Congressional Budget Office
Nonpartisan Analysis for the U.S. Congress

Using Multiple Data Sources to Learn About the Race and Ethnicity of Taxpayers

January 5, 2024

Presentation at the
American Economic Association Meeting, Committee on Economic Statistics

Rebecca Heller, Labor, Income Security, and Long-Term Analysis Division, CBO

Shannon Mok, Tax Analysis Division, CBO

James Pearce, Tax Analysis Division, CBO

Jonathan Rothbaum, Census Bureau

For more information about the conference, see www.aeaweb.org/conference/

Disclaimers

This work is released to inform interested parties of ongoing research and to encourage discussion. The views expressed in this presentation, including those related to statistical, methodological, technical, or operational issues are solely those of the authors and do not necessarily reflect the official positions or policies of the Census Bureau.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System (DMS) number: P-7527994 (7531634), Disclosure Review Board (DRB) approval number: CBDRB-FY23-SEHSD003-053).

Background

Tax Data Lacks Race and Ethnicity Information

Internal Revenue Service (IRS) data provide high-quality measures of income and are useful for studying how the tax system affects households. But these data lack information about the race and ethnicity of individuals.

Researchers have used survey data, which includes self-reported race and ethnicity data, to study how the tax system differentially affects households by race and ethnicity (Goldin and Michelmore 2022; Holtzblatt et al. 2023). Research has also shown that many surveys face reporting issues that can make drawing inferences difficult (Meyer et al. 2022).

Researchers and policymakers have criticized this lack of data (Brown 2021) and there have been new efforts to analyze this issue using administrative tax data (Fisher 2023; Cronin et al. 2023).

This presentation summarizes preliminary work by CBO and the Census Bureau as part of CBO's ongoing efforts to increase its capacity to analyze budgetary and economic outcomes for various demographic groups (CBO 2022).

Methods to Add Race and Ethnicity Information to Tax Data

- **Data statistically matched by CBO.** The agency maintains a microsimulation model based on tax data that is statistically matched to the Current Population Survey Annual Social and Economic Supplement (CPS ASEC).
  - Although the resulting data contains race and ethnicity information from the CPS ASEC, CBO has not used it for analysis because it is not clear that the statistical match preserves the relationships between income, tax liability, and race and ethnicity.
  - The lack of race and ethnicity on tax data makes comparisons difficult as it is not clear how the distribution of income by race and ethnicity might differ between administrative tax data and survey data.

- **Data linked by the Census Bureau.** The Census Bureau uses tax data to supplement the information provided by respondents in its household surveys (including the CPS ASEC). Because of this, the Census Bureau has CPS ASEC data linked to administrative tax data on an individual level.

Data

Data Statistically Matched by CBO

CBO's statistically matched approach combined records from tax return data and the CPS ASEC with a tax unit as the matching concept.

- Tax units are the basic unit of observation in the tax return data.
- In the CPS-ASEC, CBO used an algorithm based on reported relationships, ages, and income variables to construct tax units for each household.

CBO defined demographic cells on the basis of a taxpayer's age, marital status, number of children, and whether the taxpayer could be claimed as a dependent. Then, the agency assigned tax units in each data set to the appropriate cell.

For each demographic cell, CBO used the tax data to estimate a regression of total income on its components that are observed in both the tax data and the CPS ASEC. The estimated coefficients were then used to calculate predicted total income for each tax unit in each data set.

For additional details, see CBO 2017.

Data Statistically Matched by CBO (Continued)

Within each demographic category, CBO ranked tax units in each data set from highest to lowest using their predicted total income.

Next, CBO matched the tax unit from the tax data with the highest predicted total income to the CPS ASEC–constructed tax unit with the highest predicted total income.

In this way, moving from highest to lowest income within each demographic cell, tax units were matched across the two data sets accounting for differences in weights.

Unmatched tax units constructed in the CPS ASEC represent nonfilers.

For this analysis, we used the 2019 CPS ASEC and tax returns filed for 2018.

For additional details, see CBO 2017.

Data Linked by the Census Bureau

The linked data are based on the 2019 CPS ASEC sample and include people with a Protected Identification Key (PIK).

Using the PIK, survey respondents were linked with their 2018 administrative tax data, including certain fields in the IRS Forms 1040, W-2, Schedule SE, and Schedule A.

Survey respondents were matched to the primary taxpayer and the spouse on the return. If a dependent was in the CPS ASEC household and had a PIK, we linked that dependent to the return as well.

We did not have information about whether someone in the household was claimed as a dependent by someone outside the CPS ASEC household. In future work, we may be able to observe that.

For details on the Census Bureau's data linking process, see Wagner and Layne 2014.

Adjustments to the Linked Data

We made three adjustments to the data:

- Sample weights were adjusted to account for the probability that a person (or couple) had a PIK.
- We used information on the seasonality of tax filing by income group to estimate who was likely to file late and included those likely late filers as filers for this work using their survey-reported information.
- We used the filing status and tax unit composition from Form 1040 to categorize filers and the CPS ASEC tax units constructed for CBO's statistical match for nonfilers and late filers.
  - Some married couples in the CPS ASEC did not file together. In those cases, the couple was "split" and information from both tax returns was kept.
  - Thus, there are more tax units in the data linked by the Census Bureau (181 million) than in the data statistically matched by CBO (175 million).

Initial Results

Preliminary Evaluation

Initial evaluations focused on broad characteristics of tax units:

- Tax unit composition and filing status
- Adjusted Gross Income (AGI)

We also looked at some specific groups of filers:

- Tax units that claimed the Earned Income Tax Credit (EITC)
- Tax units that itemized deductions

Our analysis excluded returns filed by dependents as well as tax units headed by people under 18.

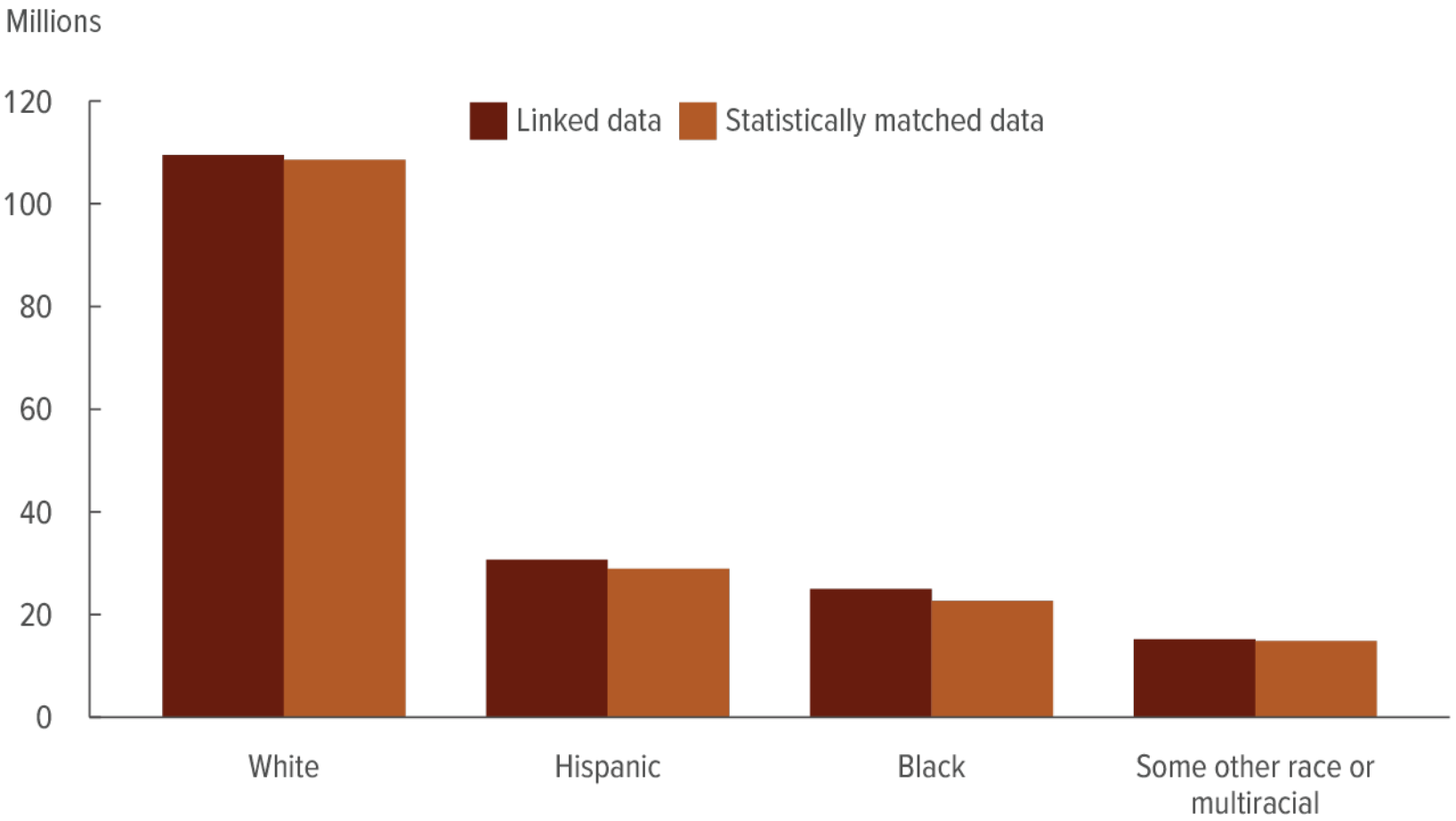
Race and Ethnicity Categories

We categorized tax units by the race and ethnicity of the primary taxpayer (the taxpayer listed first on a tax return).

- We grouped taxpayers reporting Hispanic ethnicity in the category “Hispanic” and all other taxpayers in categories using what they report about race.

- The categories “White” and “Black” consist of people who only report being in one of those groups. The category “Some other race or multiracial” consists of people who report being of another race or multiple races.

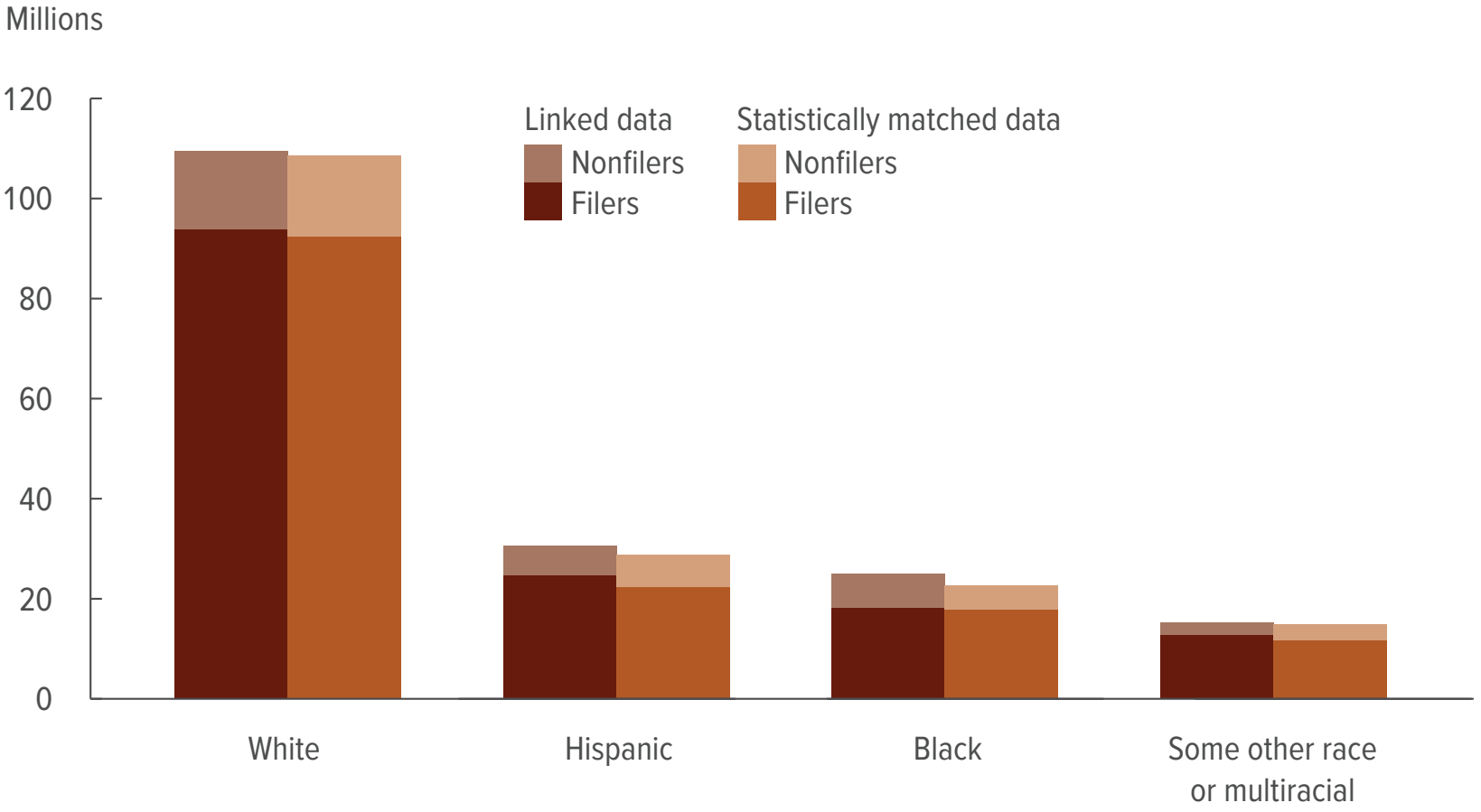
Number of Tax Units, by the Race and Ethnicity of the Primary Taxpayer, 2018



Overall, the distribution of the number of tax units by race and ethnicity in the two datasets are comparable.

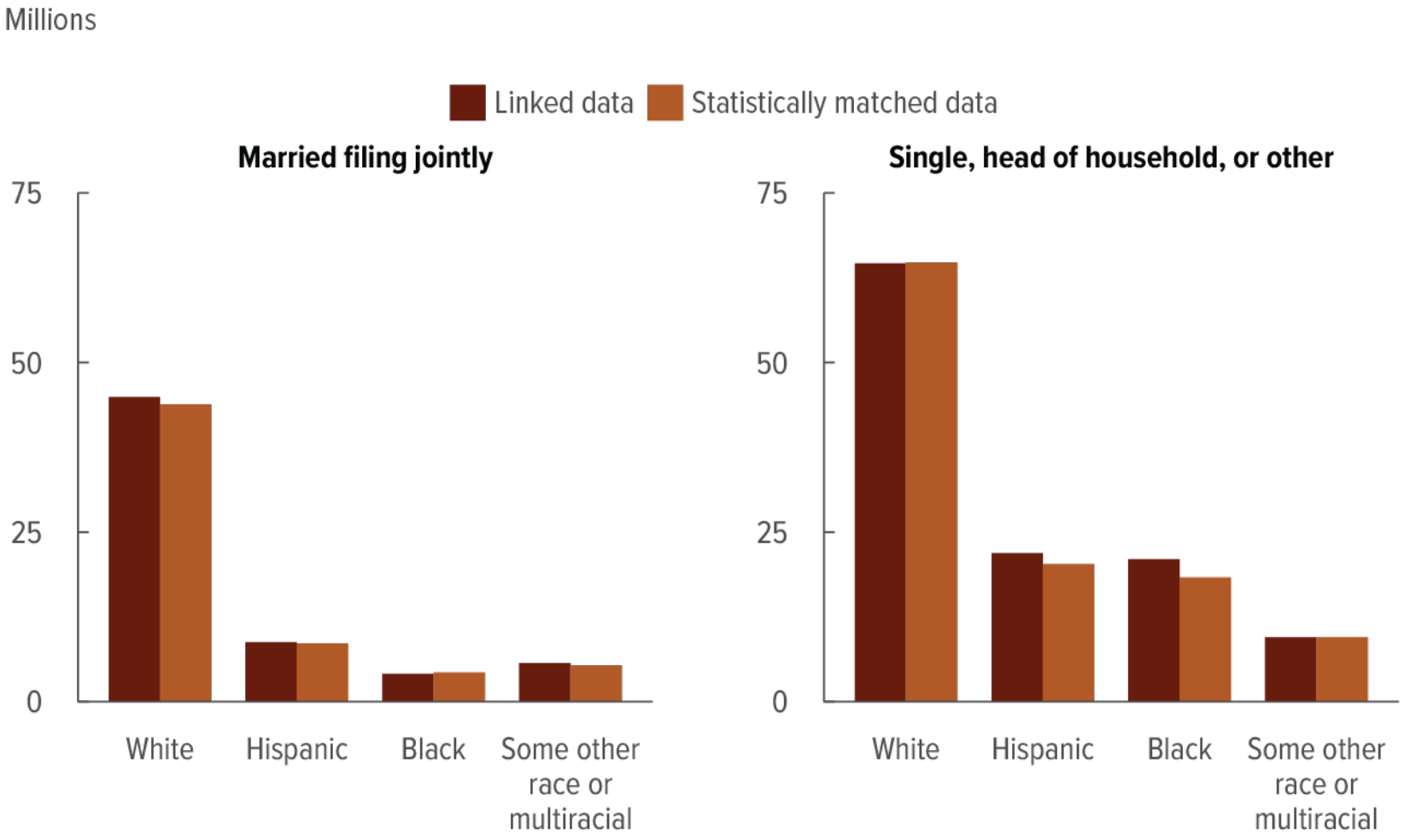
The linked data has slightly more tax units because not all married couples in the CPS ASEC file jointly. The additional tax units in the linked data seem more likely to have a Black or Hispanic primary taxpayer.

Number of Filing and Nonfiling Tax Units, 2018



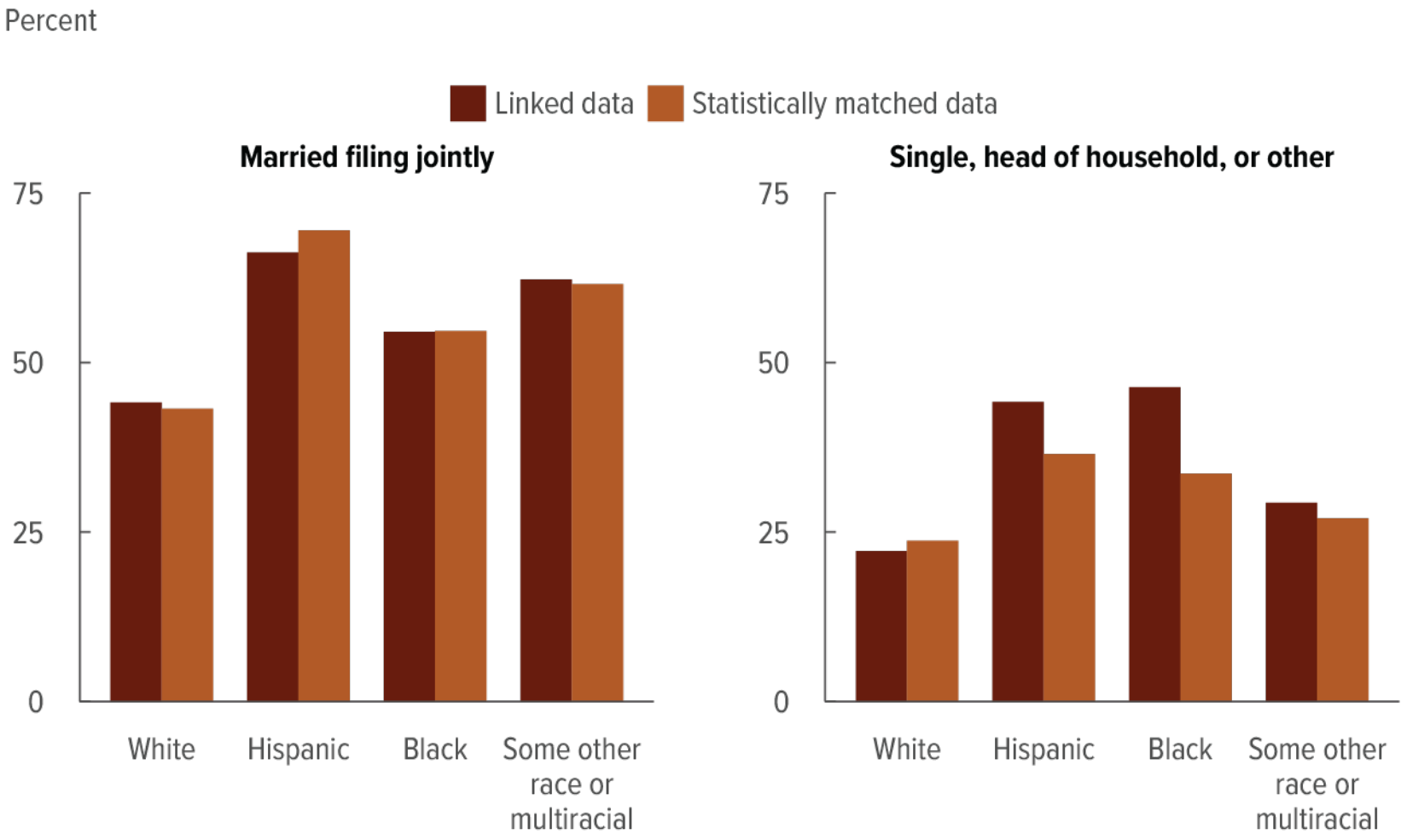
Additionally, the number of nonfiling tax units varies by race and ethnicity in the two data sets, with the biggest discrepancy occurring for tax units with a Black primary taxpayer.

Number of Tax Units, by the Filing Status and the Race and Ethnicity of the Primary Taxpayer, 2018



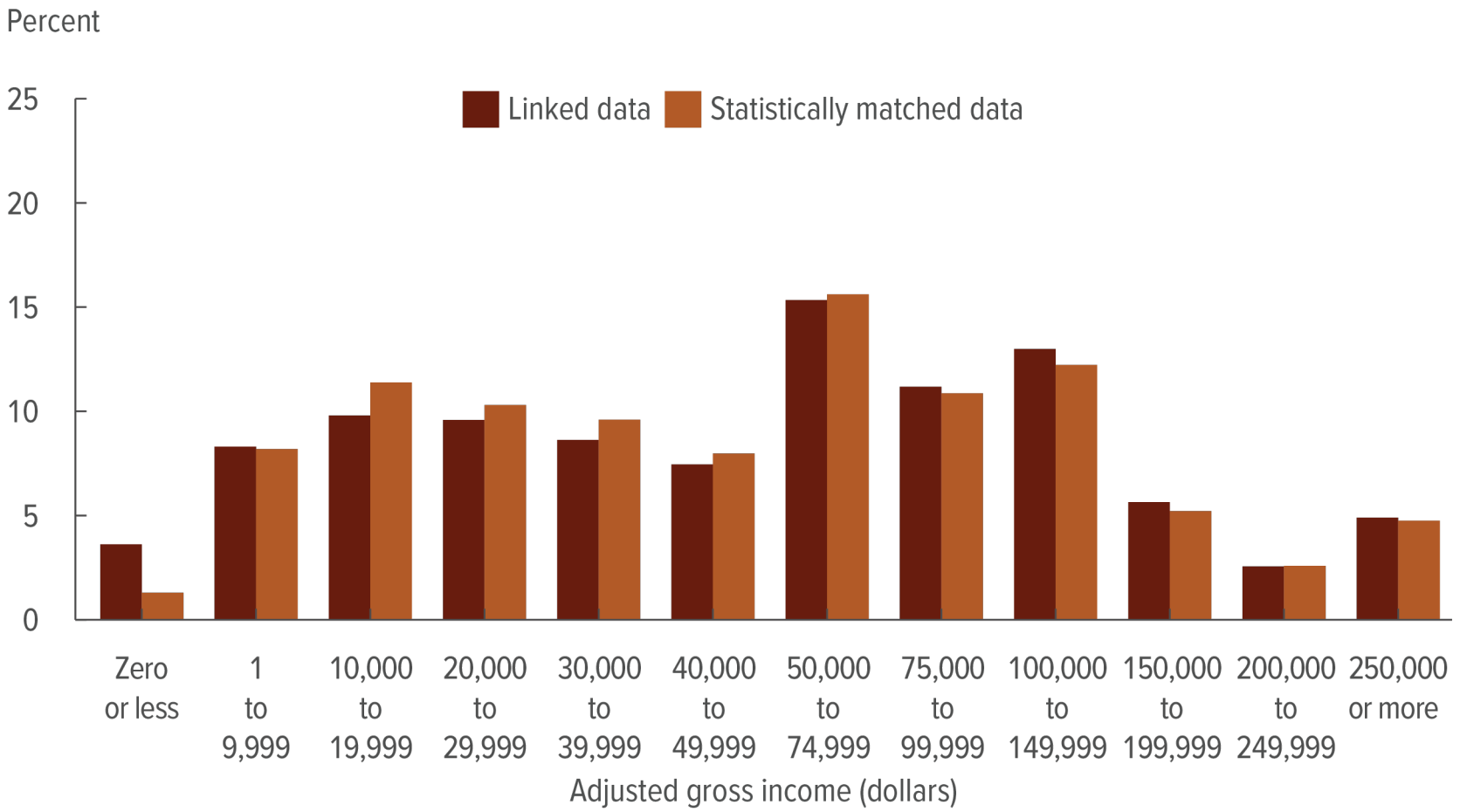
Potential differences in the number of tax units in the two data sets warrant further investigation.

Share of Filing Tax Units with Dependents, by the Filing Status and the Race and Ethnicity of the Primary Taxpayer, 2018



In both data sets, potential differences exist in the share of filing tax units with dependents, primarily in tax units that are not married filing jointly or in units in which the primary taxpayer is Hispanic. Some dependents observed in a tax unit in the linked data are outside of the household reported in the CPS ASEC.

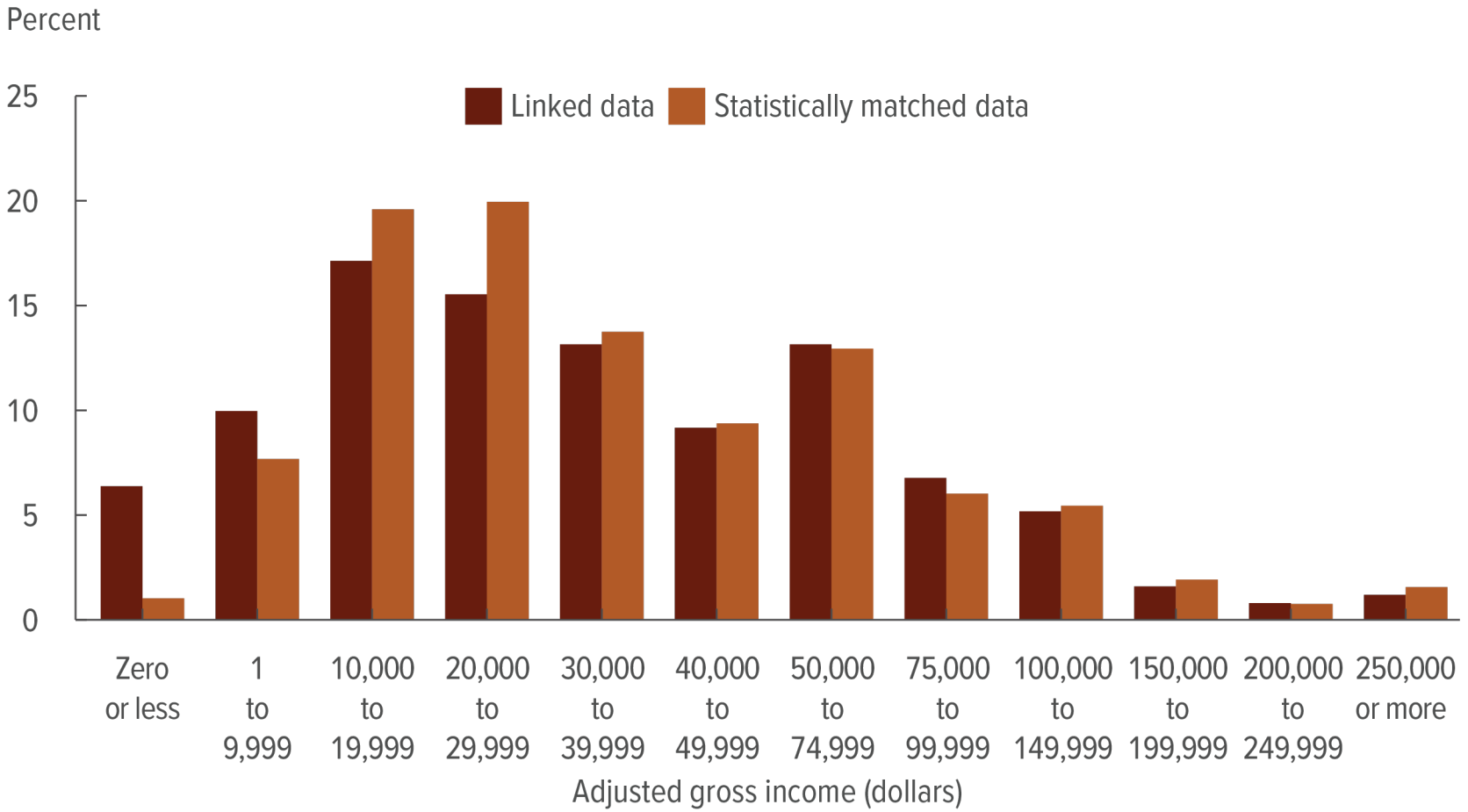
Distribution of Filing Tax Units That Have Primary Taxpayers Who Are White, by Income, 2018



The distribution of tax units headed by White primary taxpayers across income groups is broadly similar in the two data sets, though there are some potential differences, particularly for those tax units with AGI at the bottom of the distribution.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

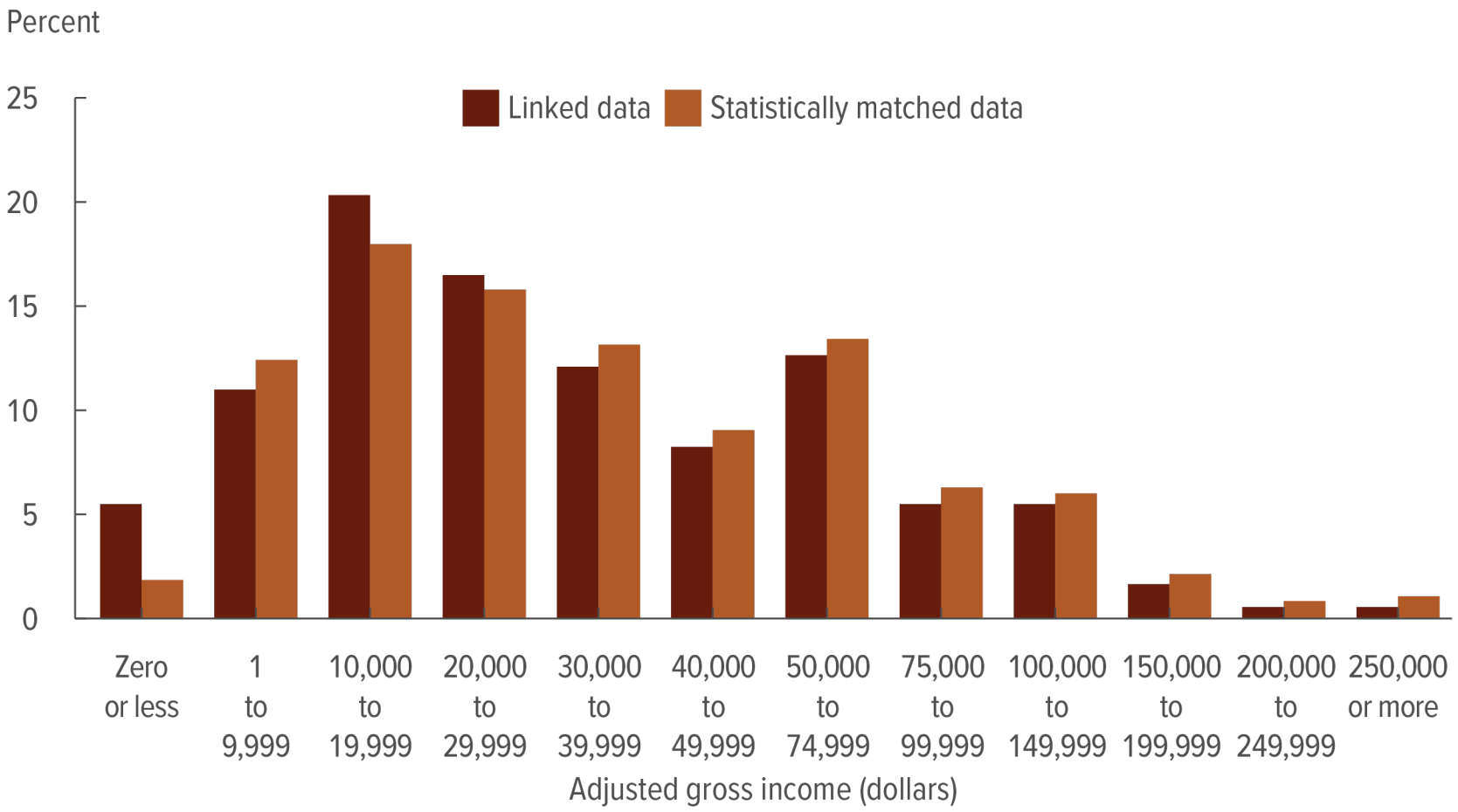
Distribution of Filing Tax Units That Have Primary Taxpayers Who Are Hispanic, by Income, 2018



The differences in the distribution of tax units across income groups between the two data sets appear to be more pronounced among tax units with Hispanic primary filers than among those with White primary filers. Those differences are largely concentrated in lower income groups.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

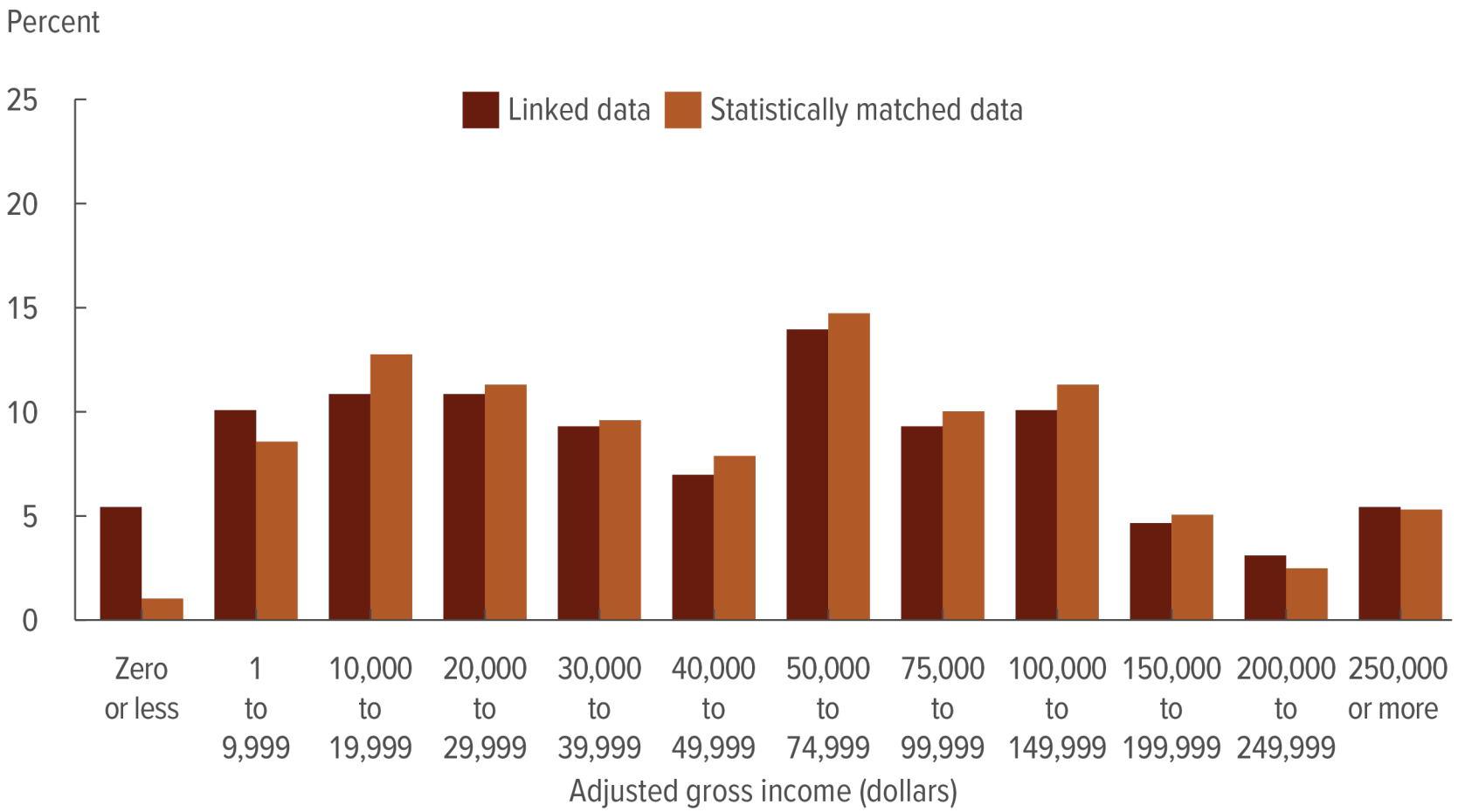
Distribution of Filing Tax Units That Have Primary Taxpayers Who Are Black, by Income, 2018



The statistically matched data appear to have a larger share of filing units with Black primary taxpayers in the higher income groups than do the linked data.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

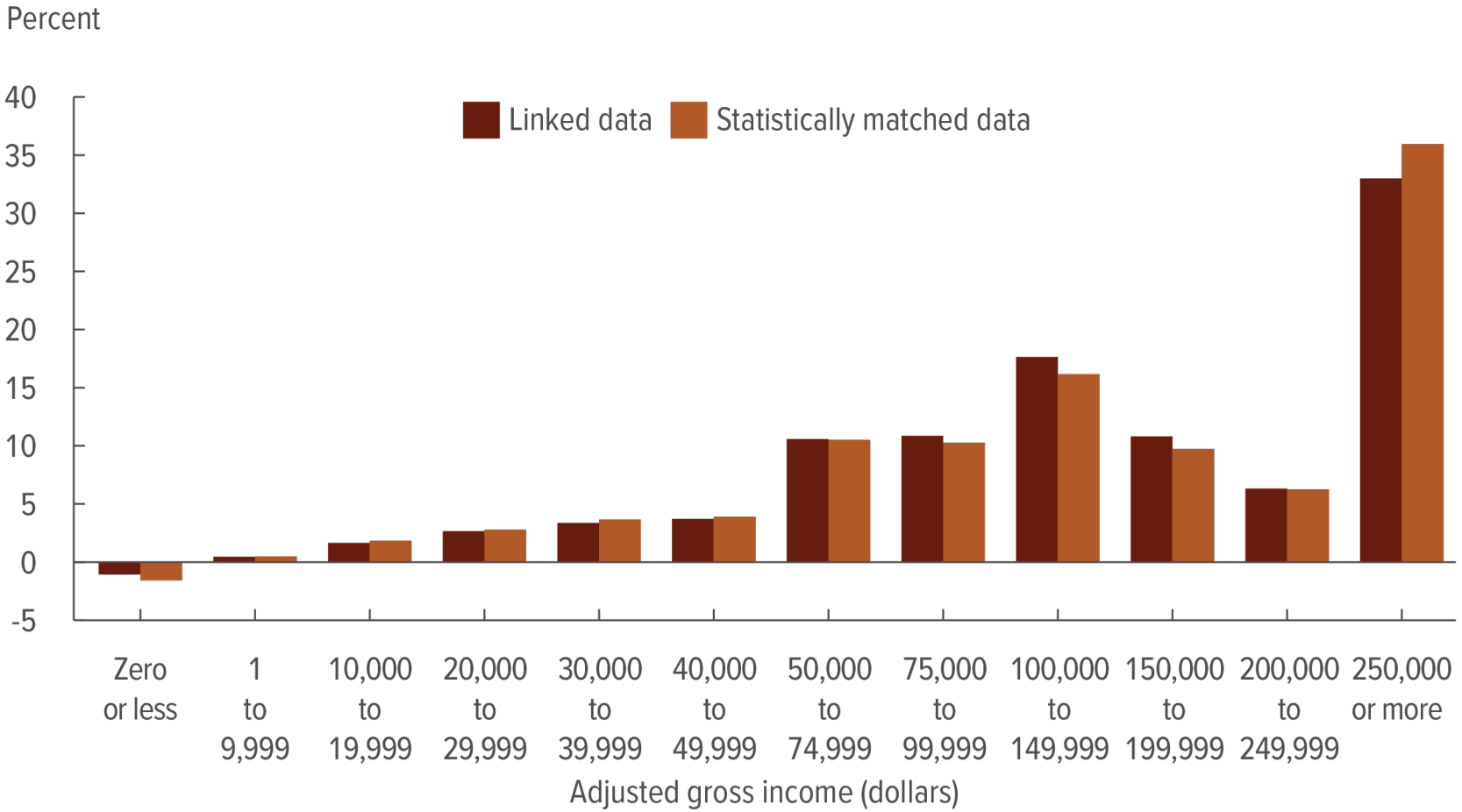
Distribution of Filing Tax Units That Have Primary Taxpayers Who Are Some Other Race or Multiracial, by Income, 2018



The linked data appear to have a larger share of filing units with primary taxpayers who are some other race or multiracial in the lowest and highest income groups than does the statistically matched data.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

Distribution of Income Across Filing Tax Units That Have Primary Taxpayers Who Are White, 2018

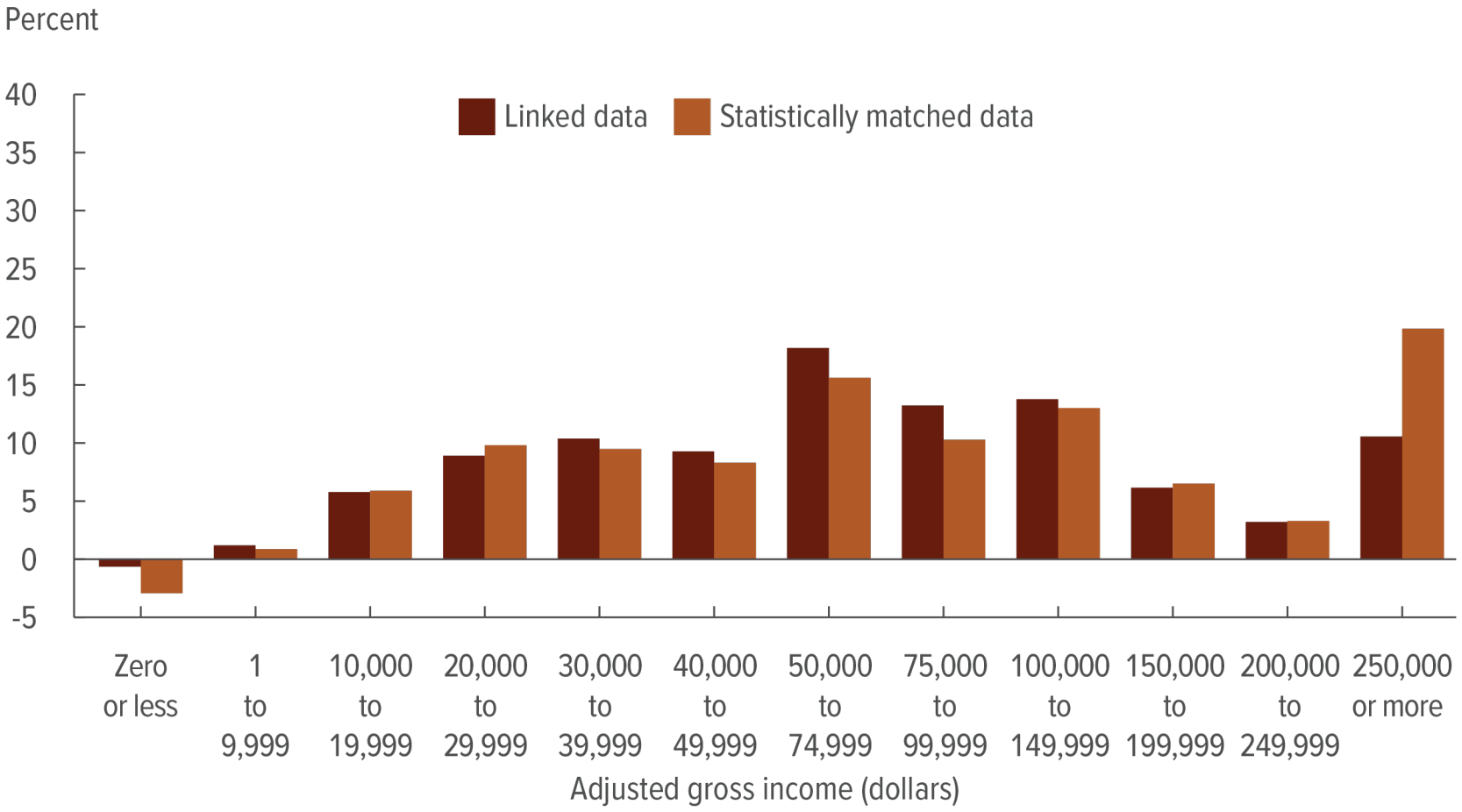


Across filing tax units with White primary taxpayers, the distribution of income in the data sets compares well. The difference in the highest income group could be attributable to differences in sampling between the SOI and the CPS ASEC.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

In the statistically matched data, tax units with White primary taxpayers account for 64 percent of tax units and 73 percent of AGI; in the data linked by the Census Bureau, such tax units account for 63 percent of tax units and 75 percent of AGI.

Distribution of Income Across Filing Tax Units That Have Primary Taxpayers Who Are Hispanic, 2018

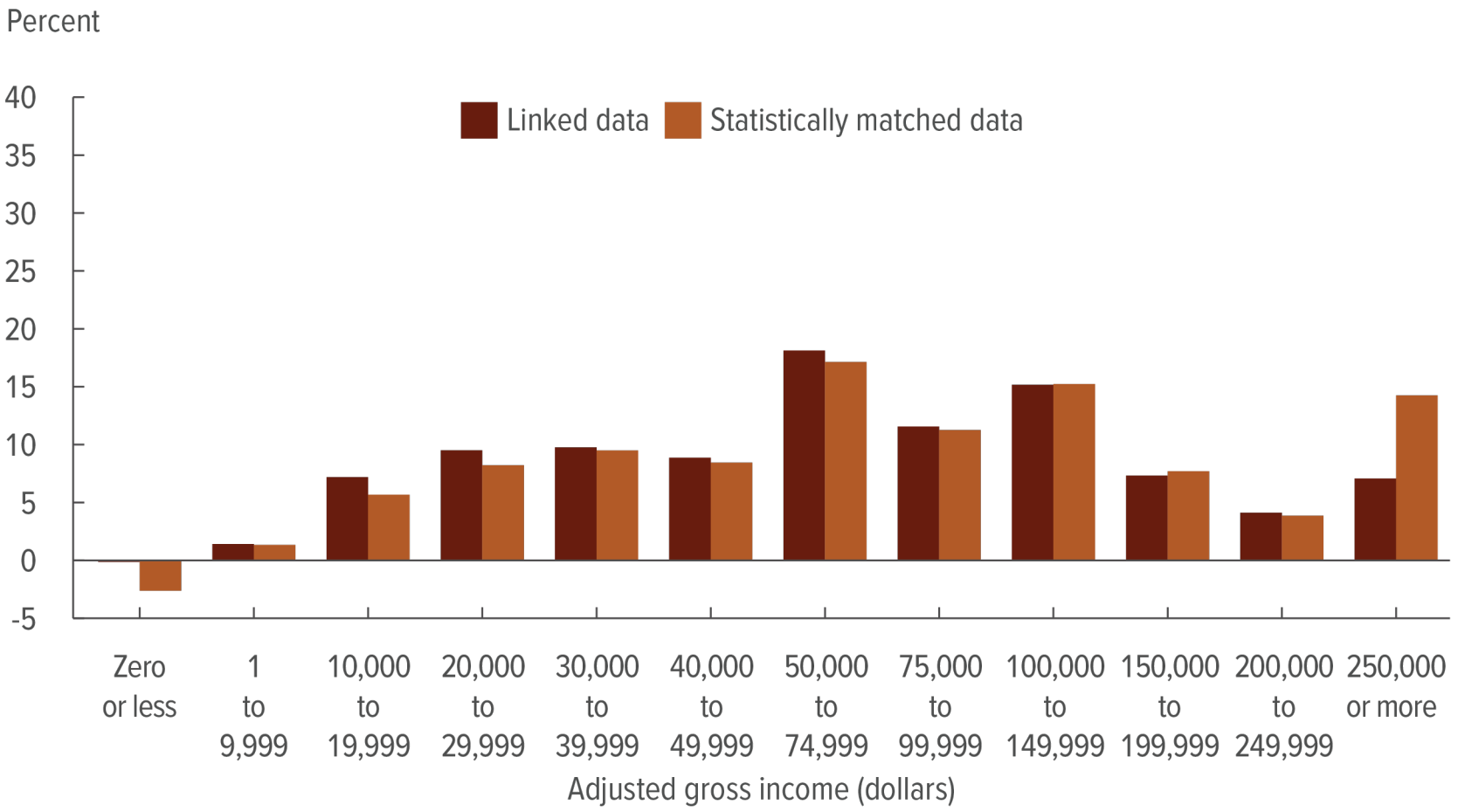


Filing tax units that have Hispanic primary taxpayers and income above $250,000 have a larger share of income in the statistically matched data than in the linked data. That difference could be attributable to differences in sampling between the SOI and the CPS ASEC.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

In the statistically matched data, tax units with Hispanic primary taxpayers account for 16 percent of tax units and 10 percent of AGI; in the data linked by the Census Bureau, such tax units account for 17 percent of tax units and 10 percent of AGI.

Distribution of Income Across Filing Tax Units That Have Primary Taxpayers Who Are Black, 2018

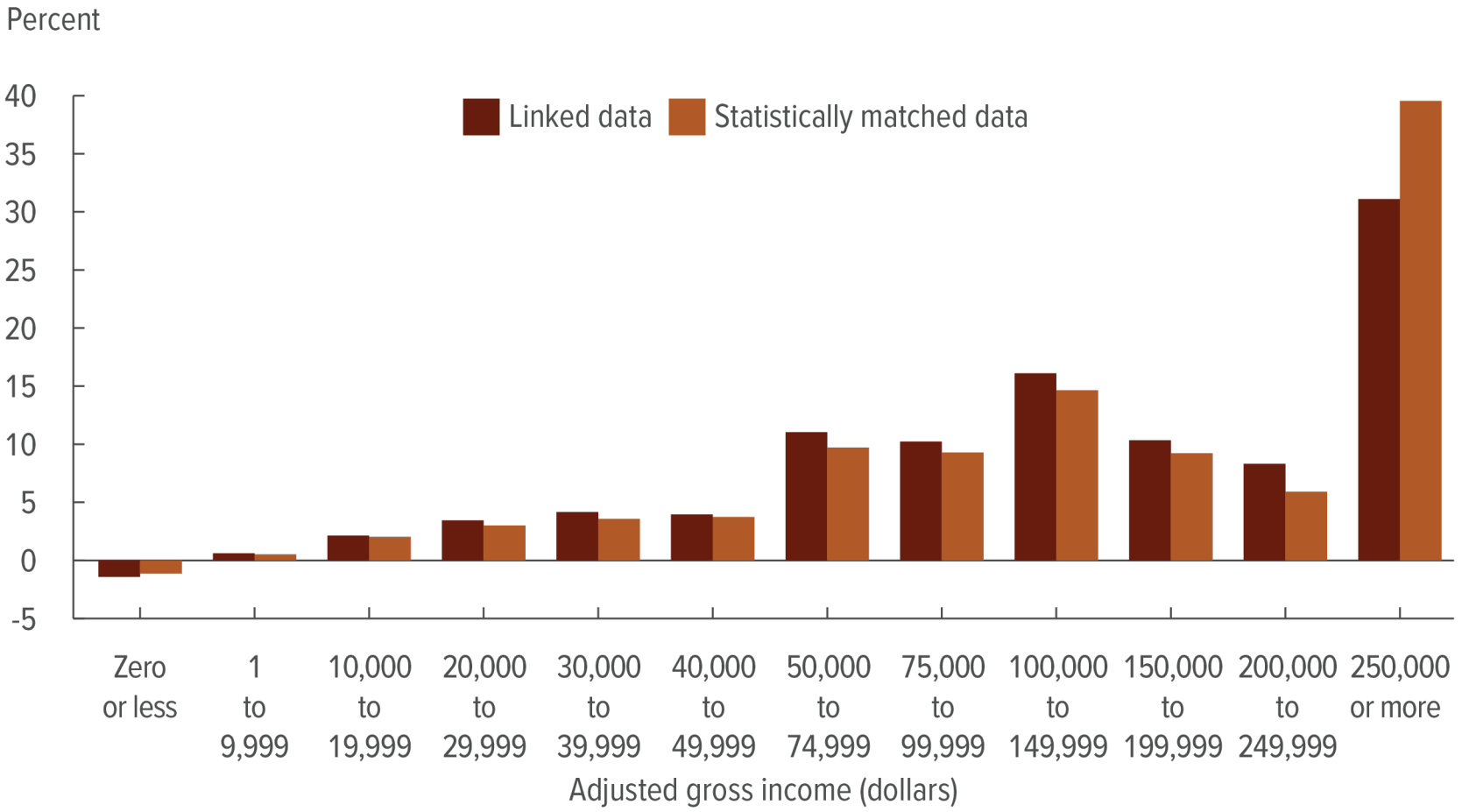


Similar to filing tax units with Hispanic primary taxpayers, the statistically matched data appear to have a larger share of income held by tax units with Black primary taxpayers in the top income category than the linked data does. That difference could also be attributable to differences in sampling between the SOI and the CPS ASEC.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

In the statistically matched data and the linked data, tax units with Black primary taxpayers account for 12 percent of tax units and 7 percent of AGI.

Distribution of Income Across Filing Tax Units That Have Primary Taxpayers Who Are Some Other Race or Multiracial, 2018

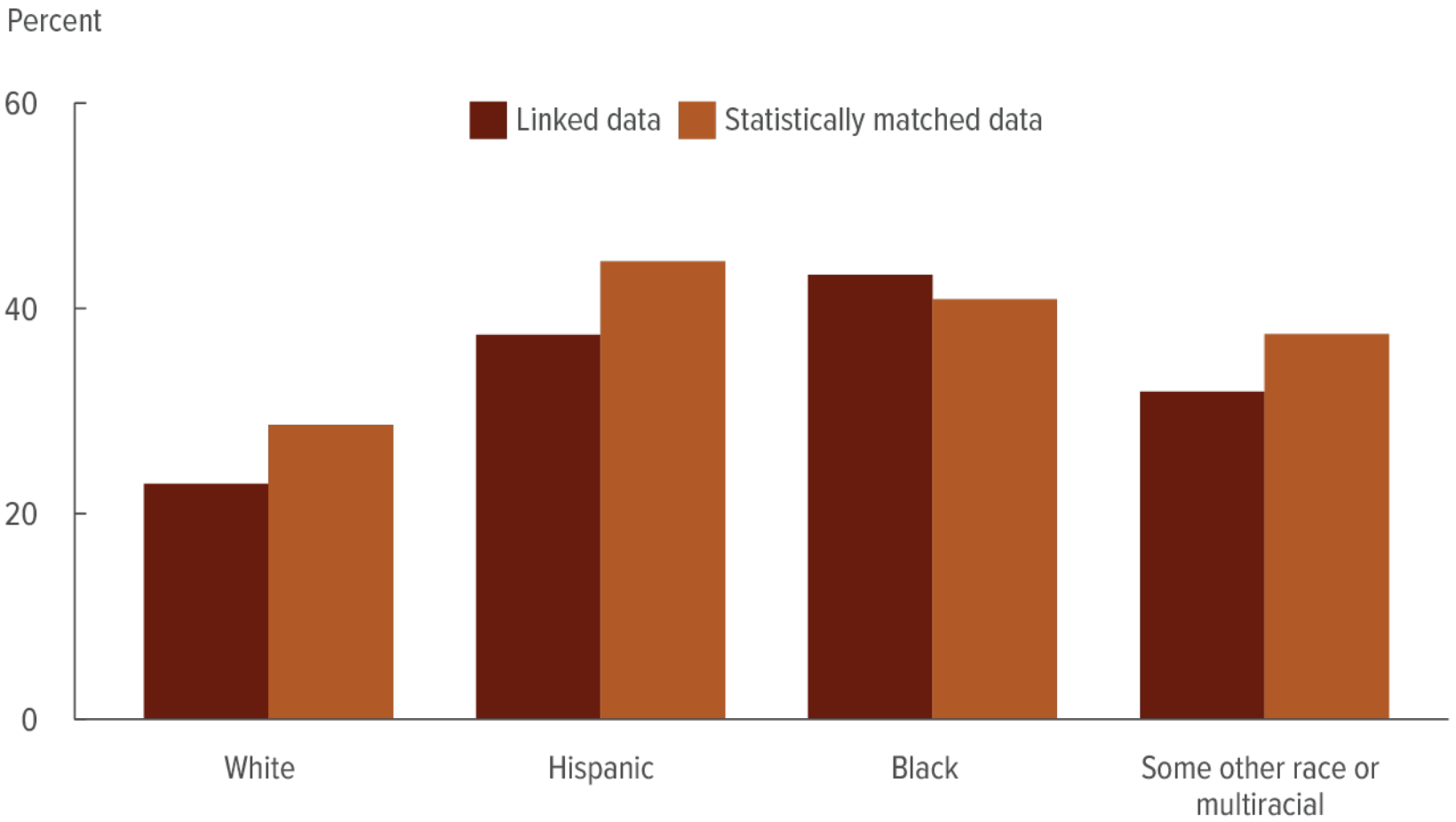


The distribution of income among filing tax units with a taxpayer who is some other race or multiracial resembles that of tax units that have a White primary taxpayer. Both data sets estimate that a large portion of income accrues to taxpayers with incomes of more than $250,000.

CBO has corrected this slide since the presentation was originally published. The corrections are described at the end of the presentation.

In the statistically matched data, tax units with primary taxpayers who are some other race or multiracial account for 8 percent of tax units and 9 percent of AGI; in the data linked by the Census Bureau, such tax units account for 9 percent of tax units and 9 percent of AGI.

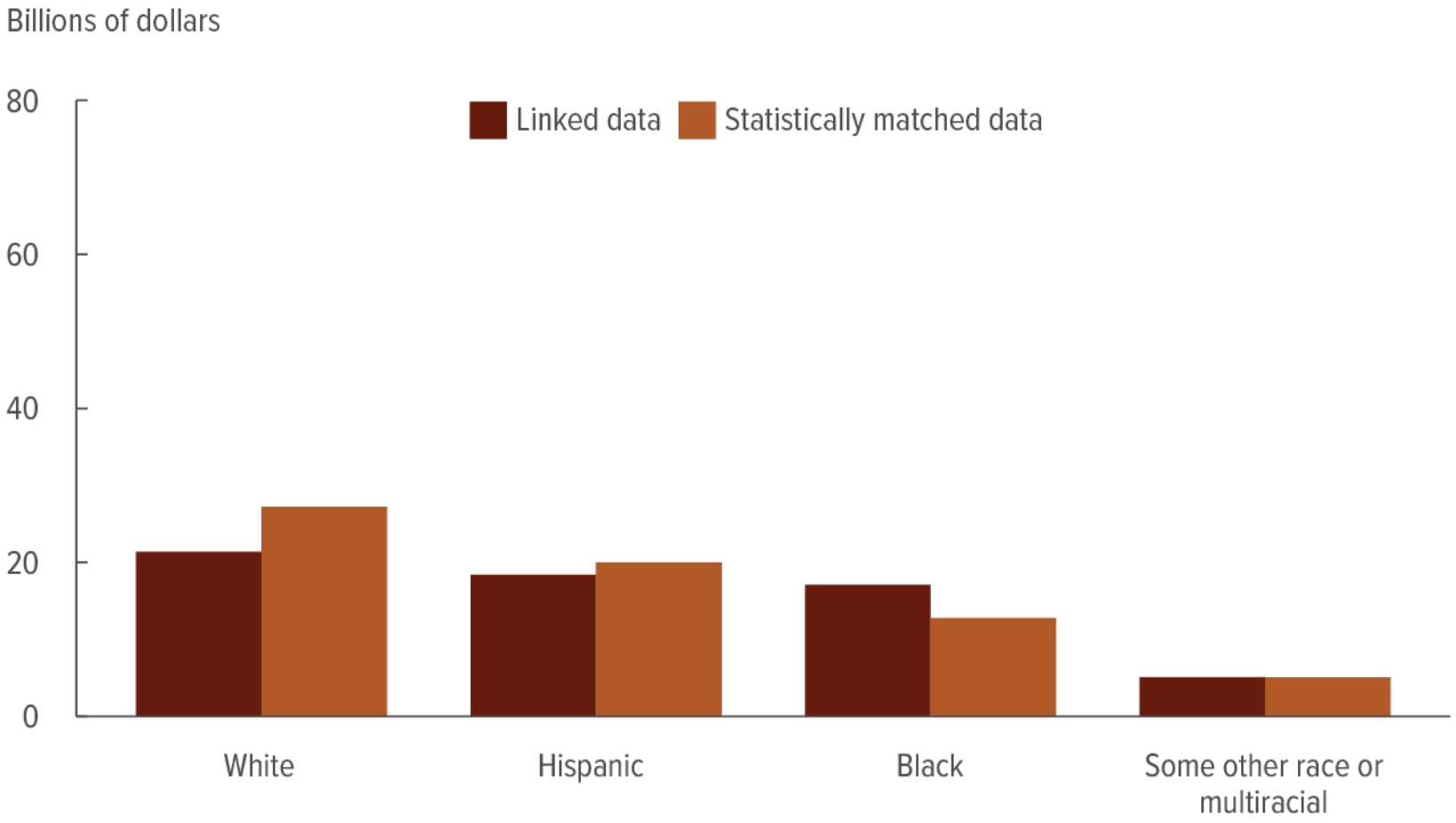
Share of Tax Units That Received EITC in 2018, by the Race and Ethnicity of the Primary Taxpayer



Overall, we estimate that 24.9 million tax units in the linked data (or 30 percent of all filing units with AGI below $50,000) and 27.4 million tax units in the statistically matched data (or 34 percent of all filing units with AGI below $50,000) received the EITC for 2018. The IRS reported that 26.5 million tax units received it for that year.

Share of tax units with EITC was restricted to tax units with AGI below $50,000. In both data sets, we calculated the EITC that a taxpayer was eligible for on the basis of earned income, filing status, and the number of EITC-qualifying children reported in tax data. In the linked data, only taxpayers who filed their 2018 tax returns in 2019 were included.

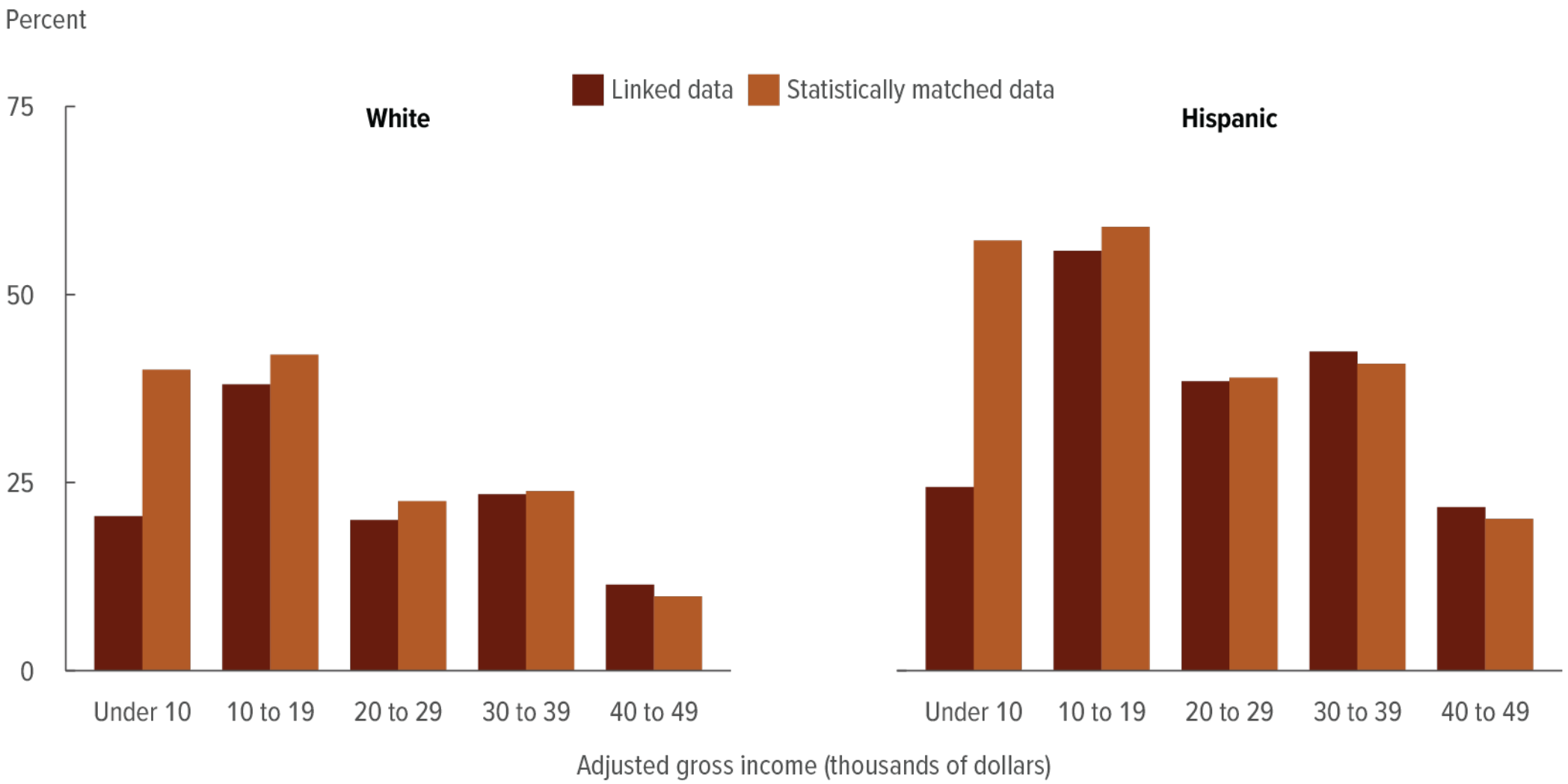
Amount of EITC Received by Tax Units in 2018, by the Race and Ethnicity of the Primary Taxpayer



The estimated amount that tax units received varies by race and ethnicity for each data set, but in total, the estimated amount of EITC that taxpayers received was $62 billion in the linked data set and $65 billion in the statistically matched data set. The IRS reported that filers claimed $65 billion of EITC in 2018.

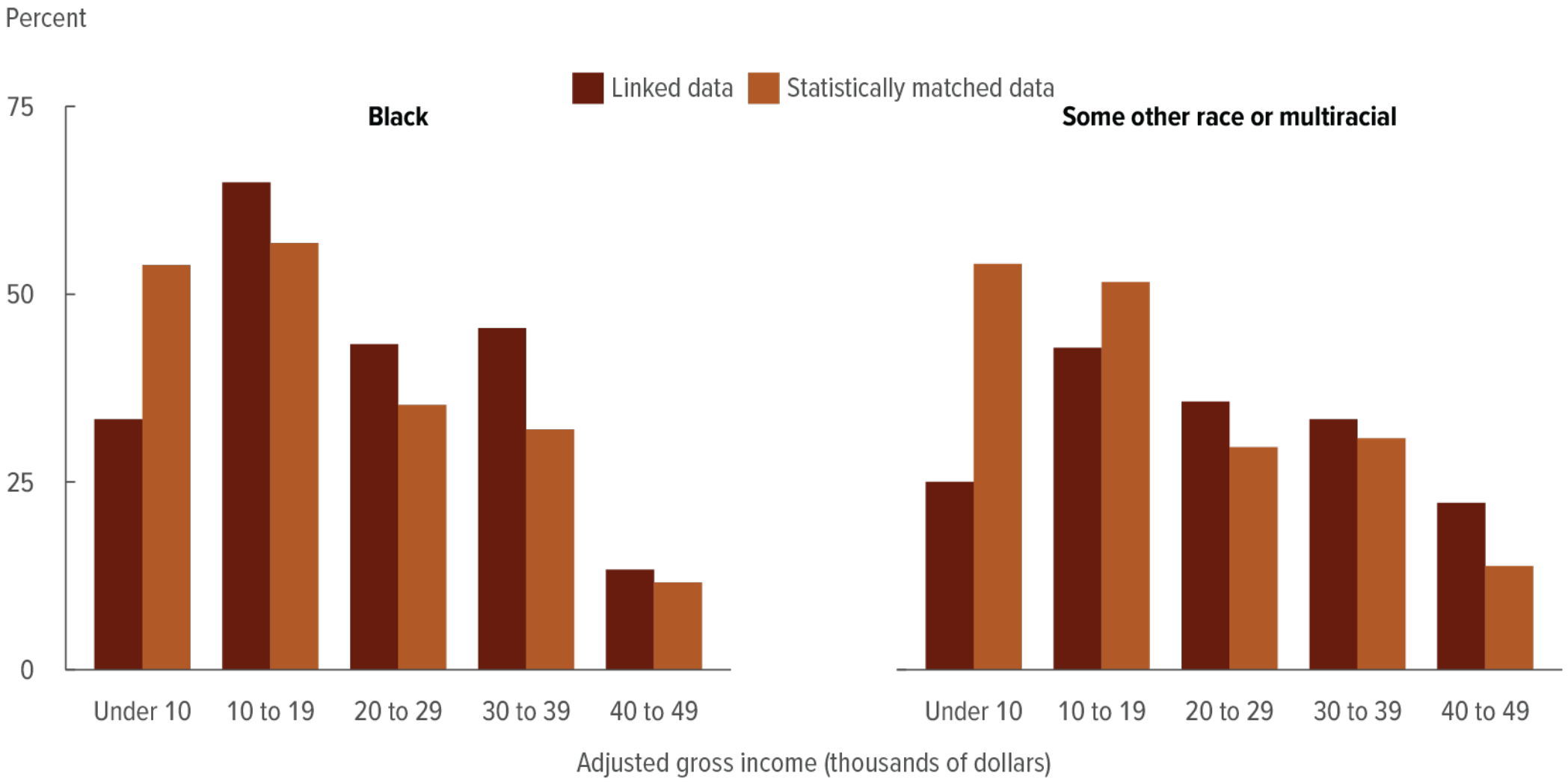
Share of tax units with EITC was restricted to tax units with AGI below $50,000. In both data sets, we calculated the EITC that a taxpayer was eligible for on the basis of earned income, filing status, and the number of EITC-qualifying children reported in tax data. In the linked data, only taxpayers who filed their 2018 tax returns in 2019 were included.

Share of Filing Tax Units Claiming EITC That Have Primary Taxpayers Who Are White or Hispanic, by Income, 2018



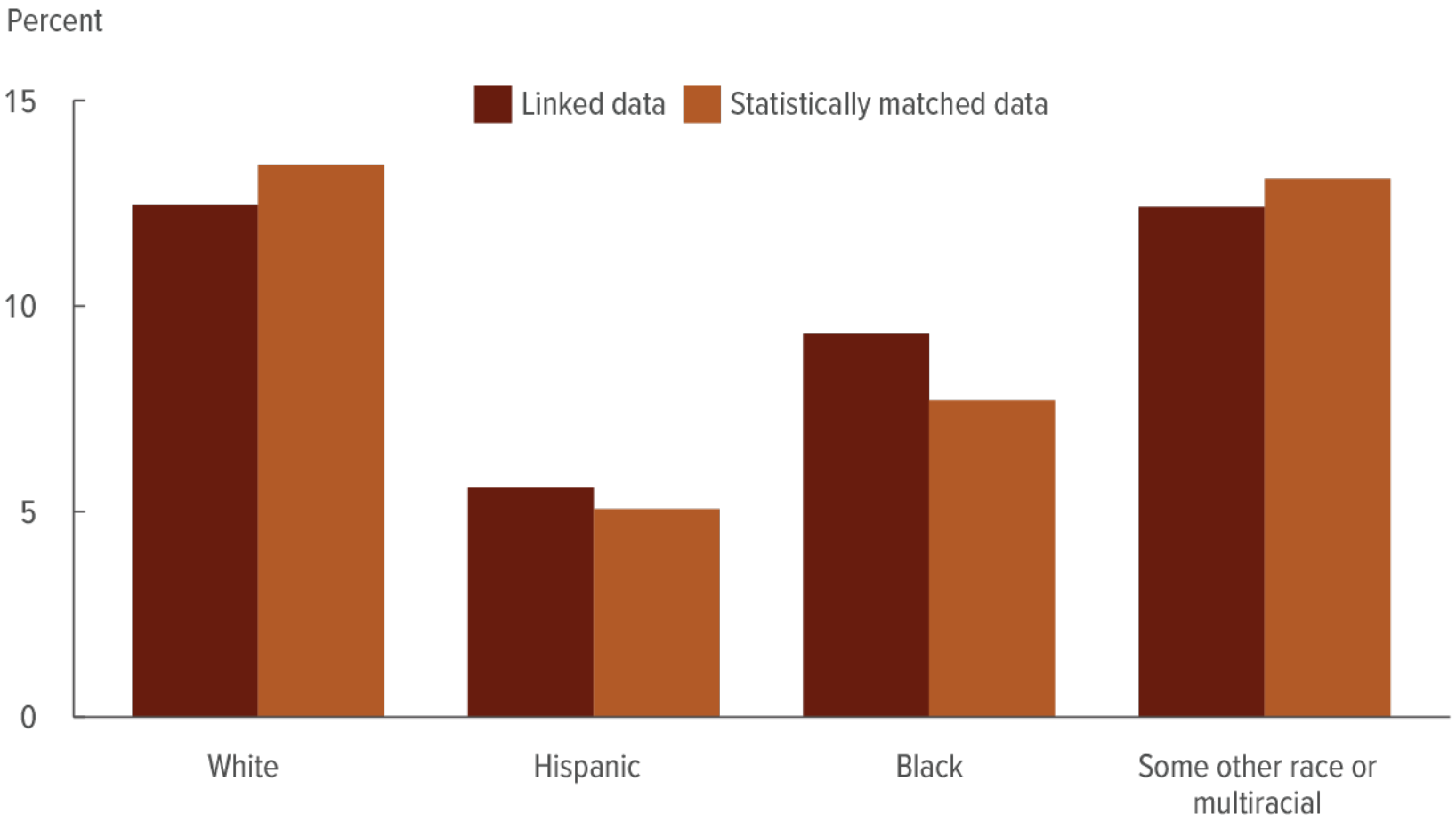
Share of tax units with EITC was restricted to tax units with AGI below $50,000. In both data sets, we calculated the EITC that a taxpayer was eligible for on the basis of earned income, filing status, and the number of EITC-qualifying children reported in tax data. In the linked data, only taxpayers who filed their 2018 tax returns in 2019 were included.

Share of Filing Tax Units Claiming EITC That Have Primary Taxpayers Who Are Black or Some Other Race or Multiracial, by Income, 2018



Share of tax units with EITC was restricted to tax units with AGI below $50,000. In both data sets, we calculated the EITC that a taxpayer was eligible for on the basis of earned income, filing status, and the number of EITC-qualifying children reported in tax data. In the linked data, only taxpayers who filed their 2018 tax returns in 2019 were included.

Share of Filing Tax Units Itemizing Deductions, by the Race and Ethnicity of the Primary Taxpayer, 2018



In the linked data set, 10.9 percent of all filing tax units itemize deductions; that share is higher, at 11.6 percent, in the statistically matched data set.

The share itemizing deductions varied by the race and ethnicity of the primary taxpayer; itemization rates were highest in tax units with a primary taxpayer who identified as White or some other race or multiracial.

In the linked data, only taxpayers who filed their 2018 tax returns in 2019 were included.

Conclusions and Next Steps

Future Work

- Initial results show that the data compare well but that additional research and more formal comparisons are warranted.

- Next steps include:
  - Aggregating tax units into households;
  - Reviewing alternative ways of categorizing households by race and ethnicity for mixed-race households;
  - Producing standard errors and creating more formal comparisons;
  - Comparing the data at a more granular level to understand how household members organize into tax units to improve the statistical matching algorithm; and
  - Evaluating whether the statistical match preserves the distribution of other demographic characteristics (for example, educational attainment).

- Future work will be released as a CBO working paper

References

Brown, Dorothy. 2021. *The Whiteness of Wealth: How the Tax System Impoverishes Black Americans—And How We Can Fix It*. New York: Crown.

Congressional Budget Office. 2017. “Statistically Matching Administrative Tax Data With Household Survey Data.” Presentation at a workshop organized by the Washington Center for Equitable Growth, Washington D.C. www.cbo.gov/publication/52914.

Congressional Budget Office. 2022. Letter to Chairman John Yarmuth, “Re: Analyzing How the Effects of Federal Policies May Differ by Race and Ethnicity.” www.cbo.gov/publication/58030.

Cronin, Julie-Anne, Portia DeFilippes, and Robin Fisher. 2023. “Tax Expenditures by Race and Hispanic Ethnicity: An Application of the U.S. Treasury Department’s Race and Hispanic Ethnicity Imputation.” Office of Tax Analysis Working Paper 122. https://home.treasury.gov/system/files/131/WP-122.pdf.

Fisher, Robin. 2023. “Estimation of Race and Ethnicity by Re-Weighting Tax Data.” Office of Tax Analysis Technical Paper 11. https://home.treasury.gov/system/files/131/TP-11.pdf.

References (Continued)

Goldin, Jacob, and Katherine Michelmore. 2022. “Who Benefits from the Child Tax Credit?” *National Tax Journal*, 75(1): 123–147. www.journals.uchicago.edu/doi/full/10.1086/717919.

Holtzblatt, Janet, Swati Joshi, Nora Cahill, and William G. Gale. 2023. “Racial Disparities in the Income Tax Treatment of Marriage.” Tax Policy Center Working Paper. https://tinyurl.com/4wr6an8m.

Meyer, Bruce D., Derek Wu, Grace Finley, Patrick Langetieg, Carla Medalia, Mark Payne, and Alan Plumley. 2022. “The Accuracy of Tax Imputations: Estimating Tax Liabilities and Credits Using Linked Survey and Administrative Data.” In *Measuring Distribution and Mobility of Income and Wealth*, edited by Raj Chetty, John Friedman, Janet Gornick, Barry Johnson, and Arthur Kennickell, 459–498. Chicago: University of Chicago Press. http://tinyurl.com/ms79vcpp.

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.” CARRA Working Paper 2014-01. https://tinyurl.com/ynn6yuzf.

Correction

The Congressional Budget Office has corrected this presentation since its original publication.

On January 31, 2024, the following changes were made:

Slides 17–24: labeling on the x-axis was adjusted. None of the values in the presentation were affected.