# Non-Random Assignment of Individual Identifiers and Selection into Linked Data

## Implications for Research

Kyle Raze
*kyle.raze@census.gov*
*Center for Economic Studies*

Nicole Perales
*nicole.perales@census.gov*
*Center for Economic Studies*

Christin Landivar
*liana.c.landivar@census.gov*
*Social, Economic, and Housing Statistics Division*

October 24, 2024

DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Disclaimer

*Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection* (Project 7506072: CBDRB-FY24-CES027-002, CBDRB-FY24-CES027-006).

# Motivation

Advances in the ability to link survey data to administrative records have generated wide-ranging benefits for measurement and inference

- Reducing measurement error from non-response, imputation, and misreporting (e.g., Bollinger, Hirsch, Hokayem, and Ziliak, 2019; Meyer, Mittag, and Goerge, 2022)
- Facilitating analyses of novel outcomes (e.g., Chetty, Hendren, Jones, and Porter, 2020)

# Motivation

Advances in the ability to link survey data to administrative records have generated wide-ranging benefits for measurement and inference

- Reducing measurement error from non-response, imputation, and misreporting (e.g., Bollinger, Hirsch, Hokayem, and Ziliak, 2019; Meyer, Mittag, and Goerge, 2022)
- Facilitating analyses of novel outcomes (e.g., Chetty, Hendren, Jones, and Porter, 2020)

Protected Identification Keys (PIKs) are unique anonymous identifiers that allow researchers to link individuals across data sets housed at the Census Bureau

- Not all individuals can be assigned a PIK

# Motivation

Advances in the ability to link survey data to administrative records have generated wide-ranging benefits for measurement and inference

- Reducing measurement error from non-response, imputation, and misreporting (e.g., Bollinger, Hirsch, Hokayem, and Ziliak, 2019; Meyer, Mittag, and Goerge, 2022)
- Facilitating analyses of novel outcomes (e.g., Chetty, Hendren, Jones, and Porter, 2020)

Protected Identification Keys (PIKs) are unique anonymous identifiers that allow researchers to link individuals across data sets housed at the Census Bureau

- Not all individuals can be assigned a PIK

**Selection into PIK assignment is likely non-random** (Bond, Brown, Luque, and O'Hara, 2014)

- Can compromise the representativeness of linked data, leading to biased population estimates
- What should researchers do about it?

# Background

The Person Identification Validation System (PVS) assigns PIKs to individuals

- PVS matches individual records in an "incoming file" (e.g., a survey) to a "reference file" using a series of cascading probabilistic modules (see Layne and Wagner, 2014 for details)
- Reference file $\approx$ crosswalk between universe of SSNs (with identifying information) and PIKs

Improvements in PVS have increased PIK rates (Bond et al., 2014)

- New modules
- Inclusion of Individual Taxpayer Identification Numbers (ITINs) in the reference file

# Background

The Person Identification Validation System (PVS) assigns PIKs to individuals

- PVS matches individual records in an "incoming file" (e.g., a survey) to a "reference file" using a series of cascading probabilistic modules (see Layne and Wagner, 2014 for details)
- Reference file ≈ crosswalk between universe of SSNs (with identifying information) and PIKs

Improvements in PVS have increased PIK rates (Bond et al., 2014)

- New modules
- Inclusion of Individual Taxpayer Identification Numbers (ITINs) in the reference file

PIK rates have been shown to vary by race, Hispanic origin, citizenship, mobility, age, and socioeconomic status (Bond et al., 2014; Meyer and Goerge, 2011; Mulrow, Mushtaq, Pramanik, and Fontes, 2011; Bollinger et al., 2019)

# Research objectives

1. Document variation in PIK rates in household surveys
2. Quantify the magnitude of linkage-induced bias
   - Bias = Difference between a restricted-sample (e.g., PIKed respondents) mean and a full-sample (“target”) mean
3. Evaluate the performance of bias mitigation methods used in the literature
   - *Most common:* Inverse probability weighting (IPW)
   - Ongoing work to incorporate newer state-of-the-art methods

# Data

American Community Survey (ACS), 2005-2022

- Large nationally representative household survey with many social, demographic, economic, and housing characteristics
- To assign PIKs, PVS probabilistically matches names, dates of birth, sex, and addresses to reference files

# Data

American Community Survey (ACS), 2005-2022

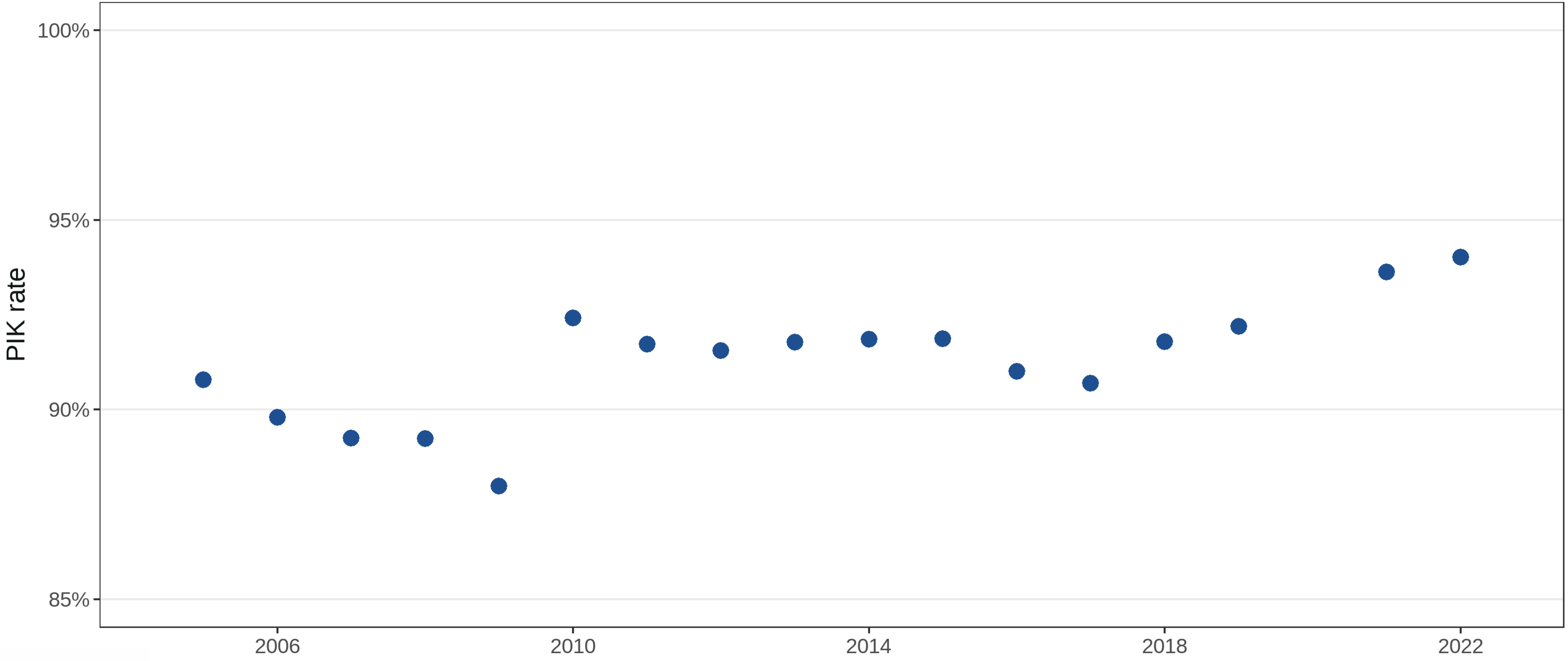
- Large nationally representative household survey with many social, demographic, economic, and housing characteristics
- To assign PIKs, PVS probabilistically matches names, dates of birth, sex, and addresses to reference files

Internal Revenue Service Form W-2 records (W-2s), 2005-2022

- Near-full coverage of formally employed workers
- Source of administrative records for an actual linkage
  - Not all ACS respondents are linked due to differences in PIK assignment or misalignment of the target population across data sources

# PIK rates

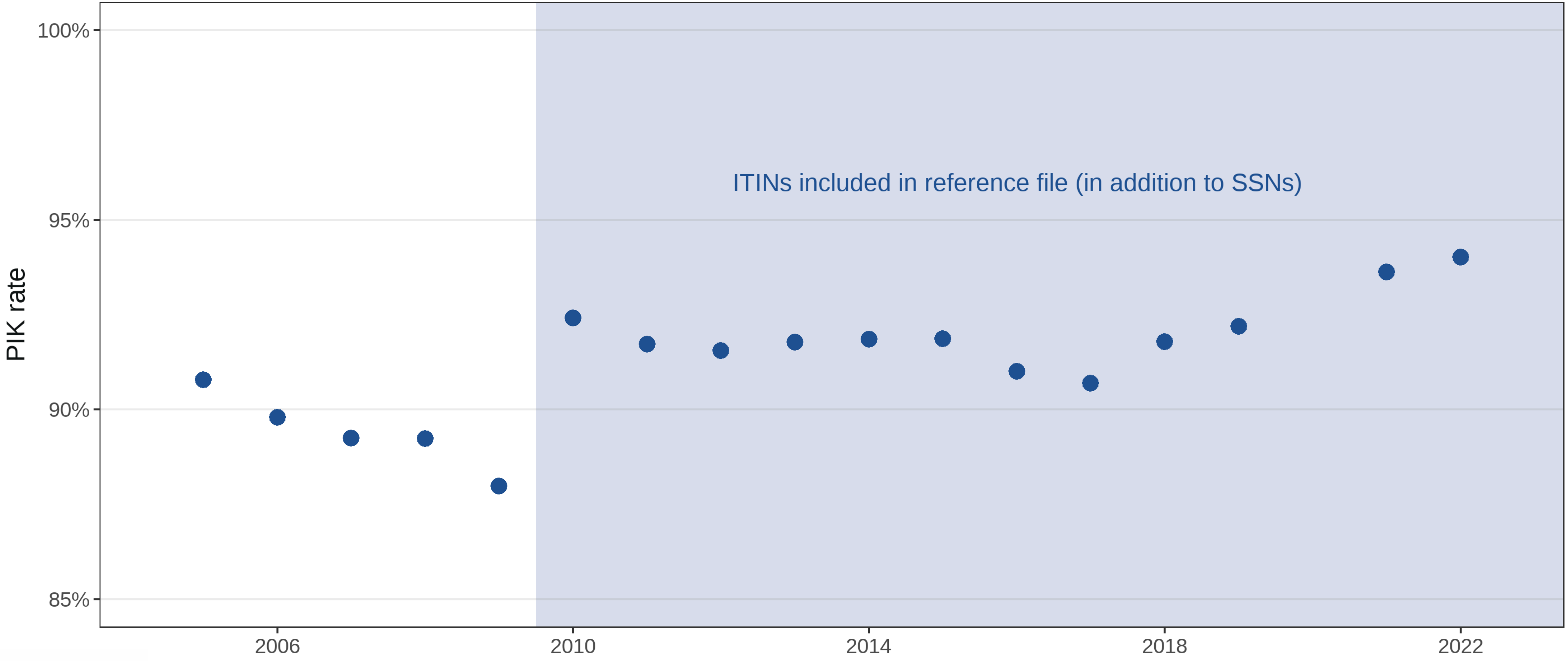
## ACS respondents




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

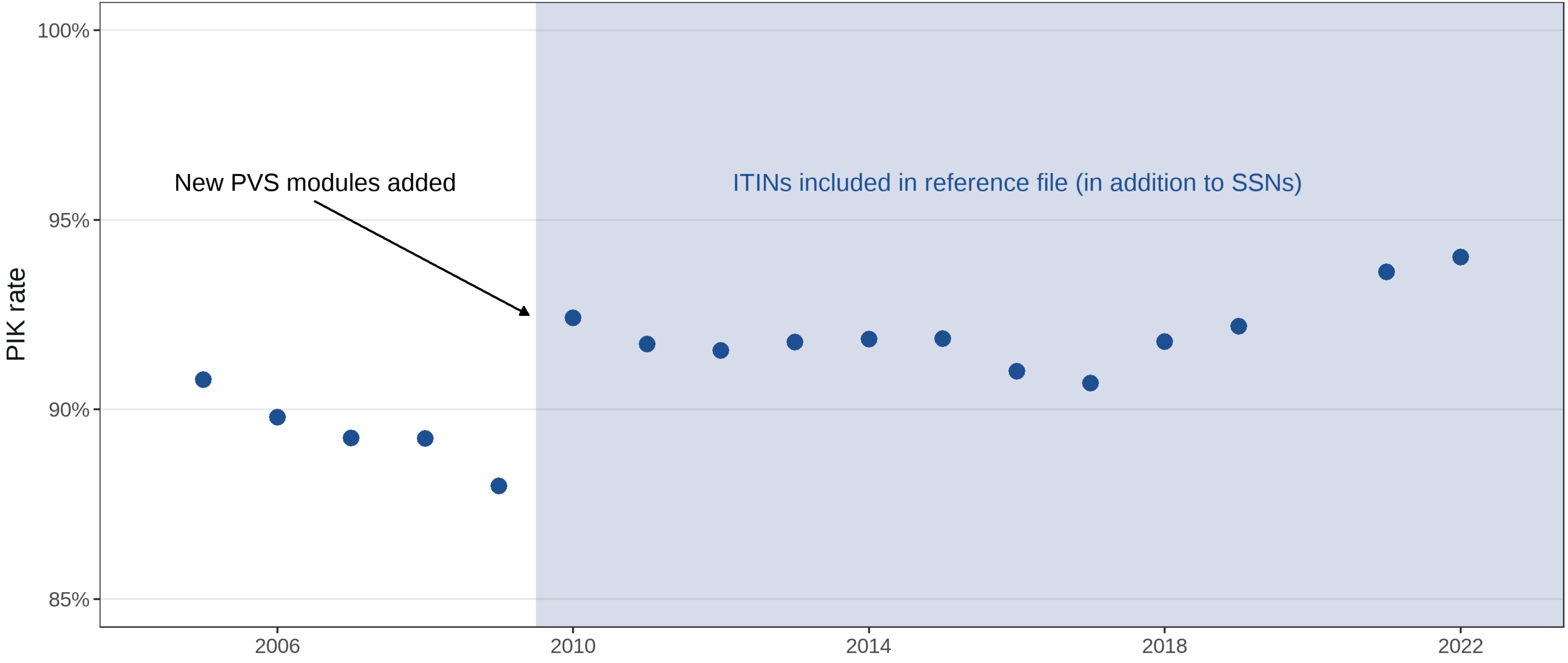
# PIK rates

## ACS respondents



# PIK rates

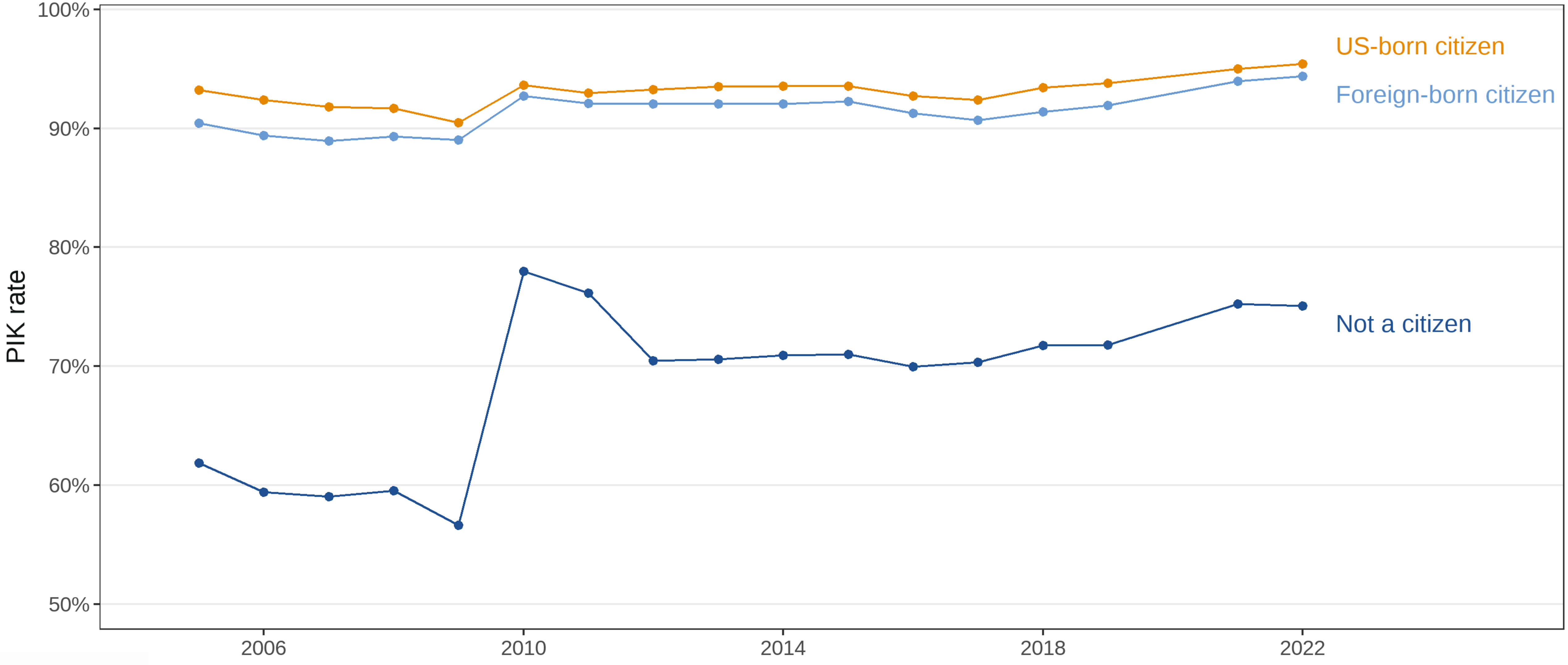
## ACS respondents




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

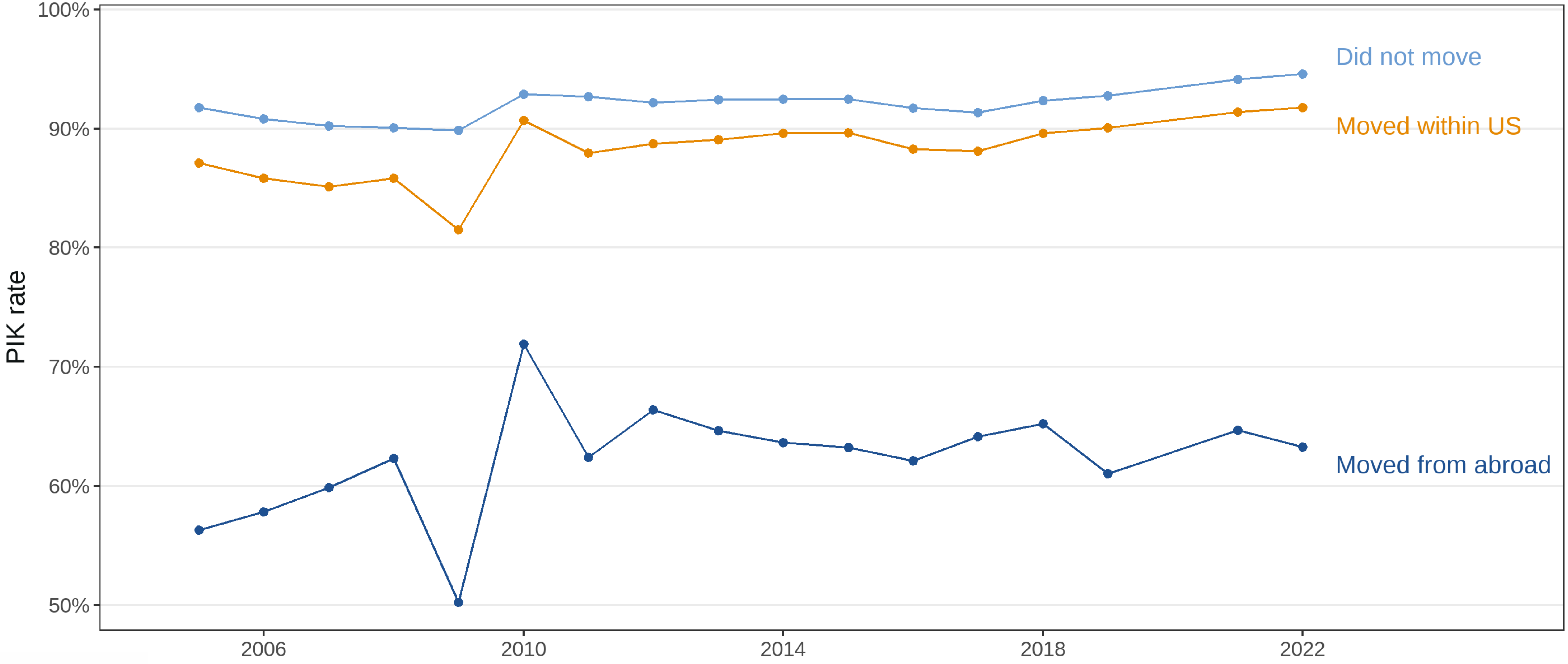
# PIK rates by demographic characteristics

Citizenship Migration Race/ethnicity Age Education



# PIK rates by demographic characteristics

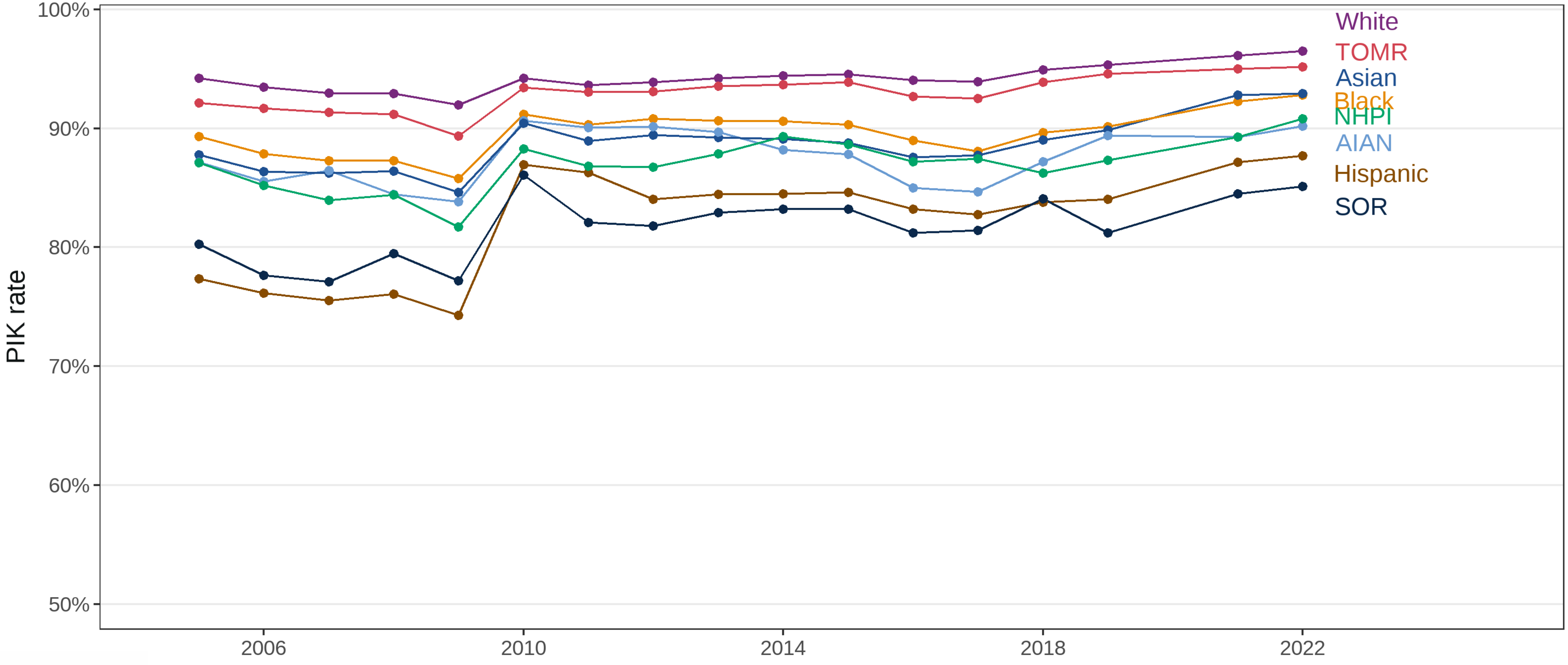
Citizenship Migration Race/ethnicity Age Education




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# PIK rates by demographic characteristics

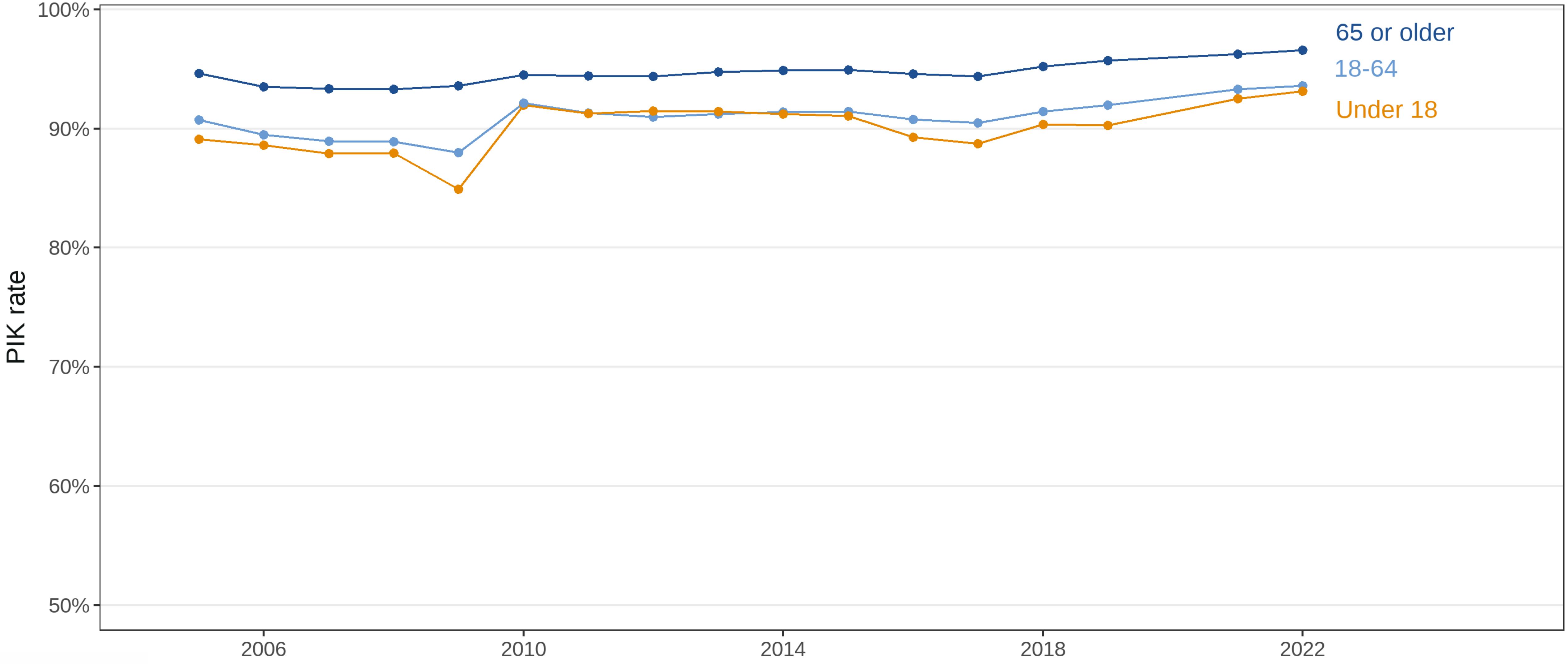
Citizenship Migration Race/ethnicity Age Education




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

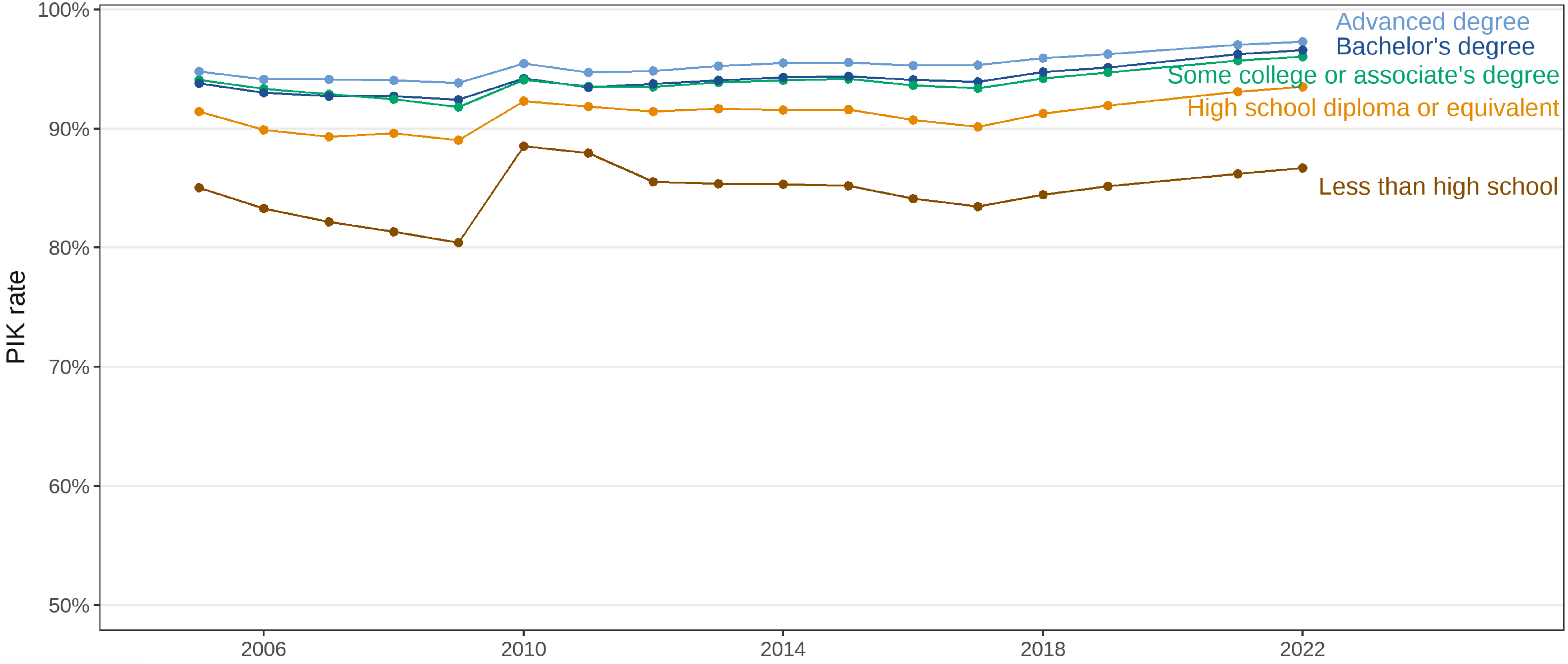
# PIK rates by demographic characteristics

Citizenship Migration Race/ethnicity **Age** Education



# PIK rates by demographic characteristics

Citizenship Migration Race/ethnicity Age **Education**




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias

**Definition**

$$\text{Linkage-induced bias} := \mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i)$$

- $y_i$ is the outcome of respondent $i \in \{1, \ldots, n\}$
- $z_i = 1$ if $i$ has a PIK
- $z_i = 0$ if $i$ does not have a PIK

# Linkage-induced bias

**Definition**

$$\text{Linkage-induced bias} := \mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i)$$

- $y_i$ is the outcome of respondent $i \in \{1, \ldots, n\}$
- $z_i = 1$ if $i$ has a PIK
- $z_i = 0$ if $i$ does not have a PIK

**Bias correction:** Wooldridge (2007) shows that Inverse Probability Weighting (IPW) estimation for missing data problems is consistent under selection-on-observables

- Meyer and Goerge (2011) and Bollinger et al. (2019) invoke selection-on-observables and use IPW to recover respresentative samples from linked data
- But IPW can be biased, inefficient, or unstable in finite samples (Busso, DiNardo, and McCrary, 2014; Li, Qin, and Liu, 2023; Liu and Fan, 2023)

# IPW steps

1. Specify a model of selection into linkage
2. Estimate the selection equation and obtain propensity scores
3. Calculate IPW weight = 1 / propensity score
4. Reweight the linked sample by multiplying IPW weights with survey weights
5. Estimate an outcome equation using the reweighted linked sample

# Evaluating the performance of IPW

**Question:** Does reweighting reduce linkage-induced bias?

**Approach:** Compare the means of linked samples and reweighted linked samples to the mean of the target sample

# Evaluating the performance of IPW

**Question:** Does reweighting reduce linkage-induced bias?

**Approach:** Compare the means of linked samples and reweighted linked samples to the mean of the target sample

- Survey outcome: Wage income
- Samples
    1. *Target:* ACS wage earners (government and private-sector workers only)
    2. *PIKed:* Target sample restricted to PIKed respondents
    3. *IPW:* PIKed sample reweighted using IPW
- Propensity scores from a logistic regression of PIK assignment on a “typical” set of predictors
    - “Basic” (e.g., observable in administrative records): sex + race/ethnicity + quartic in age
    - “Full” (e.g., only observable in surveys): “basic” + citizenship + English ability + interview mode + migration in the last year + educational attainment + marital status + disability status + region + urban/rural indicator

# Evaluating the performance of IPW

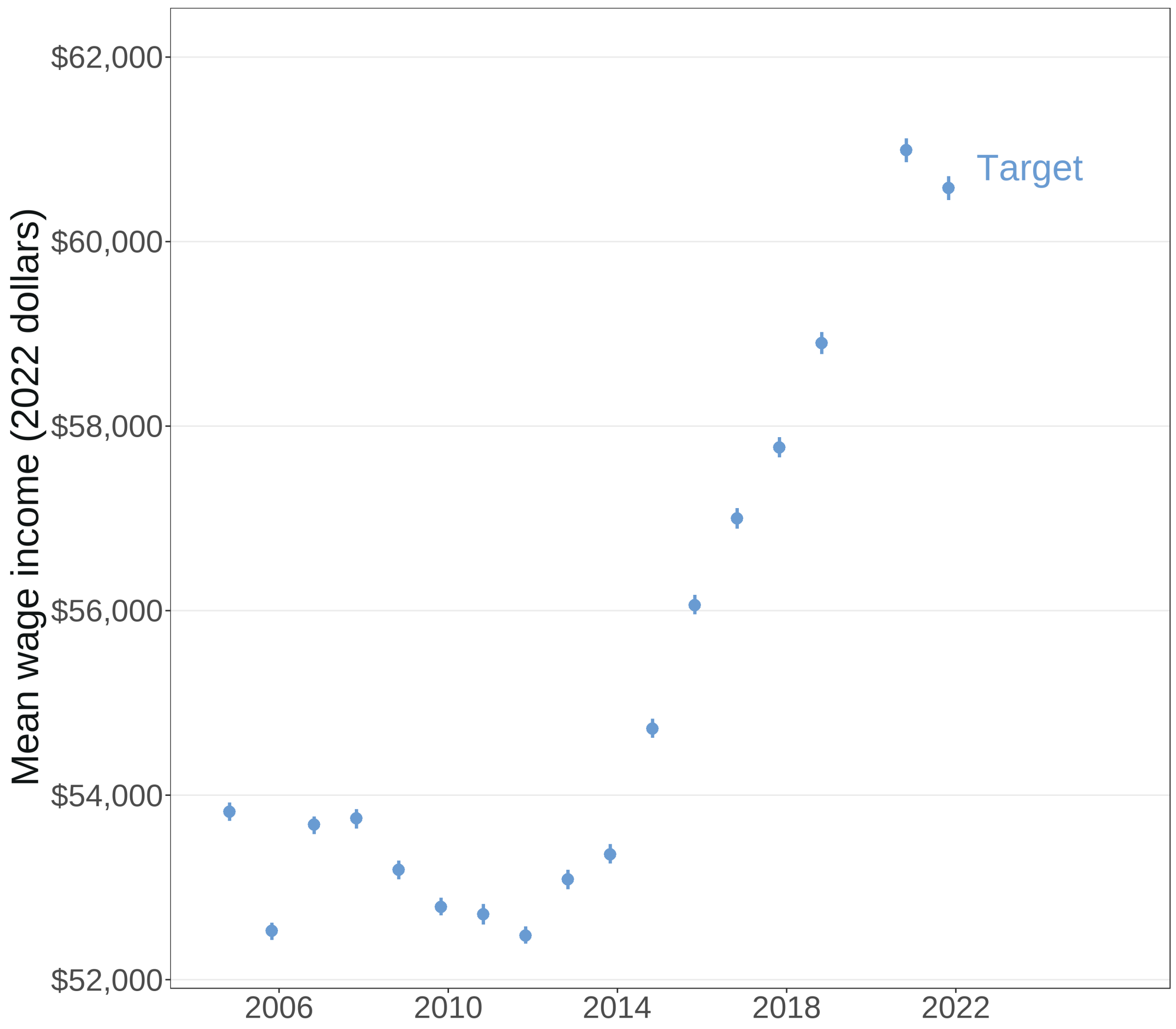
**Question:** Does reweighting reduce linkage-induced bias?

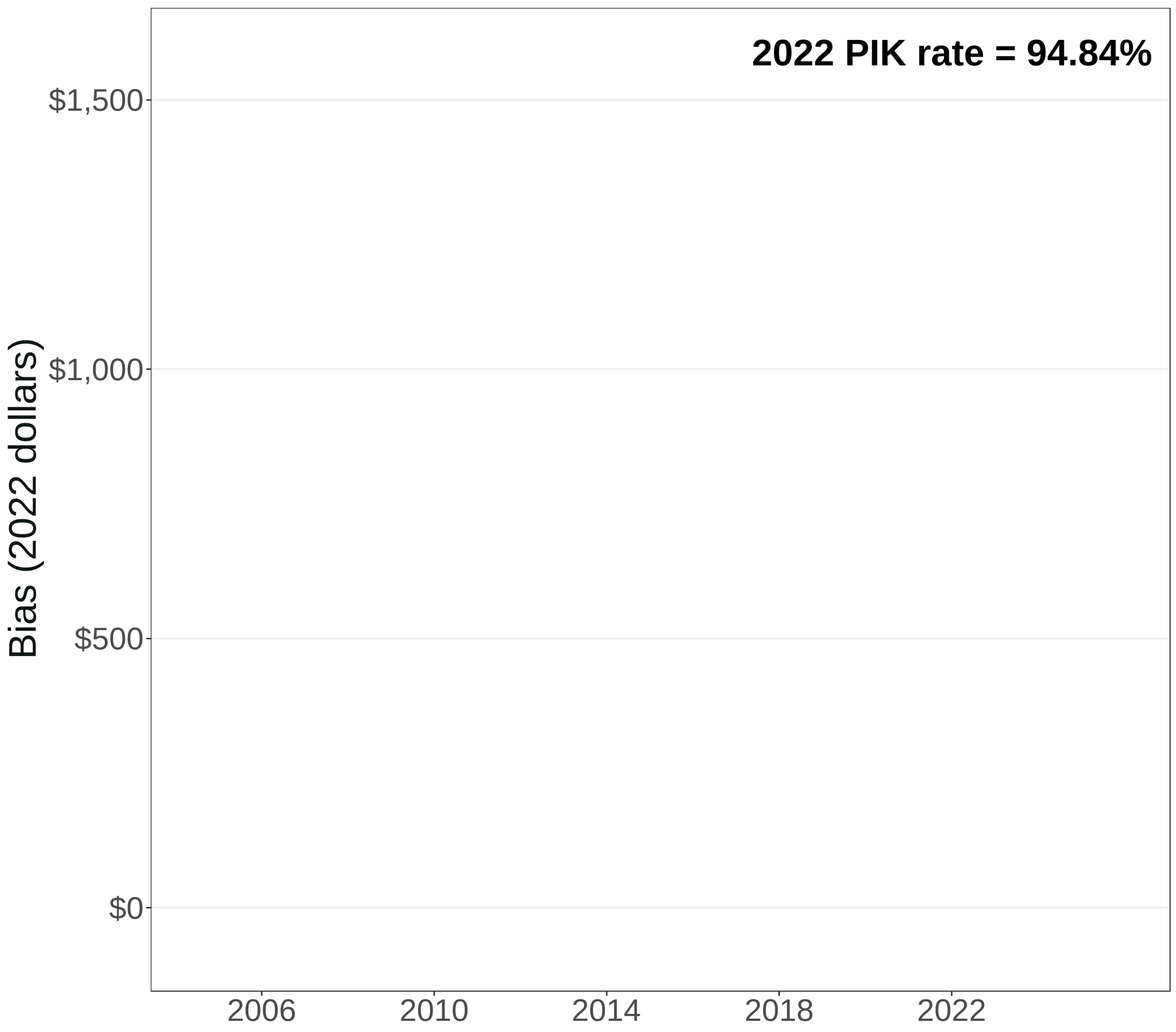
**Approach:** Compare the means of linked samples and reweighted linked samples to the mean of the target sample

- Survey outcome: Wage income
- Samples
    1. *Target:* ACS wage earners (government and private-sector workers only)
    2. *Linked:* Target sample restricted to linked respondents
    3. *IPW:* Linked sample reweighted using IPW
- Propensity scores from a logistic regression of W-2 linkage on a “typical” set of predictors
    - “Basic” (e.g., observable in administrative records): sex + race/ethnicity + quartic in age
    - “Full” (e.g., only observable in surveys): “basic” + citizenship + English ability + interview mode + migration in the last year + educational attainment + marital status + disability status + region + urban/rural indicator

# PIK-induced bias in wage income

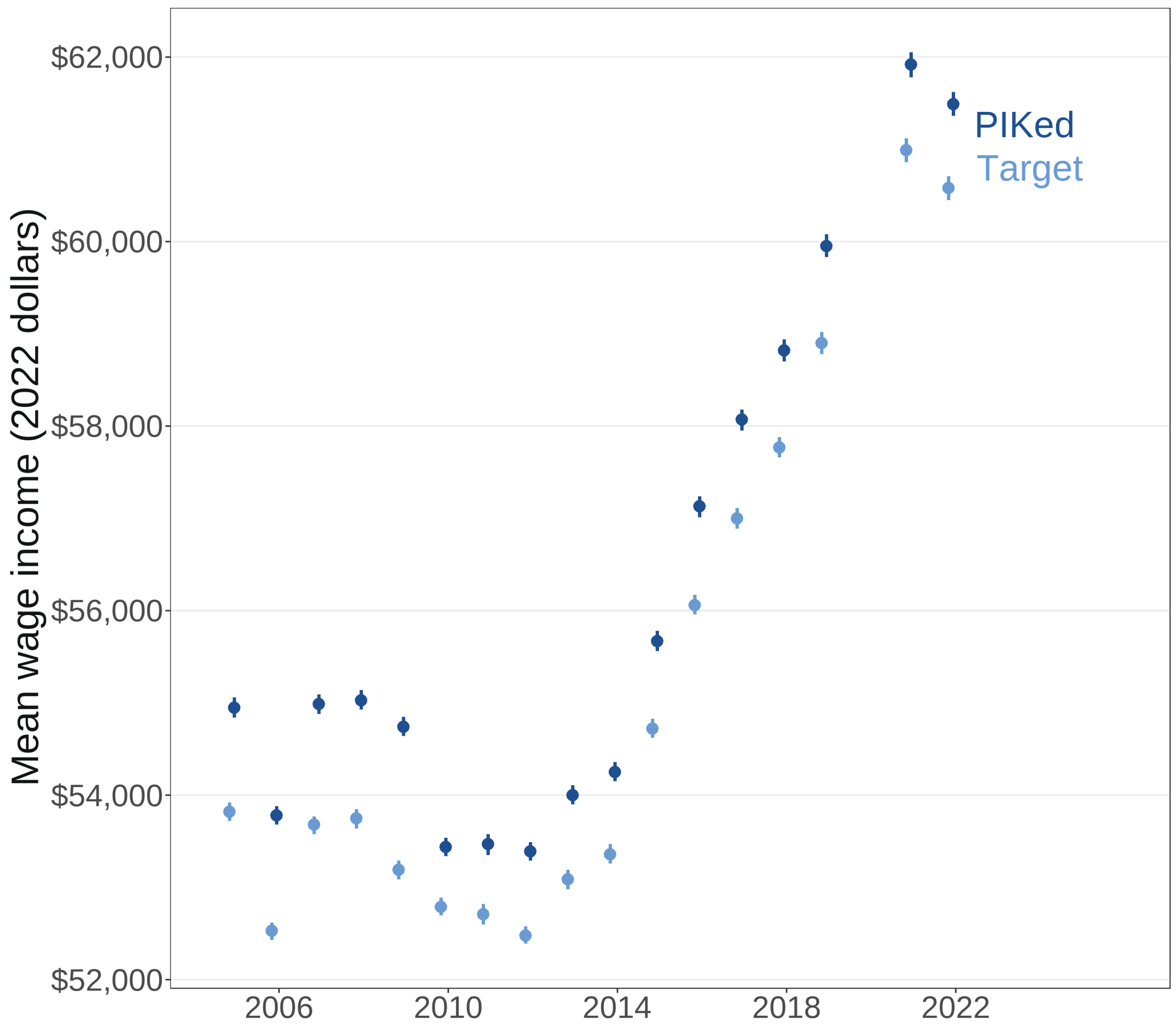
## Private-sector and government wage earners (ACS)

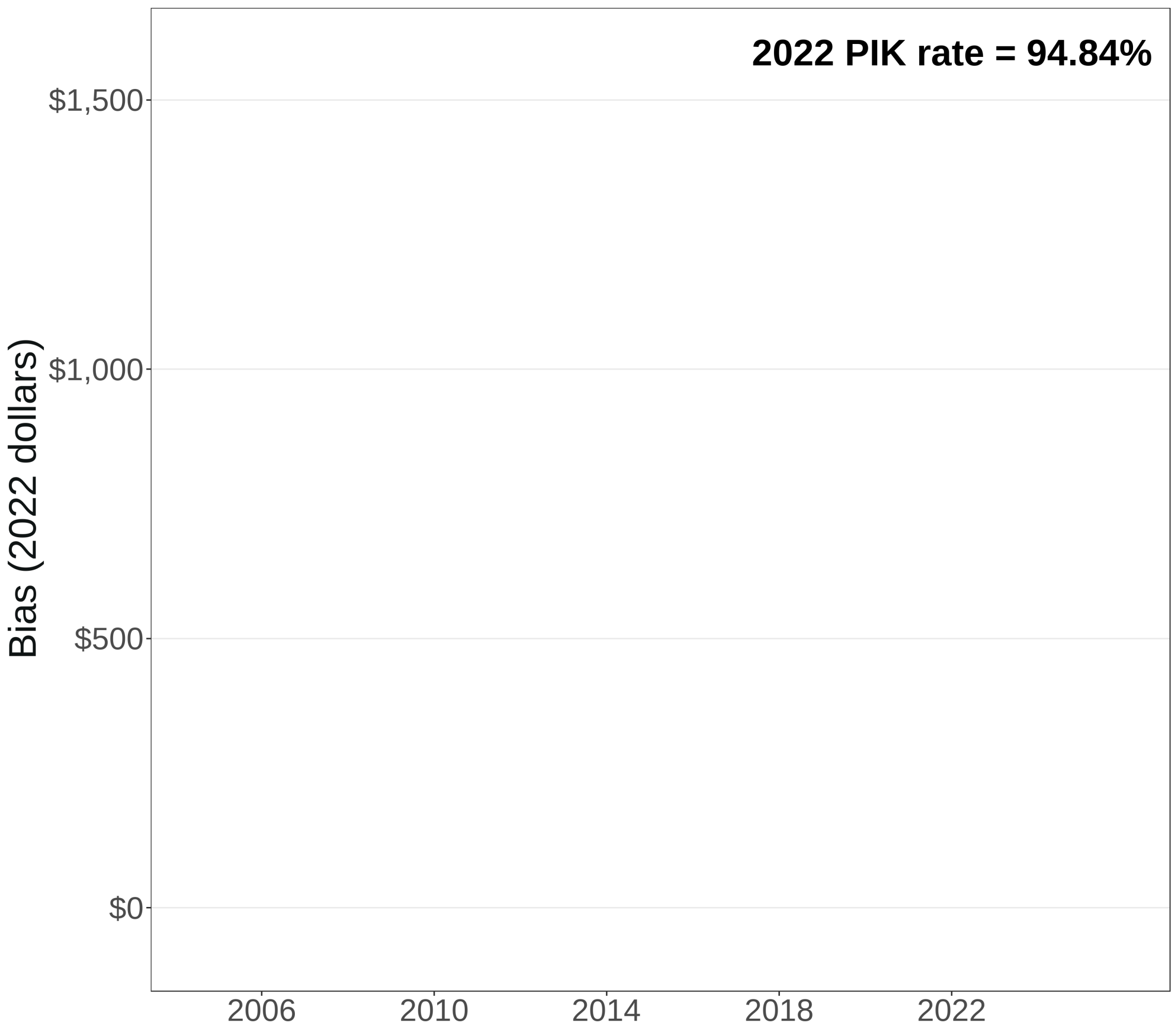


# PIK-induced bias in wage income

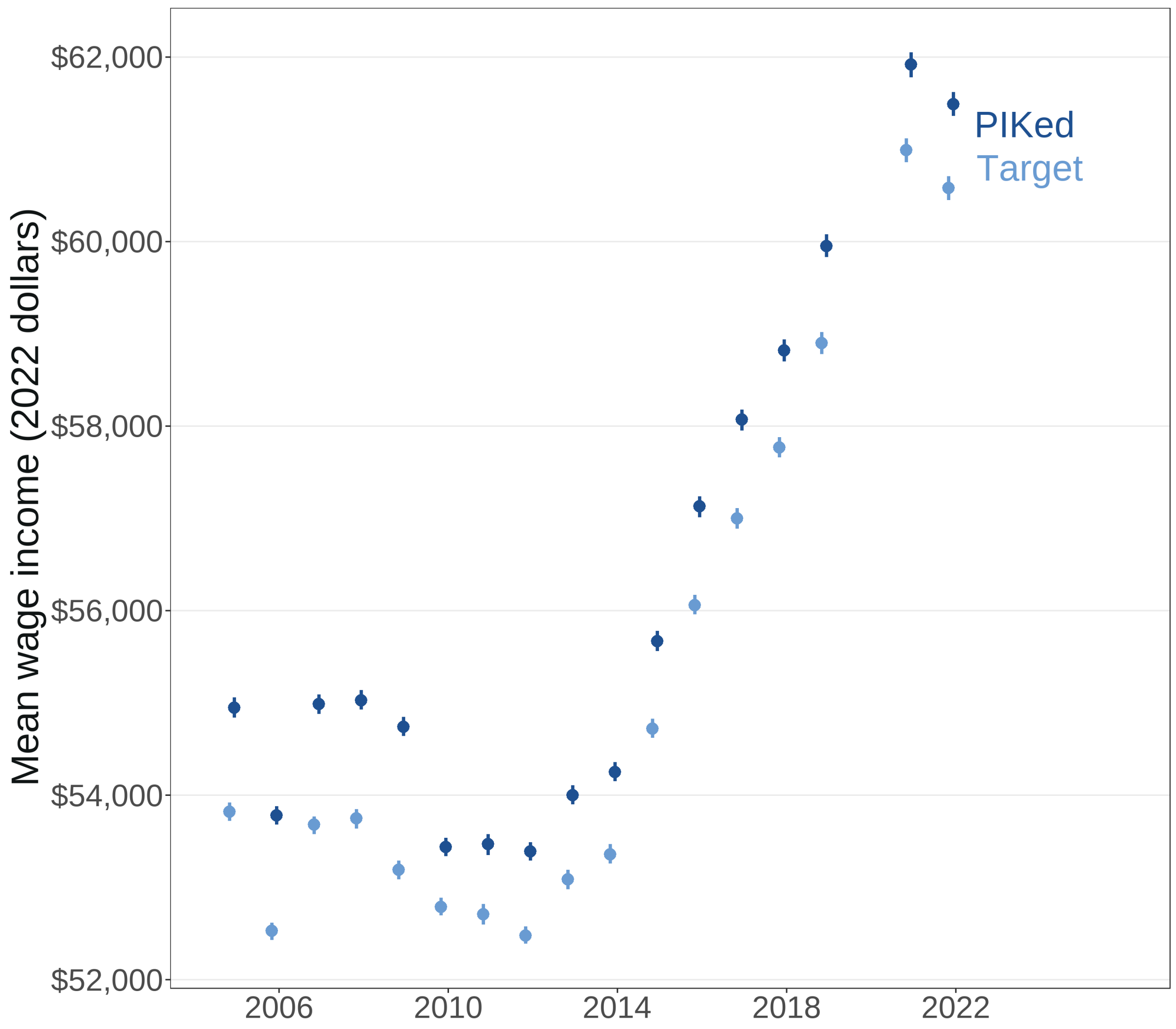
## Private-sector and government wage earners (ACS)

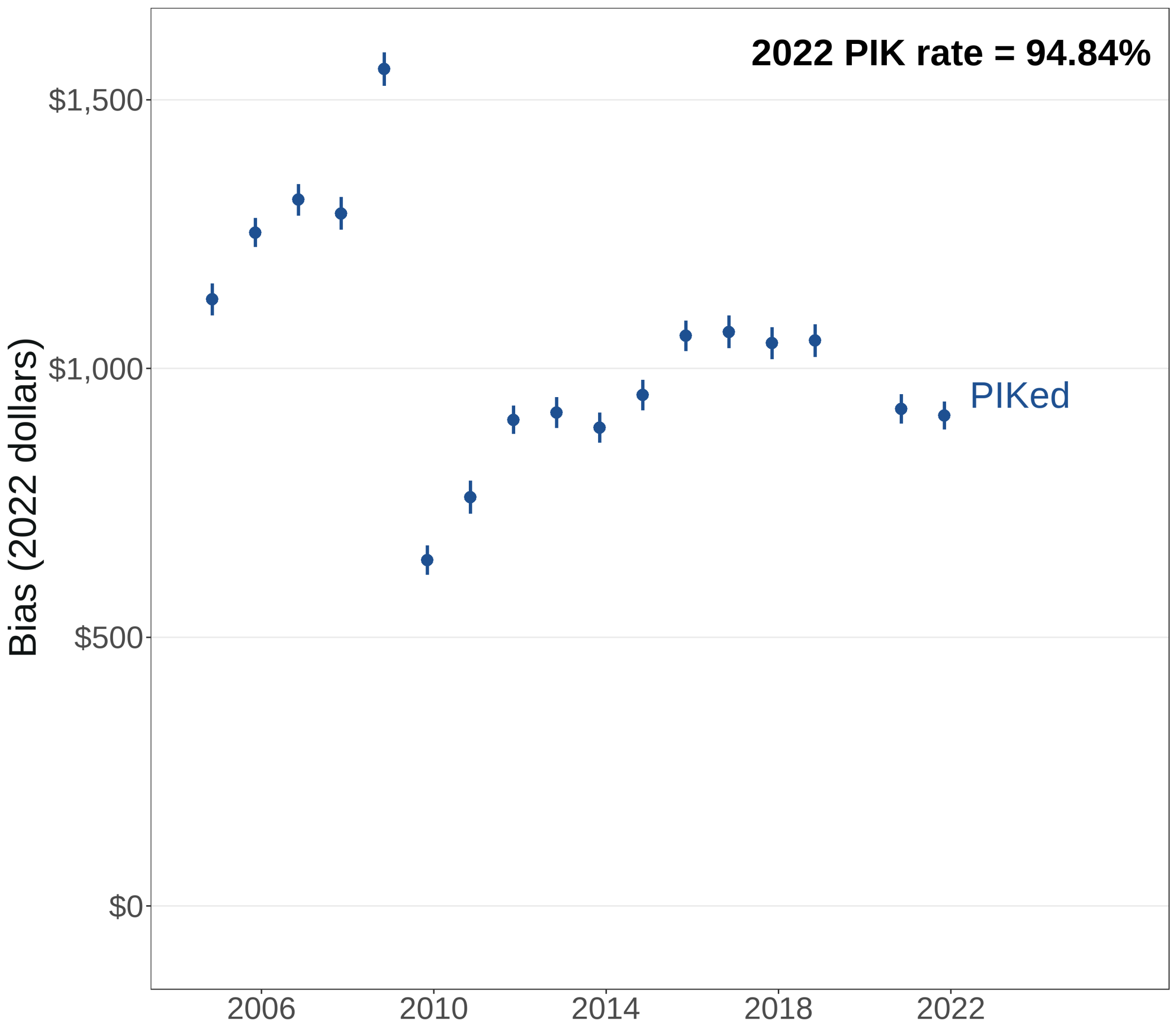


# PIK-induced bias in wage income

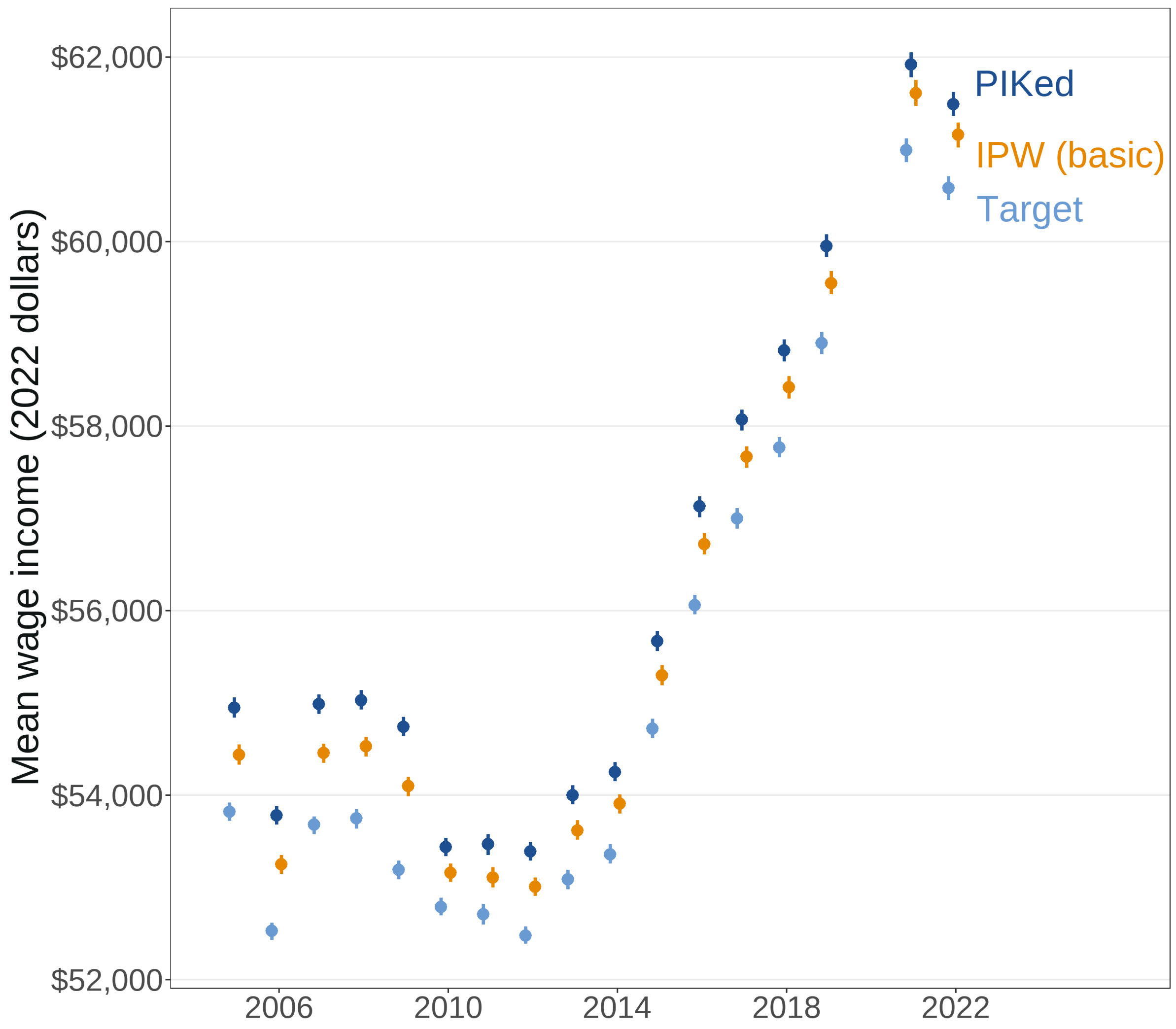
## Private-sector and government wage earners (ACS)

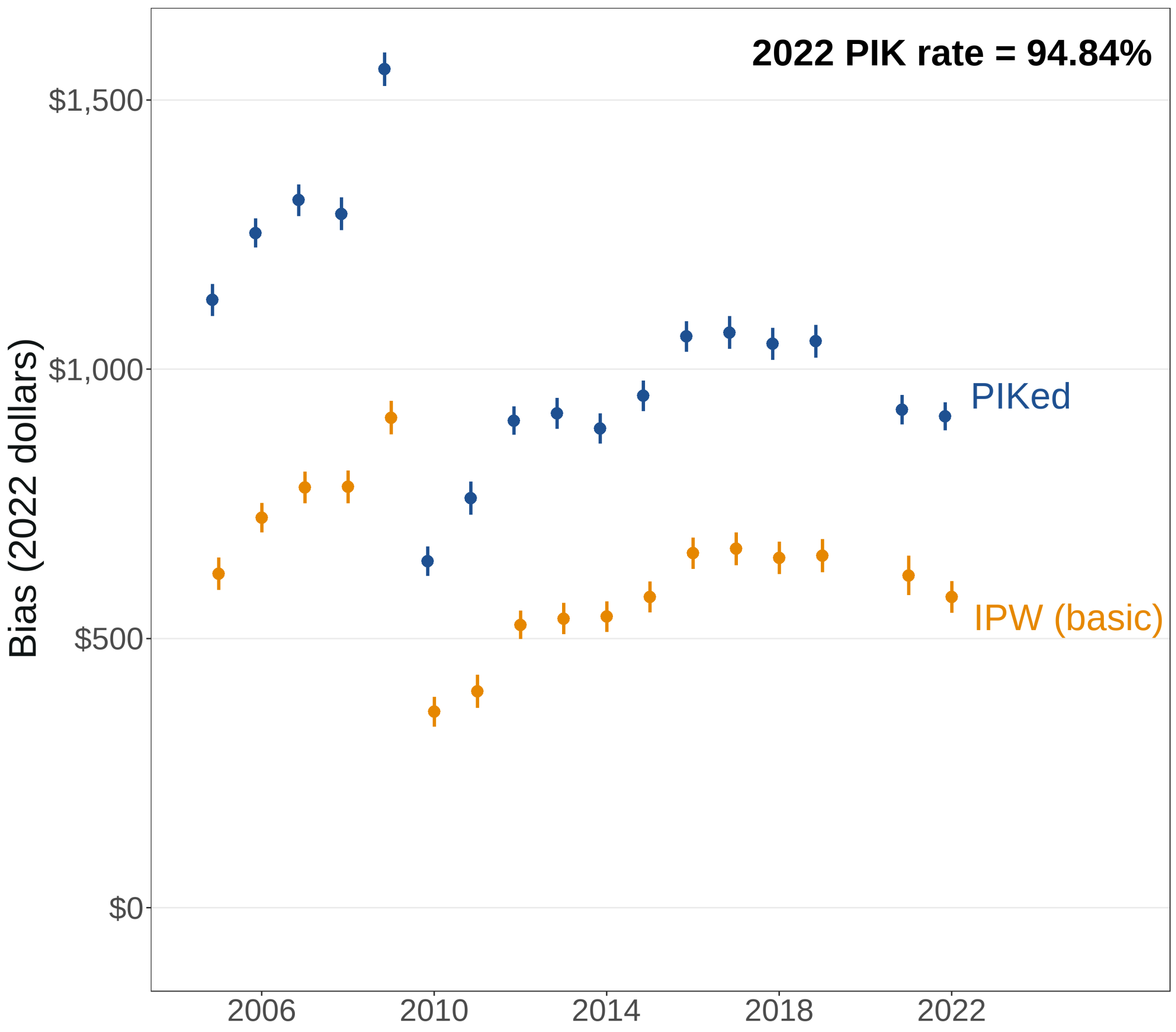


# PIK-induced bias in wage income

## Private-sector and government wage earners (ACS)

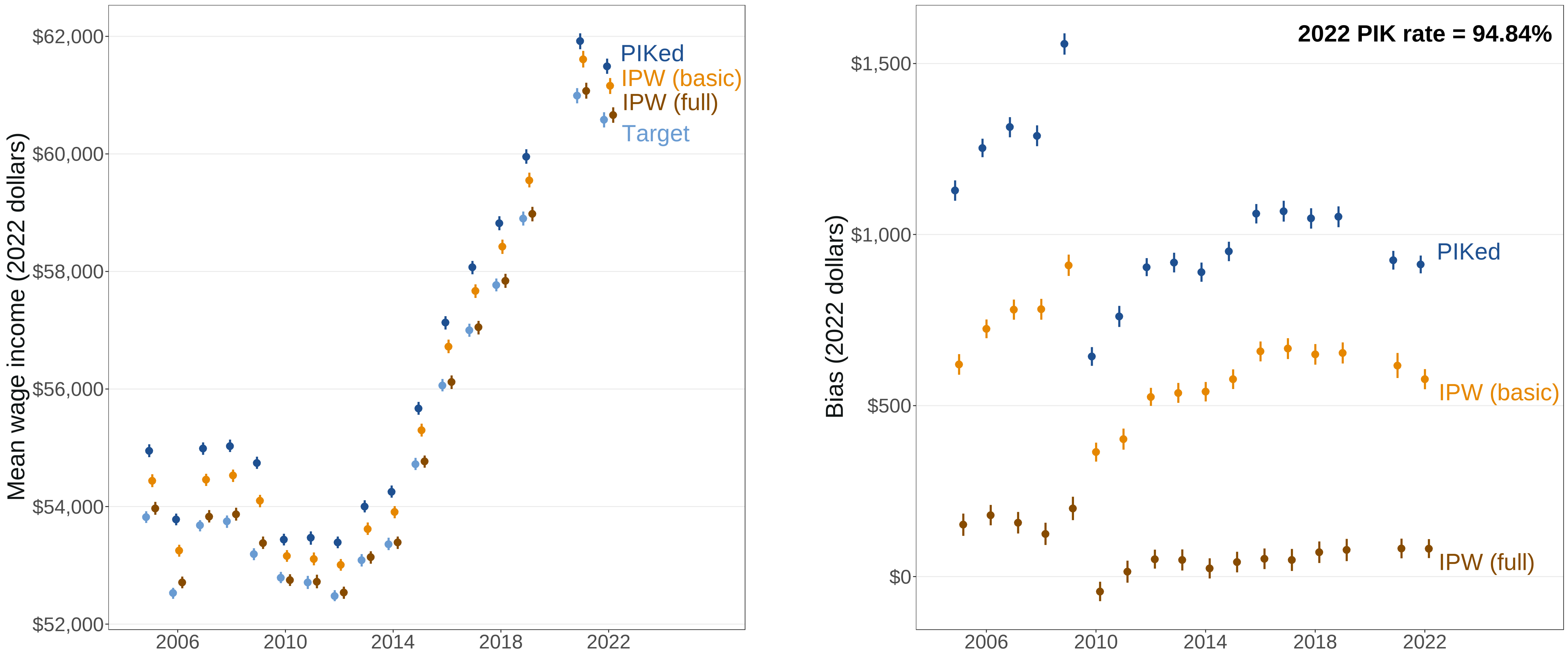



DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# PIK-induced bias in wage income

## Private-sector and government wage earners (ACS)

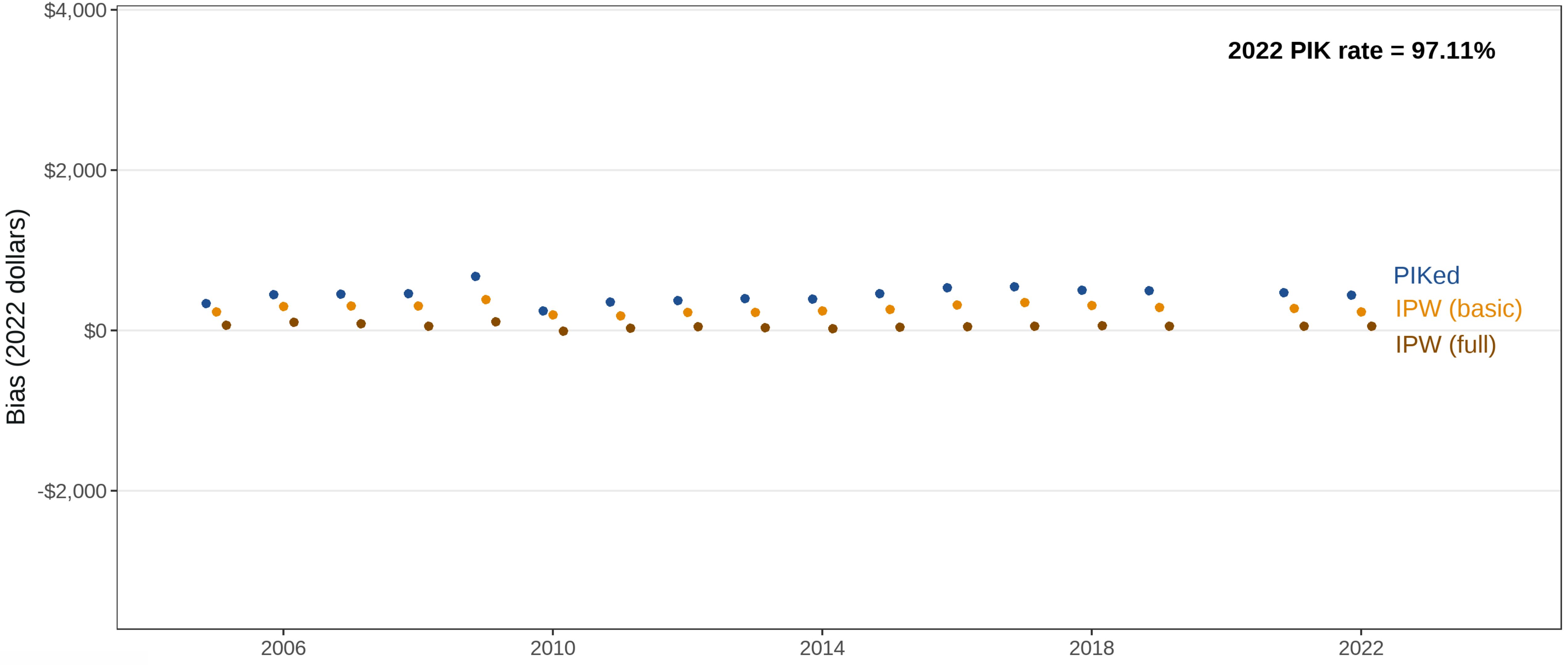





DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# PIK-induced bias in wage income by race/ethnicity

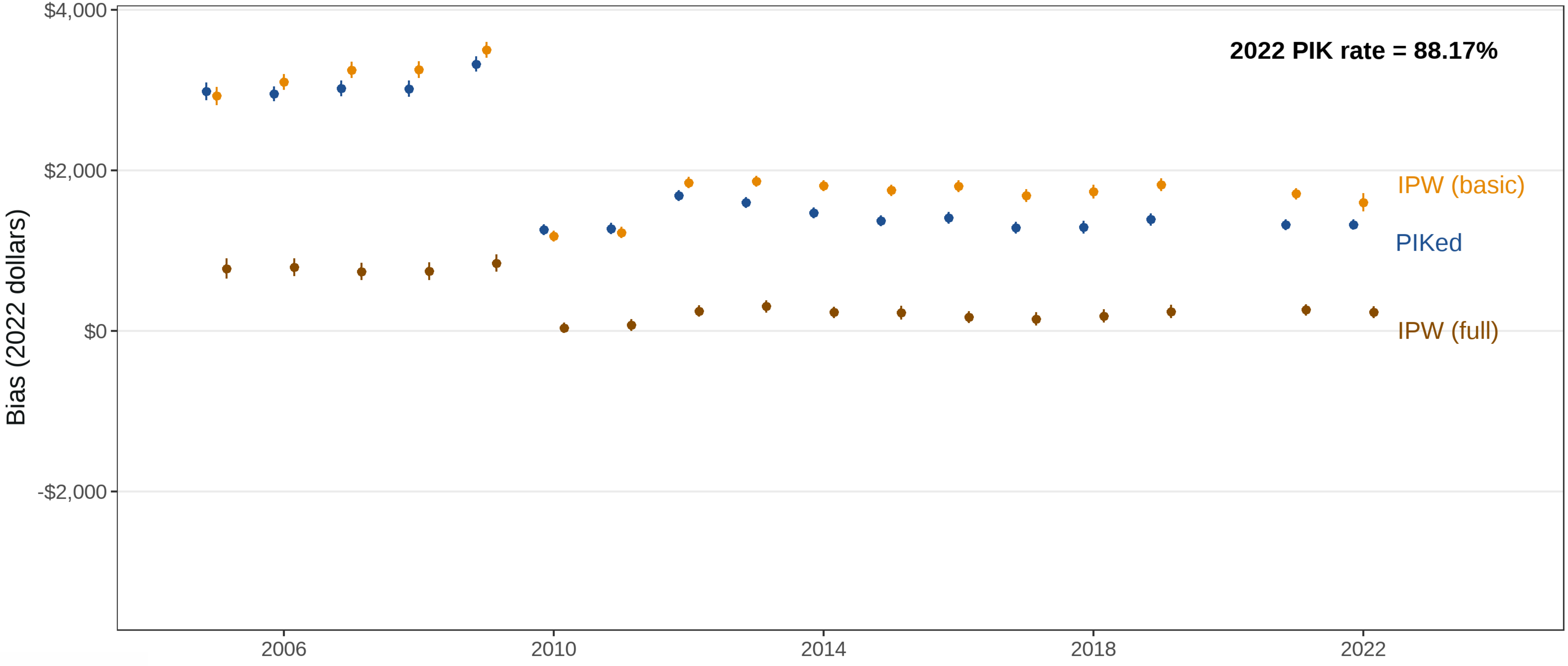
Non-Hispanic white        Hispanic of any race




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

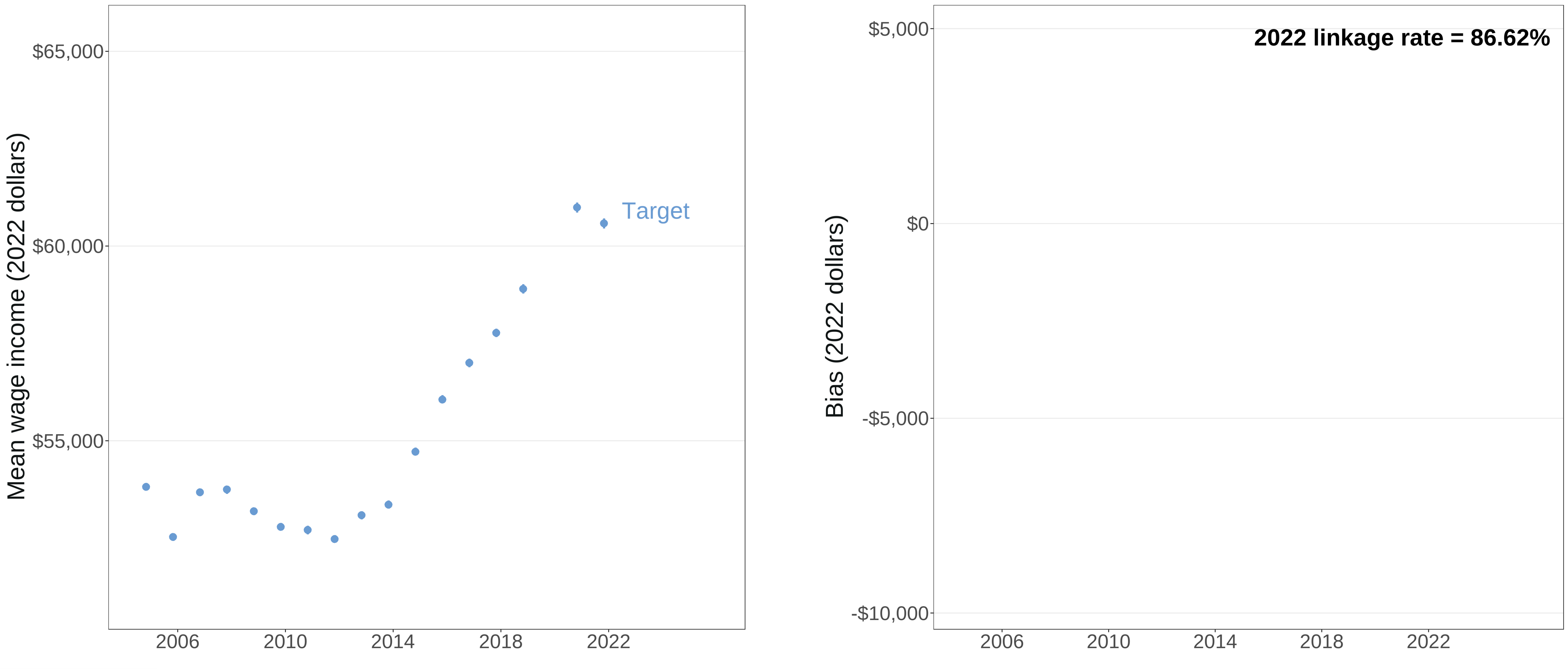
# PIK-induced bias in wage income by race/ethnicity

Non-Hispanic white     Hispanic of any race



# Linkage-induced bias in wage income

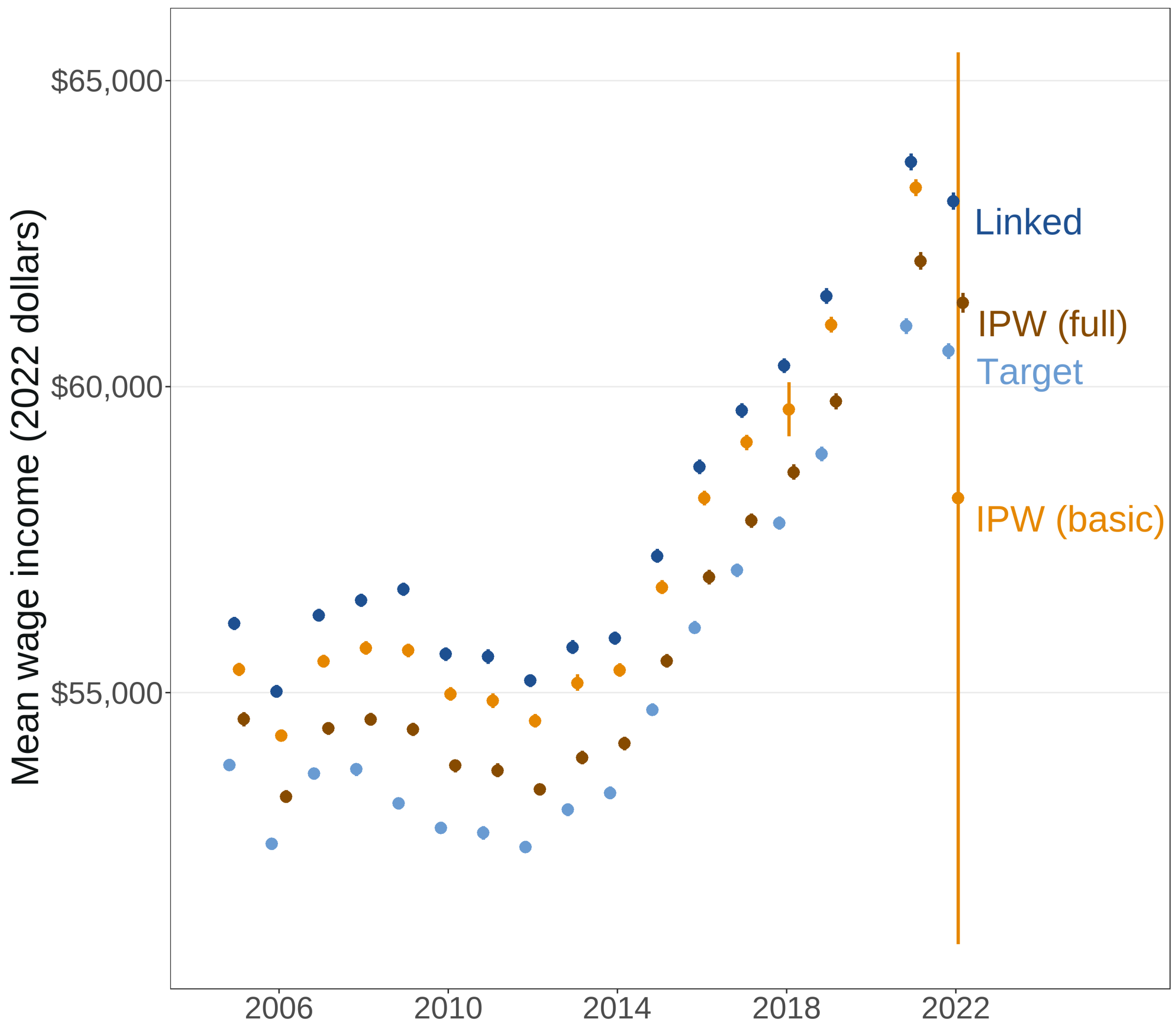
## Private-sector and government wage earners (ACS)

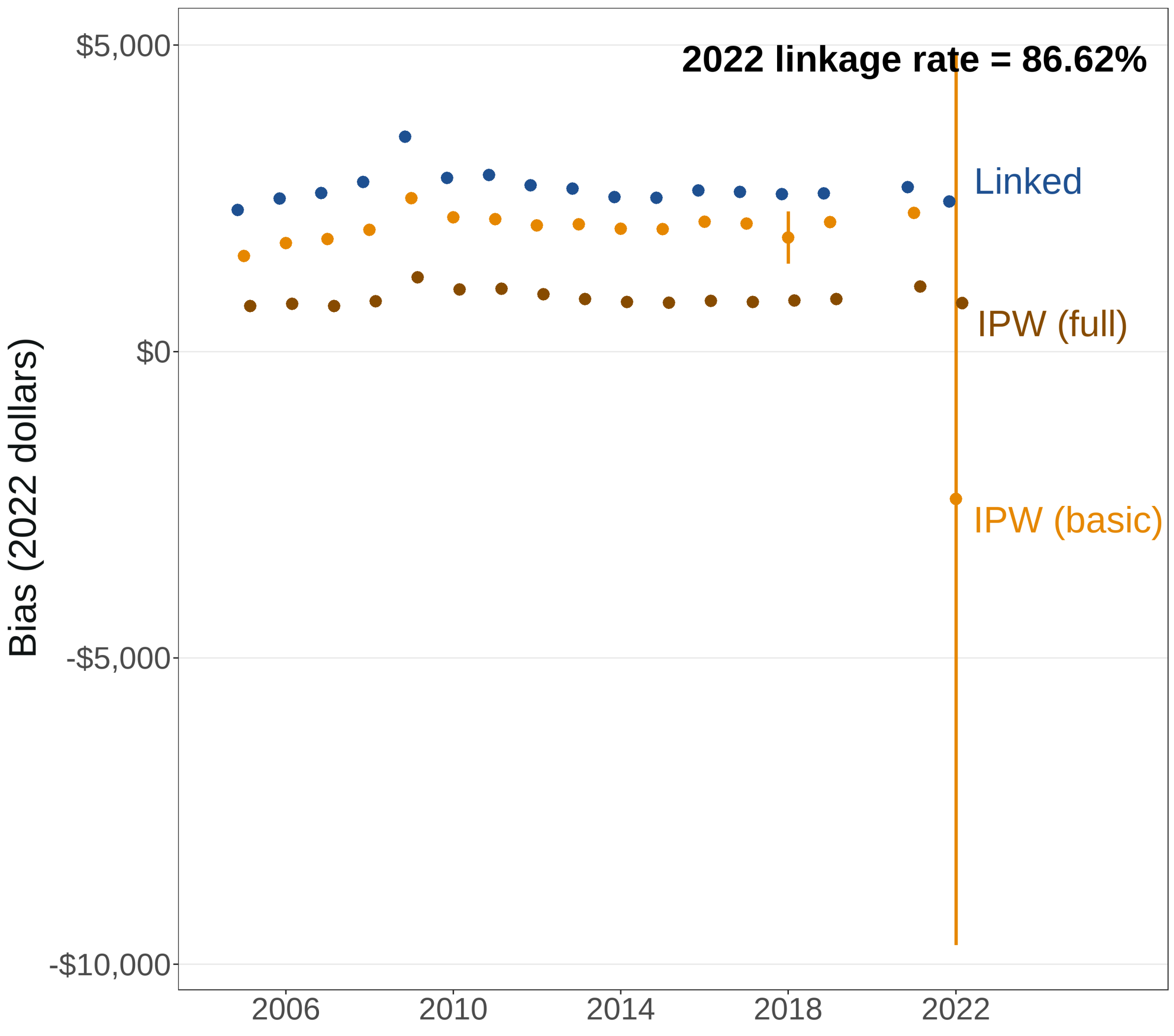



DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias in wage income

## Private-sector and government wage earners (ACS)

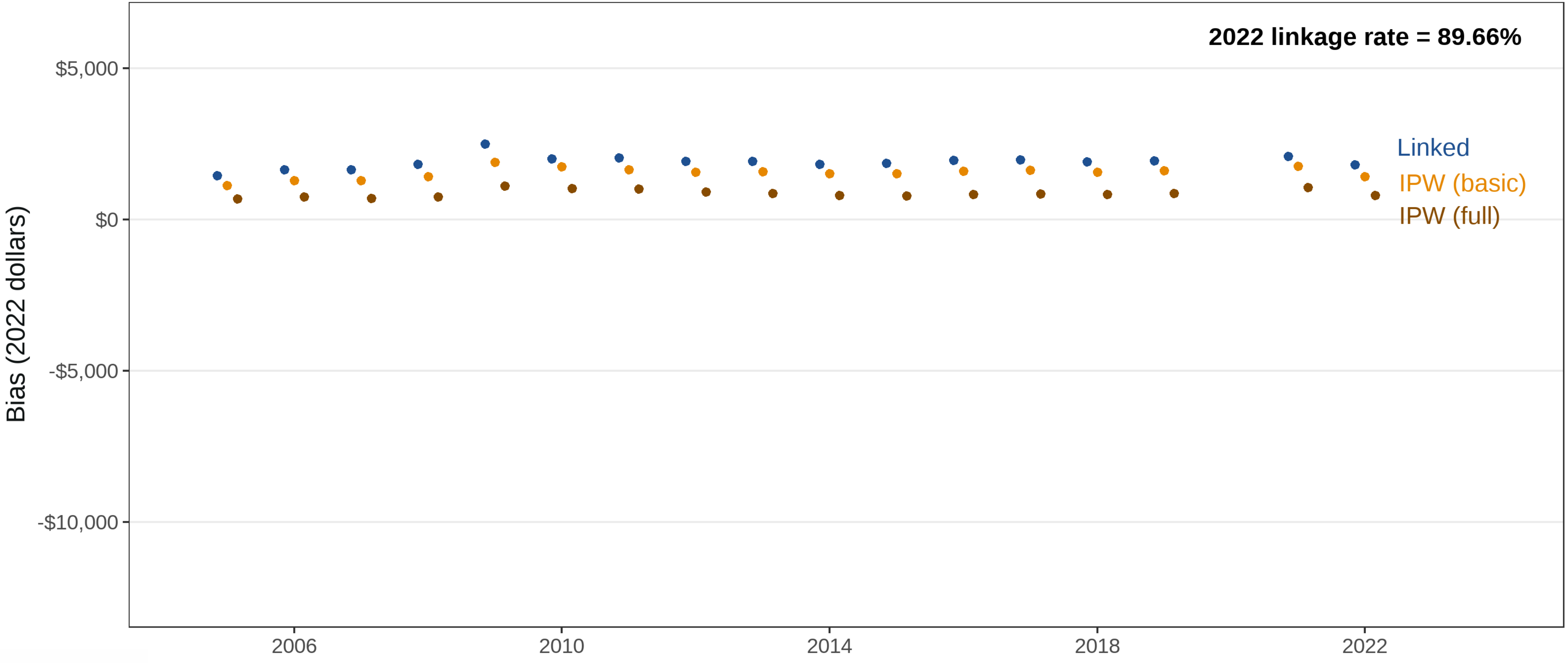



DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias in wage income by race/ethnicity

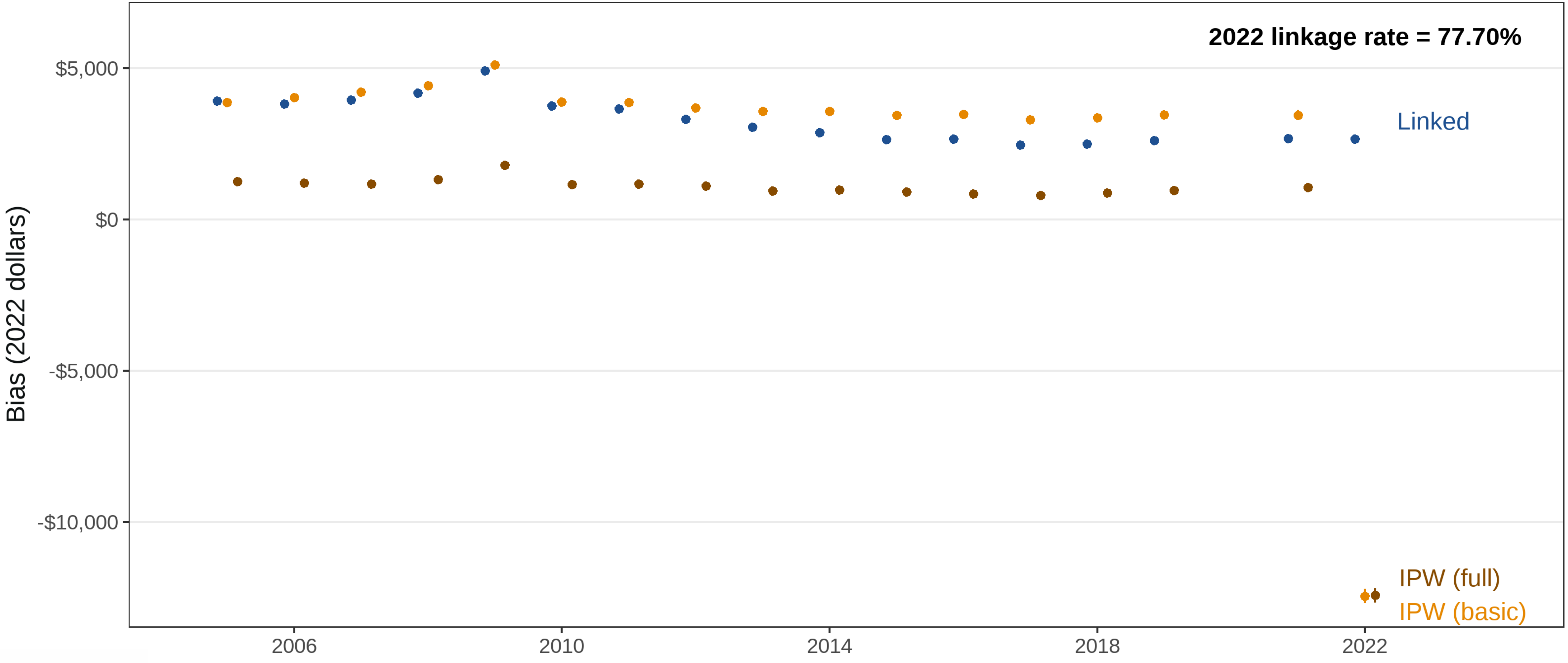
Non-Hispanic white Hispanic of any race




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias in wage income by race/ethnicity

Non-Hispanic white Hispanic of any race




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Discussion

Evidence of linkage-induced biases, even in settings with relatively high PIK rates

IPW tends to reduce linkage-induced bias, but does not necessarily eliminate it

- Underspecified models can fail to adjust for complex forms of selection into PIK assignment
    - The “basic” IPW specification accentuates linkage-induced bias for Hispanic workers
- Some evidence of instability

# Discussion

Evidence of linkage-induced biases, even in settings with relatively high PIK rates

IPW tends to reduce linkage-induced bias, but does not necessarily eliminate it

- Underspecified models can fail to adjust for complex forms of selection into PIK assignment
    - The “basic” IPW specification accentuates linkage-induced bias for Hispanic workers
- Some evidence of instability

**Next steps**

1. Incorporate additional correction techniques
    - Entropy balancing (Hainmueller, 2012; Bee et al., 2023)
    - Gradient-boosted IPW (McCaffrey, Ridgeway, and Morral, 2004; Cefalu et al., 2024)
    - Worst-case bounds for binary outcomes (Horowitz and Manski, 1995)
2. Extend analysis to the Current Population Survey (CPS)

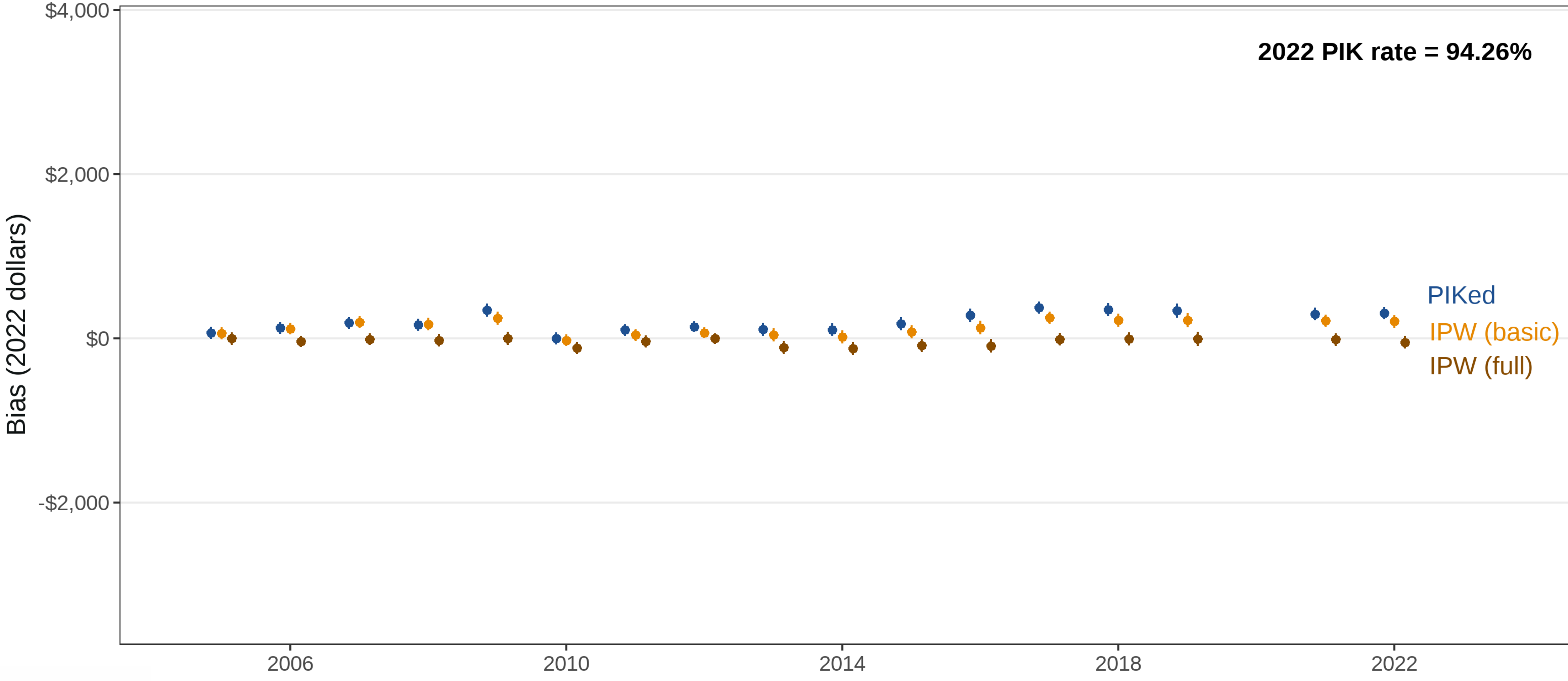
# Thank you!

Kyle Raze
*kyle.raze@census.gov*

# Appendix

# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

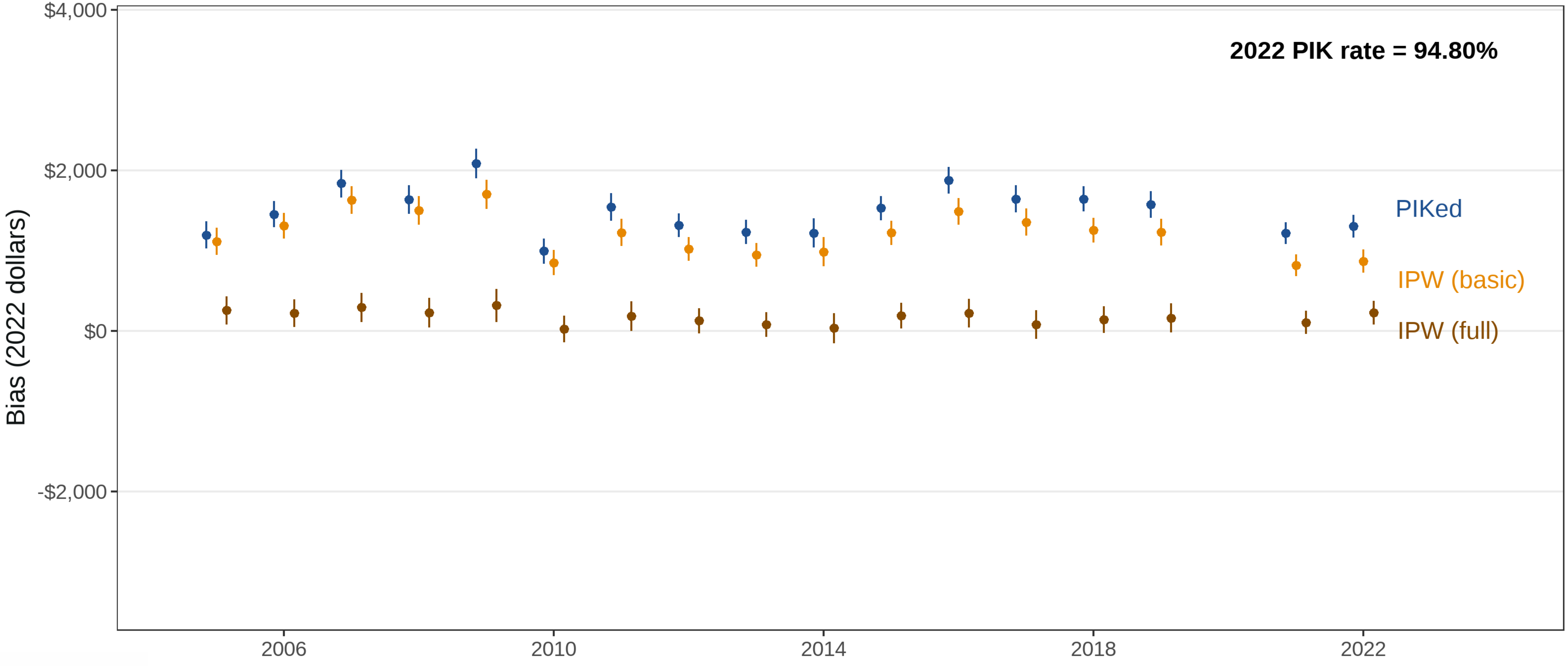
Black Asian TOMR AIAN SOR NHPI




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

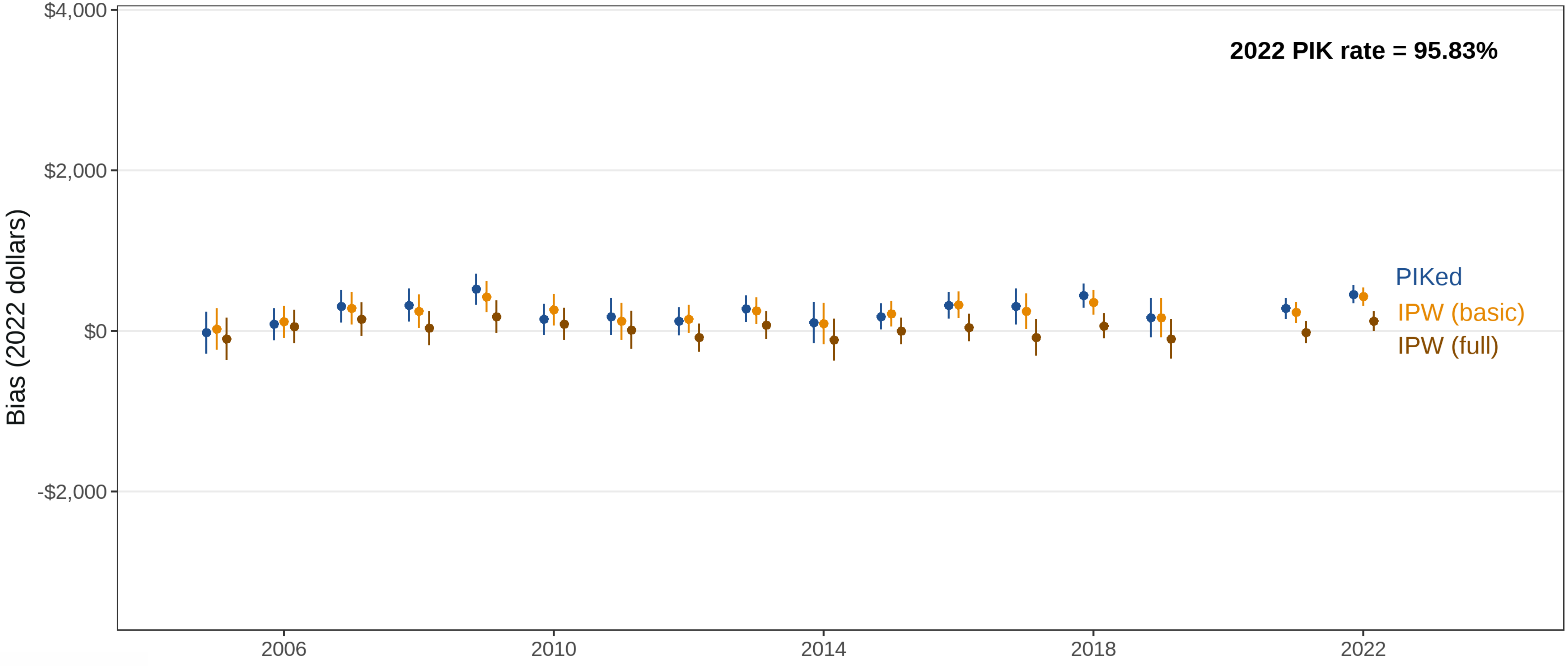
# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian TOMR AIAN SOR NHPI



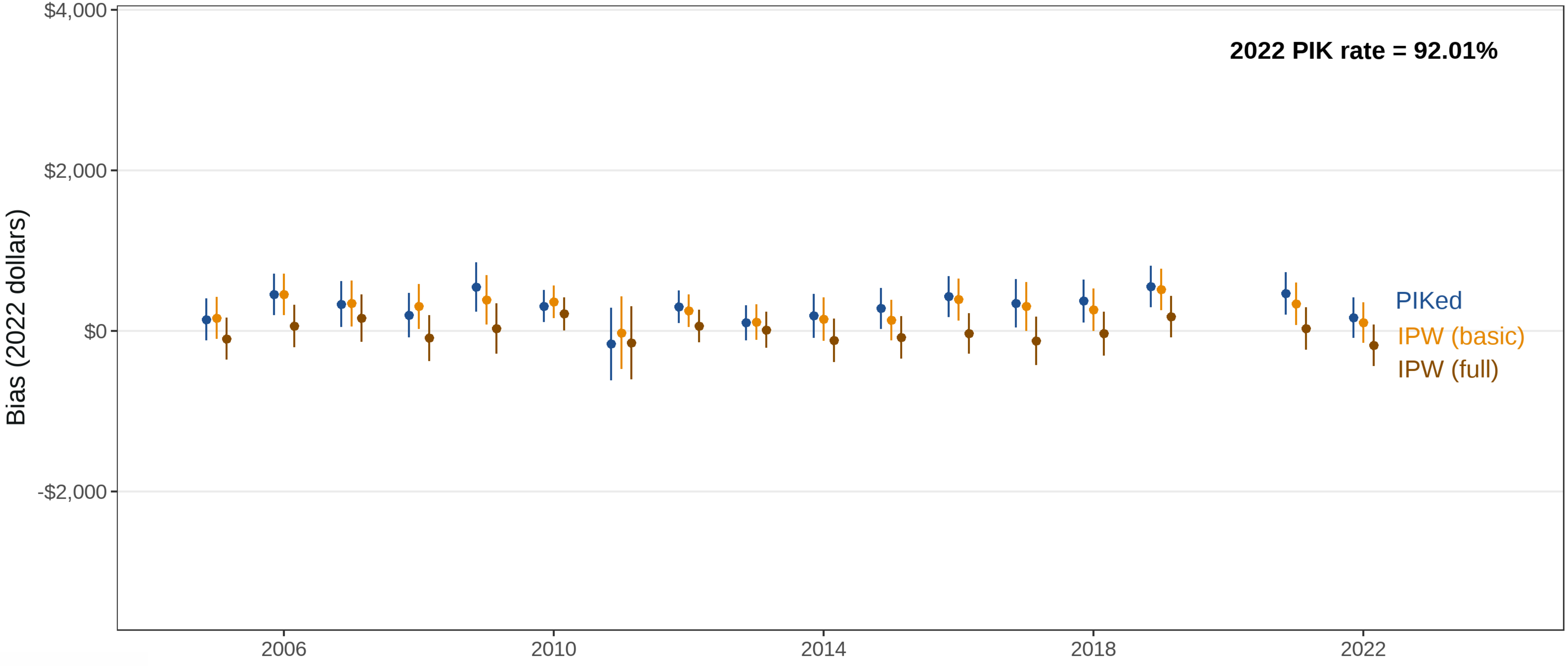
# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian TOMR AIAN SOR NHPI



# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

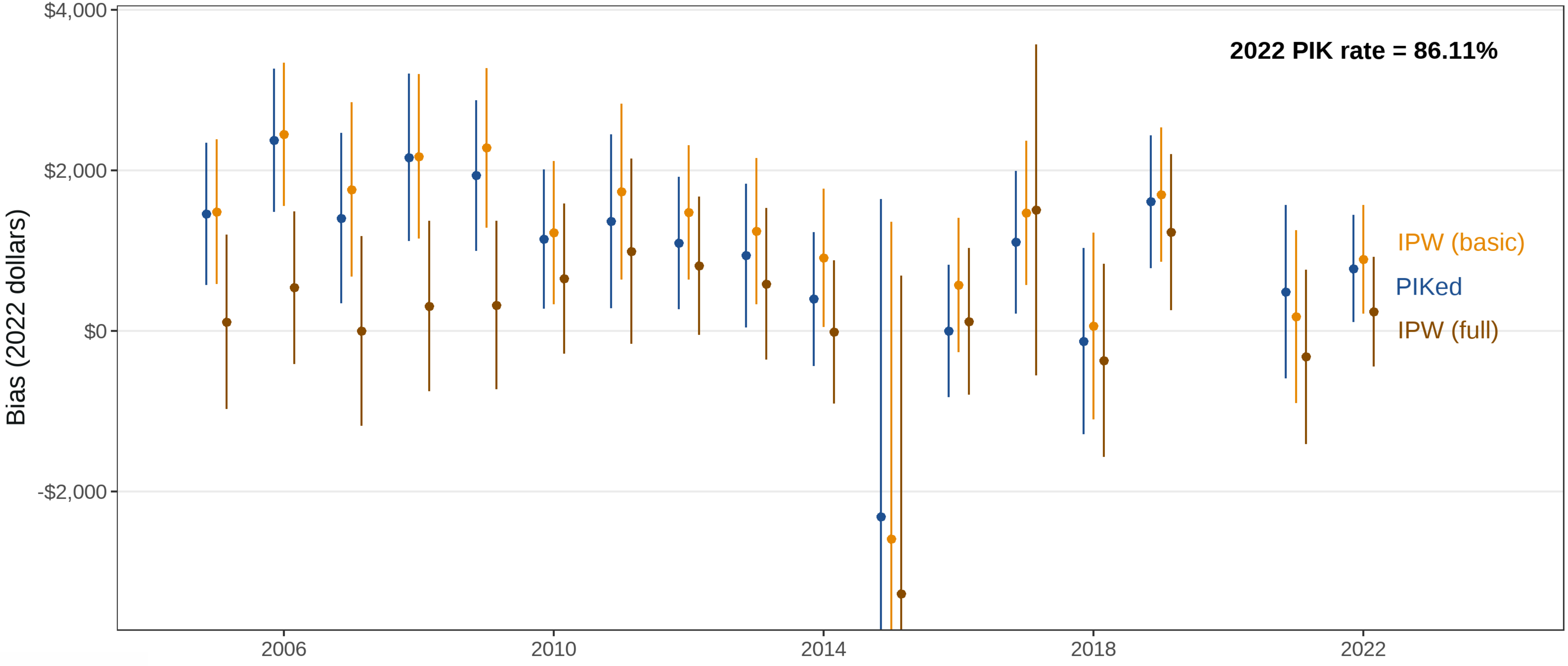
Black Asian TOMR AIAN SOR NHPI




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

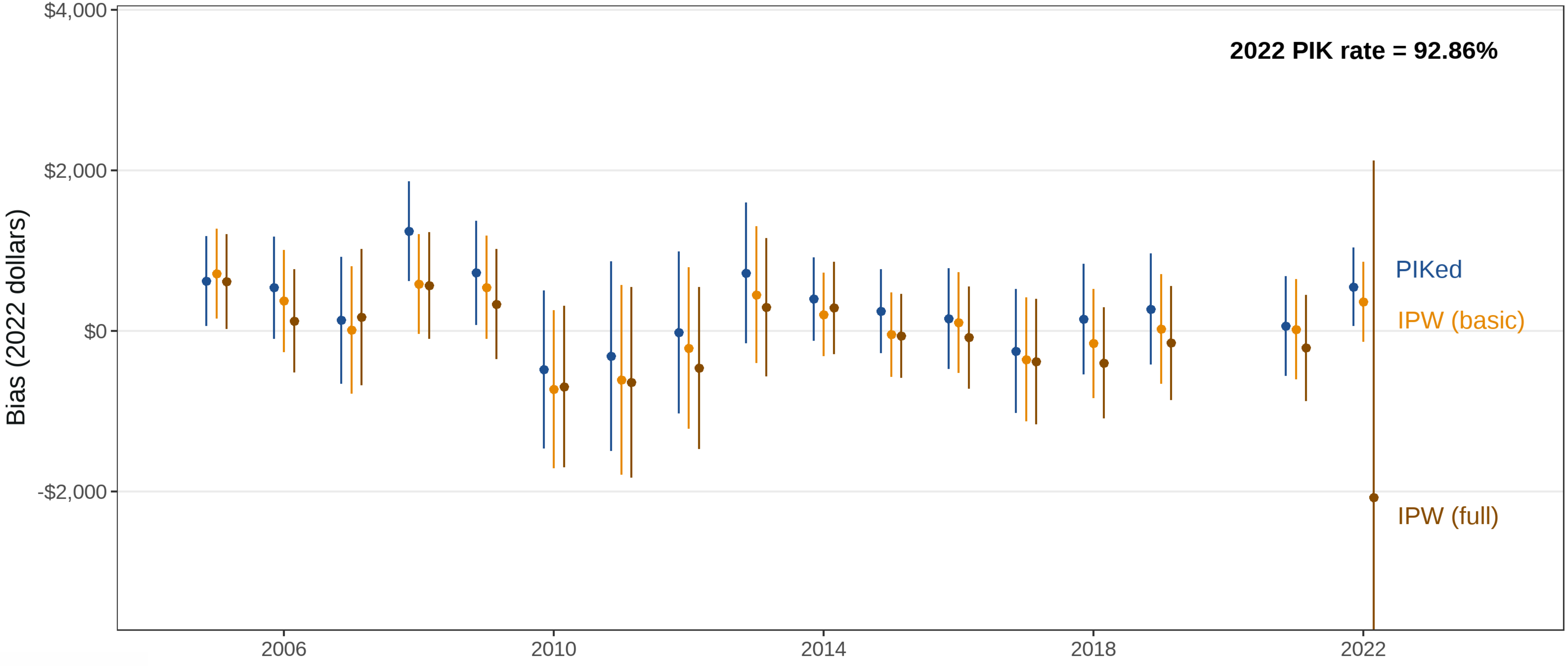
Black Asian TOMR AIAN SOR NHPI




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# PIK-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

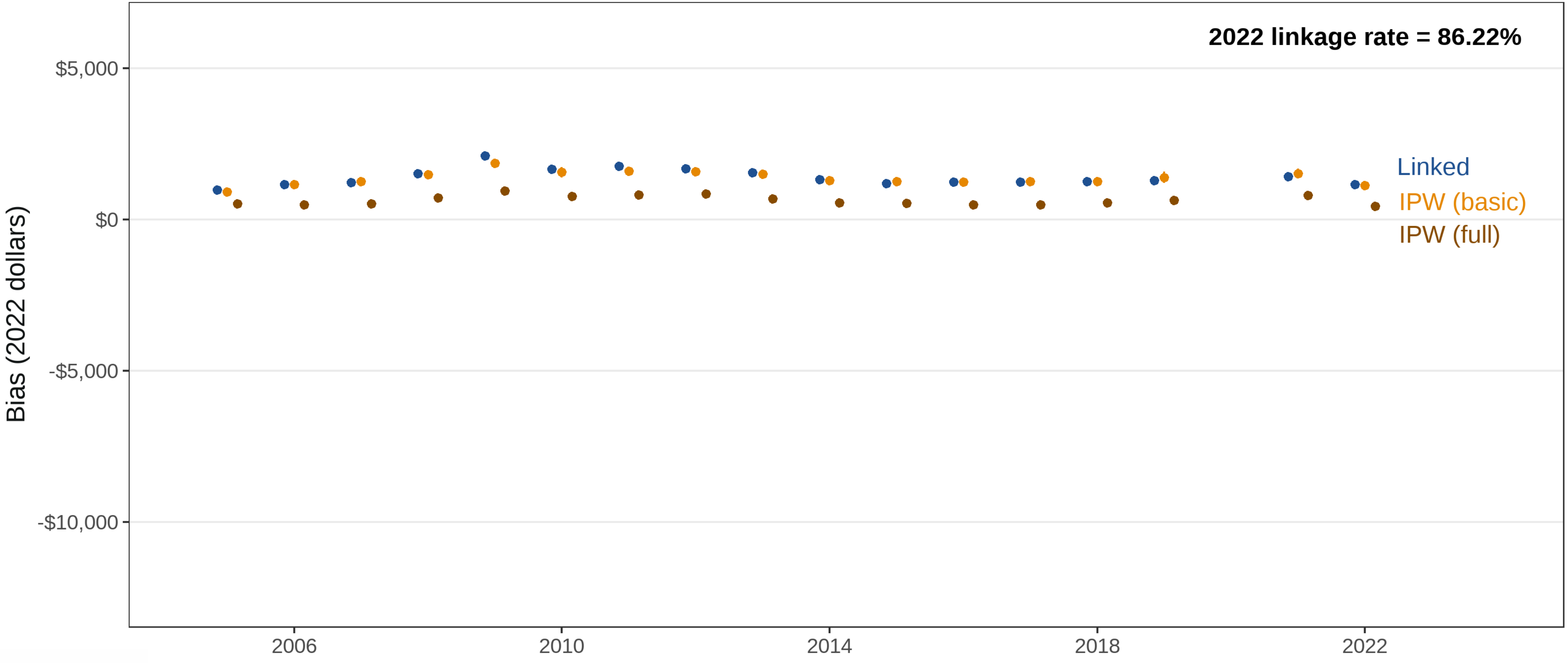
Black Asian TOMR AIAN SOR **NHPI**




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian TOMR AIAN SOR NHPI

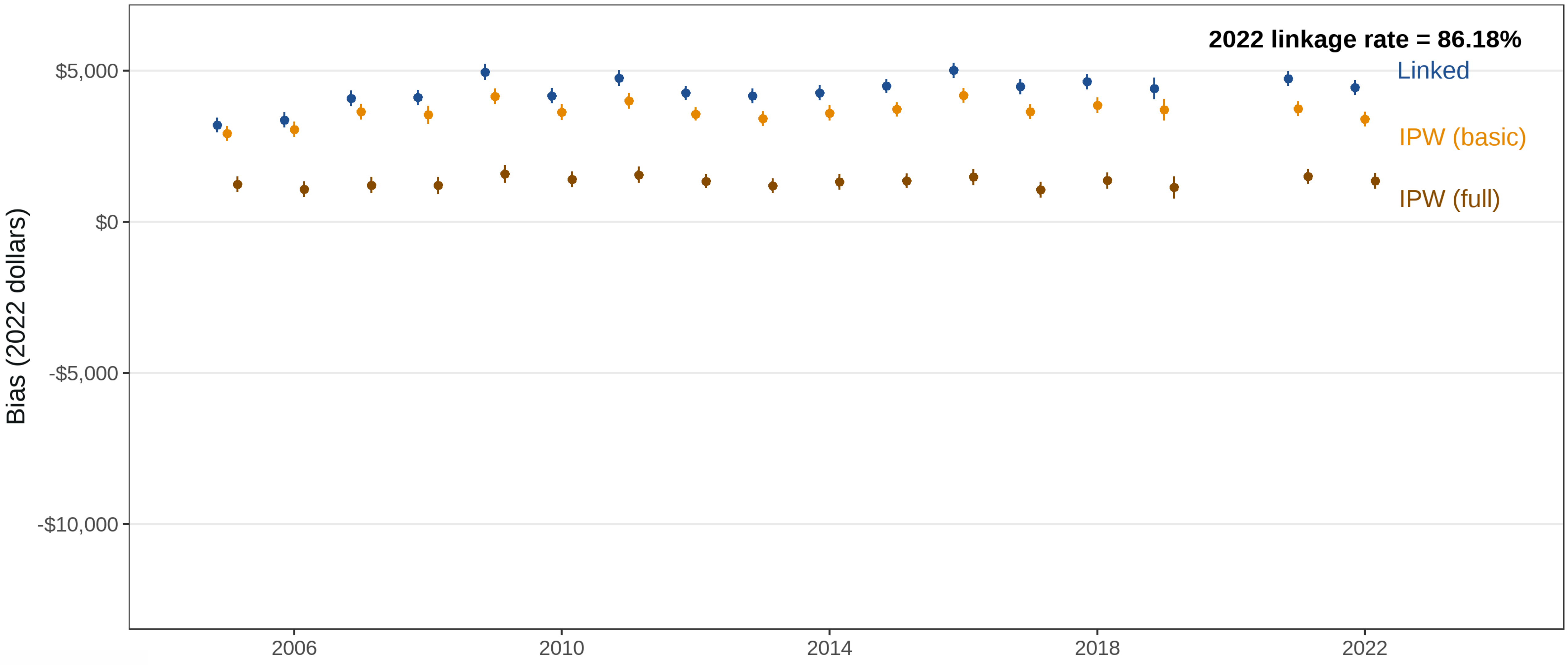



United States Census Bureau

DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

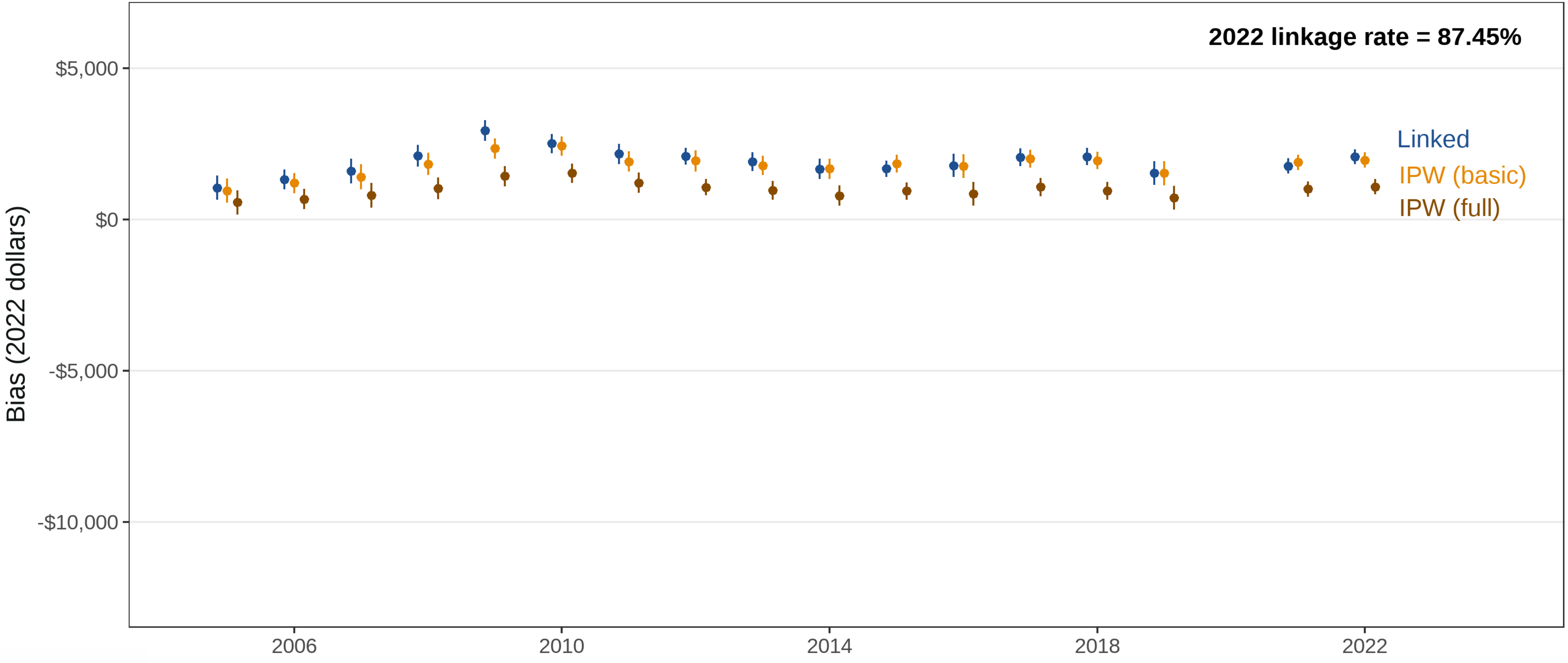
Black Asian TOMR AIAN SOR NHPI




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

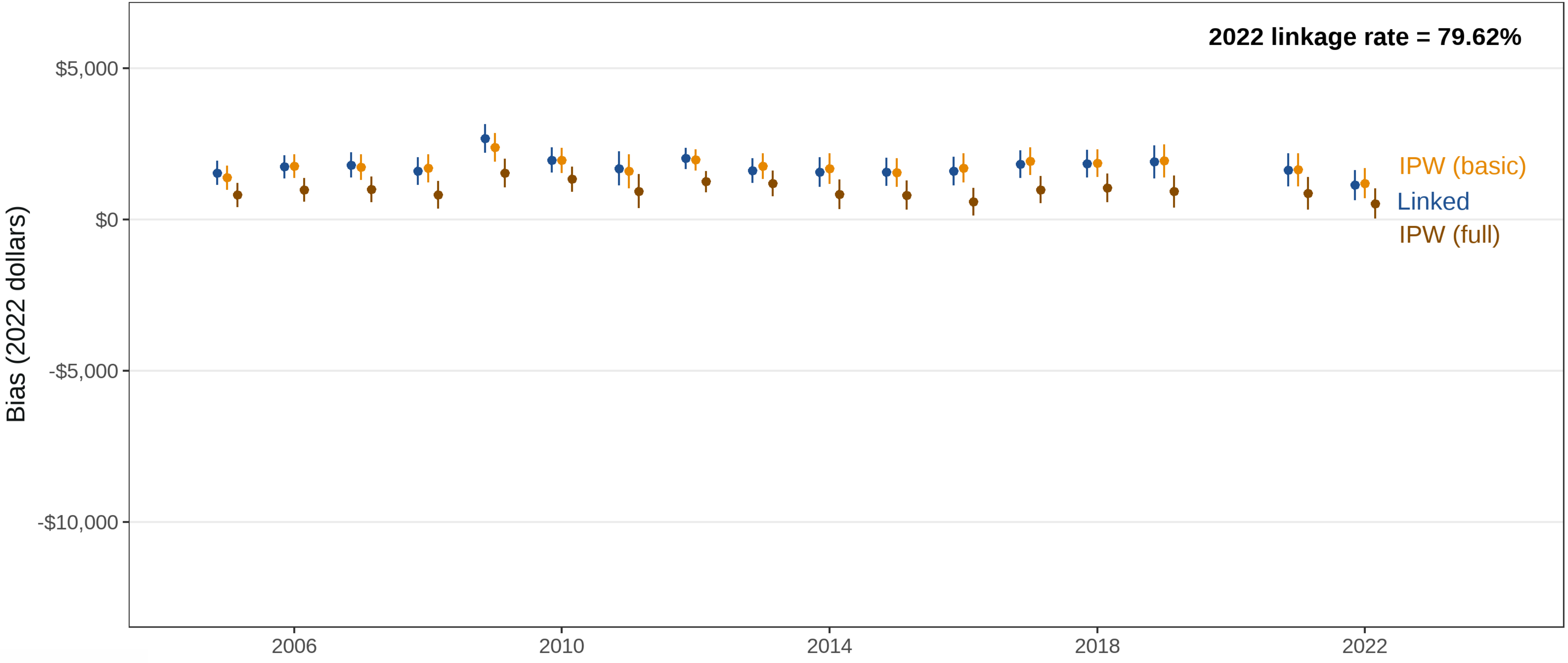
# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian **TOMR** AIAN SOR NHPI



# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

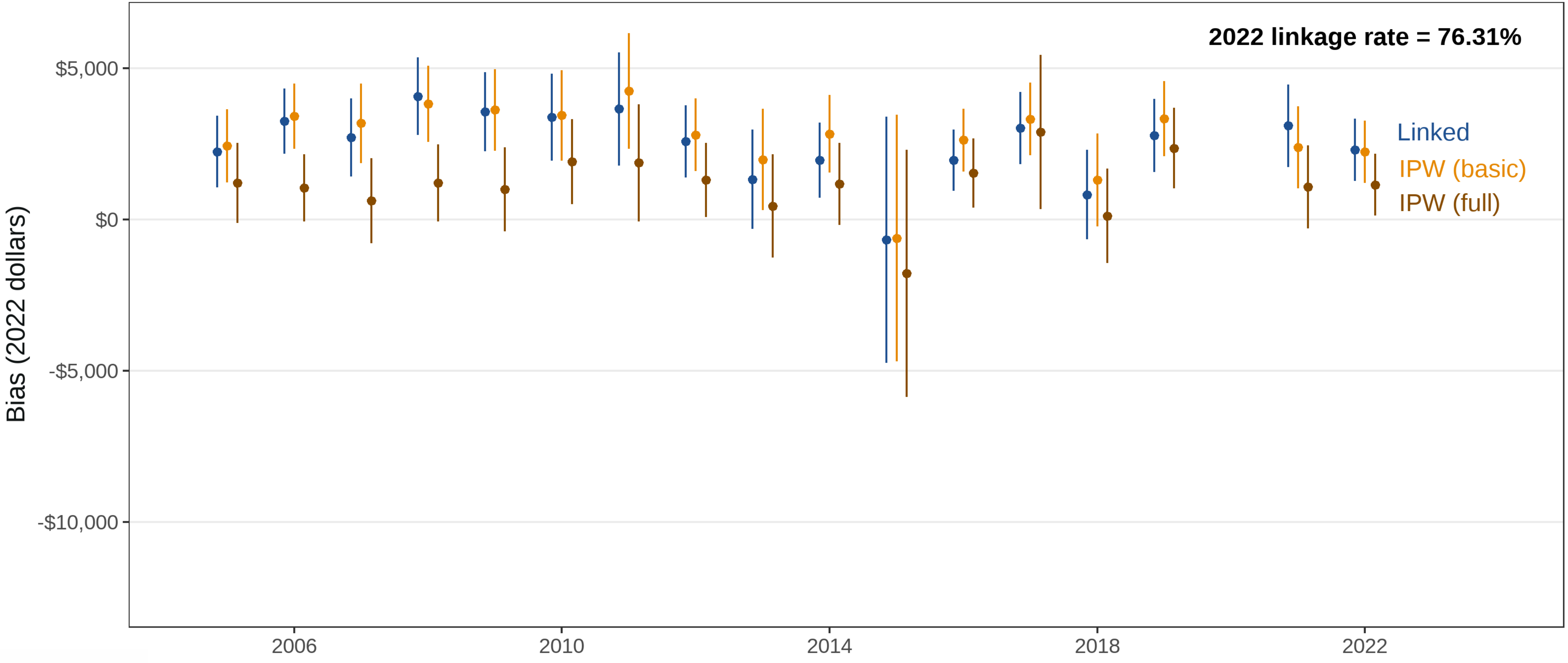
Black Asian TOMR AIAN SOR NHPI




DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006

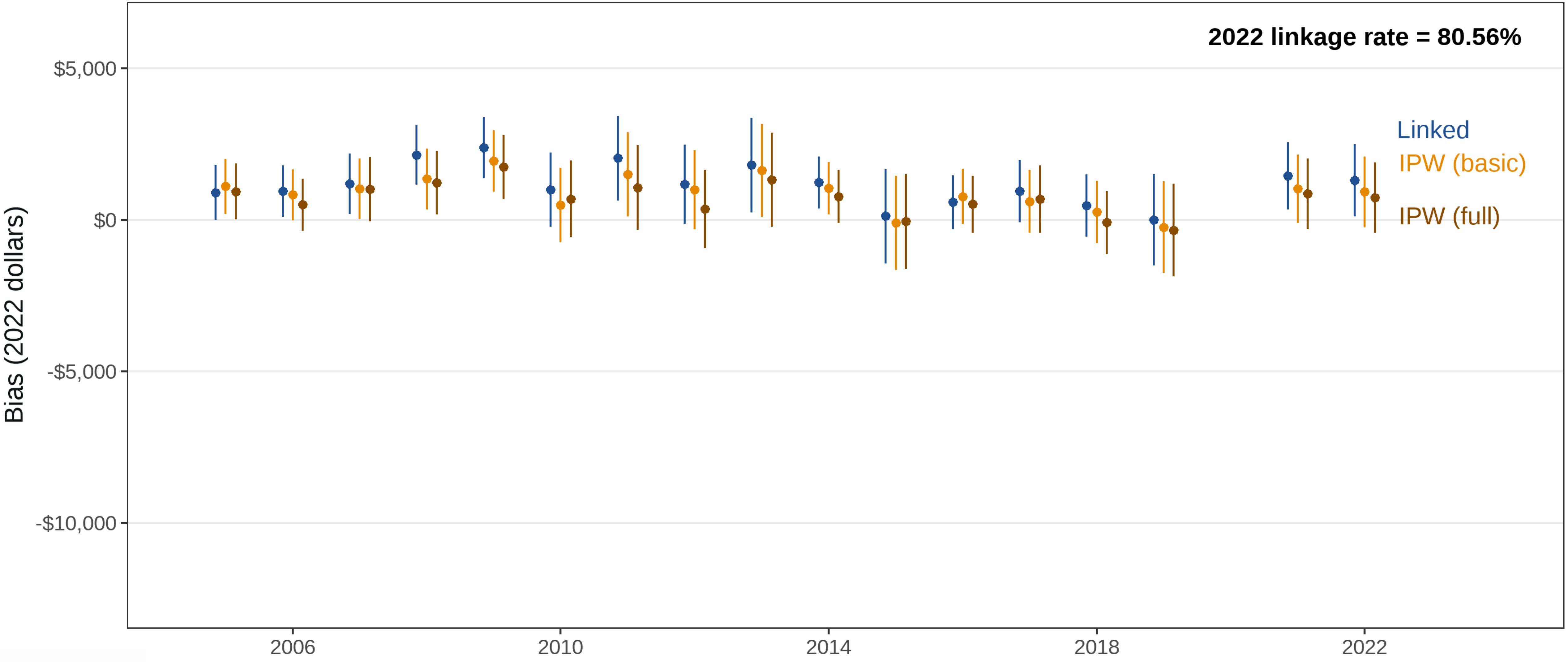
# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian TOMR AIAN SOR NHPI



# Linkage-induced bias in wage income by race/ethnicity (other non-Hispanic groups)

Black Asian TOMR AIAN SOR **NHPI**






DRB Clearance Numbers CBDRB-FY24-CES027-002 and CBDRB-FY24-CES027-006