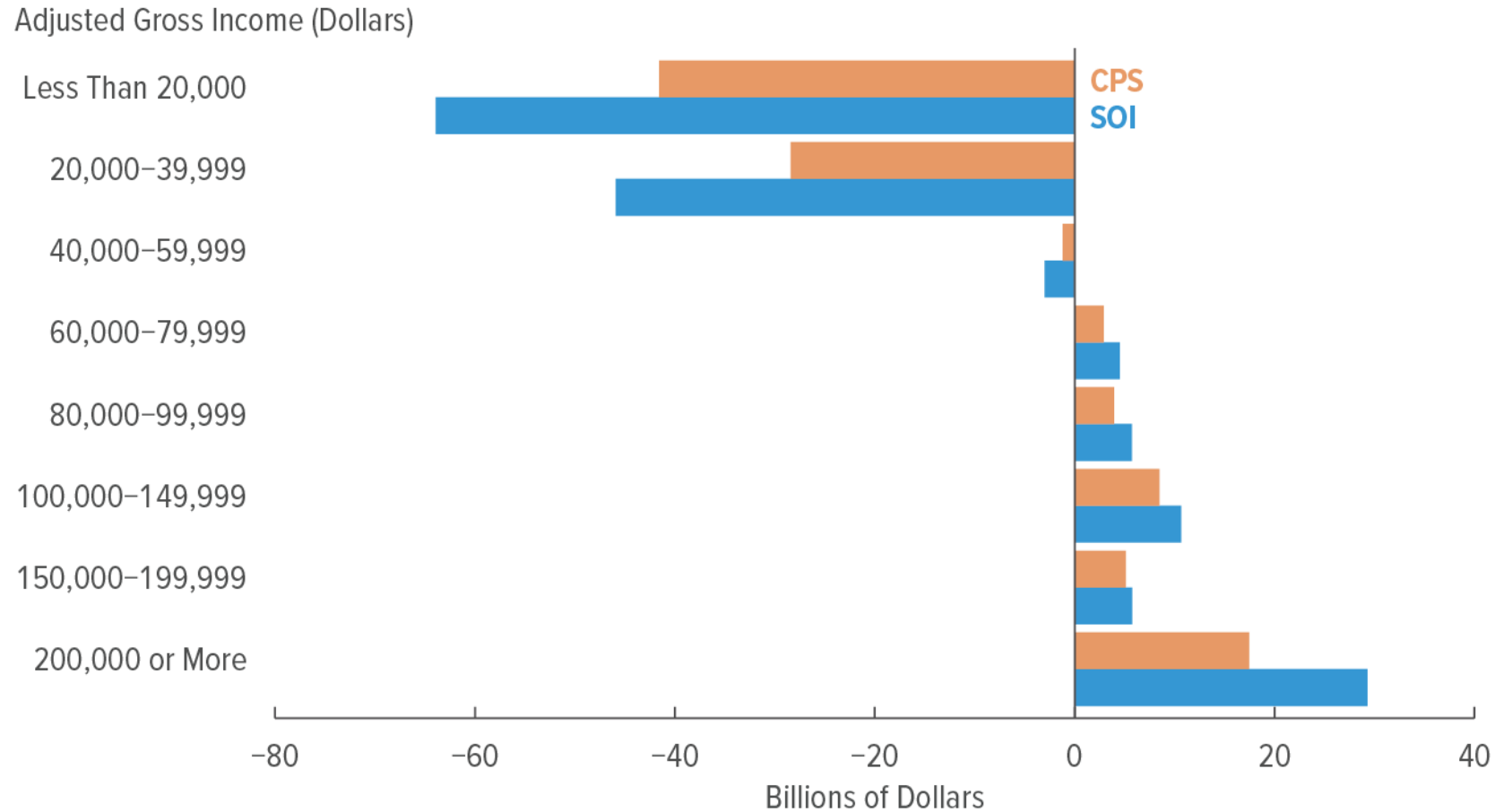
Percent



Tax liability after credits as a share of total income, after adjusting for the size of the tax unit, is generally similar using CPS or SOI data. The largest differences are for taxpayers with very low or very high incomes.

For Unmarried Taxpayers With Dependents, Tax Liability After Credits Is Larger in the CPS Than in the SOI

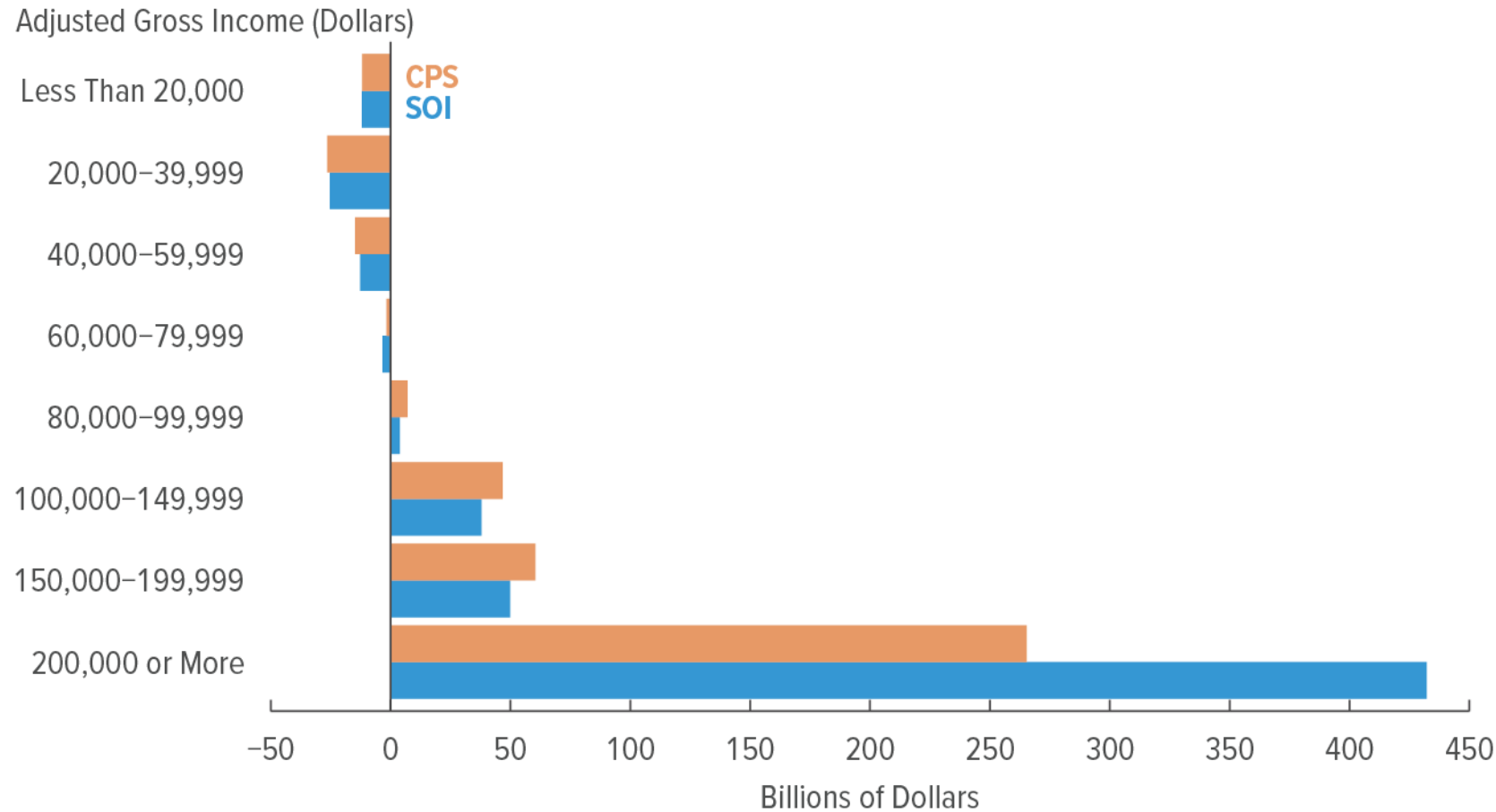
Tax Liability After Credits, Unmarried Men and Women With Dependents



For unmarried taxpayers with dependents in 2018, tax liability after credits was \$24 billion (or 41 percent) greater in the CPS than in the SOI, owing to higher refundable credits in the SOI. The largest differences were among taxpayers with income below \$40,000. For unmarried taxpayers with dependents whose AGI was \$60,000 or more, tax liability after credits was greater in the SOI than in the CPS.

For Married Couples With Dependents, Tax Liability After Credits Is Smaller in the CPS Than in the SOI

Tax Liability After Credits, Married Couples With Dependents



For married taxpayers with dependents in 2018, tax liability after credits was \$145 billion (or 31 percent) smaller in the CPS than in the SOI. (That measure was also smaller in the CPS than in the SOI for unmarried taxpayers in the top income group.) For married taxpayers with income below \$40,000, tax liability after credits was slightly smaller in the CPS than in the SOI.

Estimates of the EITC Using the CPS Are Lower Than Those From the SOI

Distribution of the Earned Income Tax Credit

	CPS	SOI
Number of Recipients (Millions)	14	20
Share of Recipients (Percent)		
Married couples	39	24
Unmarried men	14	19
Unmarried women	47	57
Benefits (Billions of dollars)	41	63
Share of Benefits (Percent)		
Married couples	41	24
Unmarried men	12	17
Unmarried women	47	58

When the amount of the EITC is estimated on the basis of CPS data, it is about one-third lower than the actual amounts of the credit determined from the SOI. Tax-filing units that are eligible for the EITC on the basis of CPS data are more likely to be headed by a married couple than the SOI data indicate. Both sources of data show that among single parents, roughly three-quarters of those eligible for the EITC are women.

Estimates of the Child Tax Credit Using the CPS Are Similar to SOI Estimates

Distribution of the Child Tax Credit

	CPS	SOI
Number of Recipients (Millions)	46	49
Share of Recipients (Percent)		
Married couples	62	52
Unmarried men	11	15
Unmarried women	27	32
Benefits (Billions of dollars)	221	222
Share of Benefits (Percent)		
Married couples	67	57
Unmarried men	8	13
Unmarried women	25	30

When the amount of the CTC is estimated on the basis of CPS data, it is much closer to the actual amounts of the credit determined from the SOI. Tax-filing units that are eligible for the CTC on the basis of CPS data are more likely to be headed by a married couple than the SOI data indicate. Both sources of data show that among single parents, those who are eligible for the CTC are much more likely to be women than men. The share of benefits reflects the share of dependents in each taxpayer group.

Using Merged Data for Estimates of the EITC and CTC

Merged Data Allow Demographic Information From the CPS to Be Appended to Income Data From Tax Returns

The merged data are used to show how the tax credits would appear to be distributed if marital status and sex from the CPS were used to categorize taxpayers but if credit amounts were based on income and tax-unit composition from the SOI.

Sex is reported in the CPS and SOI data but is not currently controlled for in the statistical match used to create the merged data.

That example illustrates the challenges in distributing tax benefits by other characteristics (such as race and ethnicity) that are only reported in the CPS. The relationships between additional demographic characteristics, tax-unit composition, and income may not be maintained in the merged file.

Estimates of the EITC Using Merged Data Are Similar to Those From the SOI

Distribution of the Earned Income Tax Credit, by Data Set

	CPS	SOI	Merged
Number of Recipients (Millions)	14	20	20
Share of Recipients (Percent)			
Married couples	39	24	27
Unmarried men	14	19	22
Unmarried women	47	57	52
Benefits (Billions of dollars)	41	63	63
Share of Benefits (Percent)			
Married couples	41	24	26
Unmarried men	12	17	20
Unmarried women	47	58	54

When the distribution of EITC benefits by marriage status and sex is categorized using CPS demographic information that has been statistically matched, it is closer to the SOI data than it would be if only the CPS data were used. However, a larger share of benefits is attributed to unmarried men in the merged data than in the SOI or CPS data alone.

Estimates of the CTC Using Merged Data Are Also Similar to Those From the SOI

Distribution of the Child Tax Credit, by Data Set

	CPS	SOI	Merged
Number of Recipients (Millions)	46	49	49
Share of Recipients (Percent)			
Married couples	62	52	55
Unmarried men	11	15	16
Unmarried women	27	32	29
Benefits (Billions of dollars)	221	222	222
Share of Benefits (Percent)			
Married couples	67	57	59
Unmarried men	8	13	13
Unmarried women	25	30	28

When the distribution of CTC benefits by marriage status and sex is categorized using CPS demographic information that has been statistically matched, it is closer to how benefits are actually distributed based on tax return information than it would be if only the CPS data were used. However, the ratio of benefits for unmarried women to that of unmarried men is lower based on the merged data than it is in the SOI or CPS data, thus understating the distribution of benefits to unmarried women compared with men.

About This Document

This document was prepared to enhance the transparency of CBO's work and to encourage external review of that work. In keeping with CBO's mandate to provide objective, impartial analysis, the document makes no recommendations.

Shannon Mok prepared the document with guidance from John McClelland and Joseph Rosenberg. Edward Harris, Alexandra Minicozzi, and Natalie Tawil provided comments. Tess Prendergast fact-checked the document.

Jeffrey Kling and Robert Sunshine reviewed the document. Christine Bogusz edited it and prepared the text for publication. R. L. Rebach created the graphics. The document is available on CBO's website at www.cbo.gov/publication/57540.