Don’t Leave Data on the Table

A Census Simulation Integrating Administrative and Survey-Collected Data

David Brown and Marta Murray-Close
U.S. Census Bureau

October 22, 2024

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 reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release. DRB Clearance Number: CBDRB-FY24-0484

2020 Census

- Address-based data collection
  - Began with list of potentially inhabitable addresses
  - Determined if each address was occupied, vacant, or uninhabitable
  - Sought survey-style response about people at each occupied address
- Used survey-style data for an address if available
- Used administrative records (AR) when could not obtain a survey-style response and AR quality was judged to be sufficiently high

2020 Census AR Enumeration

- 4.59% of housing structures resolved through AR
    - 3.20% occupied
    - 1.15% vacant
    - 0.24% delete

- Some people added may have been counted elsewhere
    - To link a person across records, assign a Protected Identification Keys (PIK)
    - Find higher-than-average duplication rate
        - 9.0% of AR Enumeration PIKs
        - 2.1% of all 2020 Census PIKs

2020 Enhanced Demographic Frame (EDF)

- Person-based data collection
  - Created list of verified people in more than 20 AR data sources
  - Verified = Assigned a PIK

- Used predictive model to select person’s most likely reference-date address from their set of AR addresses

- Integrated AR and previously collected survey-style data to obtain demographic information for each person

- Excluded people not alive on reference date and people living outside the U.S.

Data left on the table

2020 Census

- AR for addresses on Census address list
  - People living at an address
  - Demographics of residents
- AR for addresses not on Census address list
- AR for addresses that could not be assigned an address identifier

2020 EDF

- 2020 Census data for people in recent AR
  - Reference-date address
  - Current demographics
- 2020 Census data for verified people not in recent AR
- 2020 Census data for unverified people

United States® Census Bureau

This study: Integrating 2020 Census data with the 2020 EDF

2020 Census

- AR for addresses on Census address list
  - People living at an address
  - Demographics of residents
- AR for addresses not on Census address list
- AR for addresses that could not be assigned an address identifier

2020 EDF

- 2020 Census data for people in recent AR
  - Reference-date address
  - Current demographics
- 2020 Census data for verified people not in recent AR
- 2020 Census data for unverified people

Simulation methodology

- Begin with EDF people who were alive and were U.S. residents on Census Day
  - By construction, all are verified

- Add verified 2020 Census people who were alive on Census Day
  - Unduplicate by PIK, prioritizing record from earliest 2020 Census response

Simulation methodology

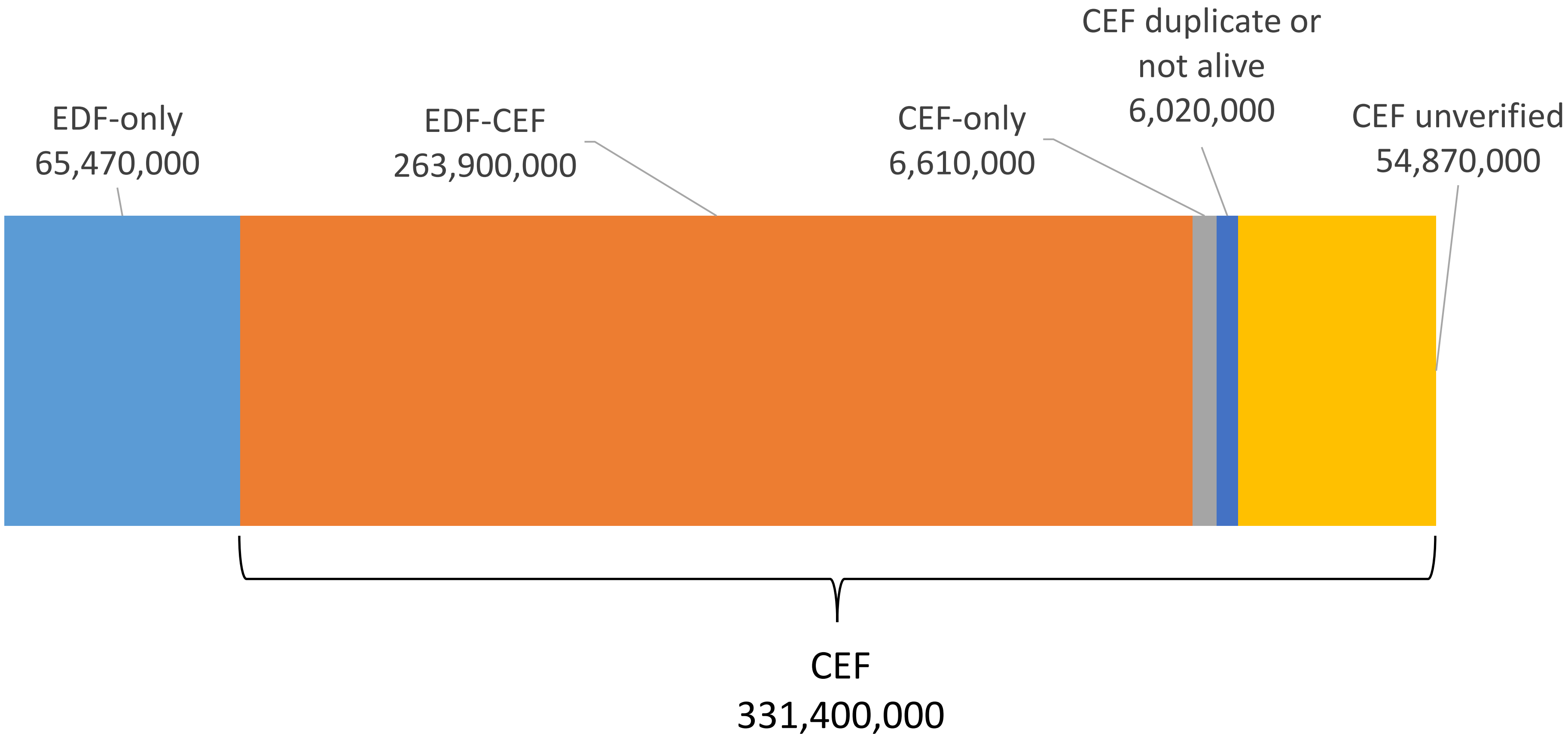
- Prioritize address information as follows:
  1. 2020 Census address
  2. EDF address
  - In the future, could select address closest to Census Day

- Prioritize demographic information as follows:
  1. As-reported 2020 Census value
  2. As-reported EDF value
  3. Edited or imputed 2020 Census value
  4. Imputed EDF value

Simulation methodology

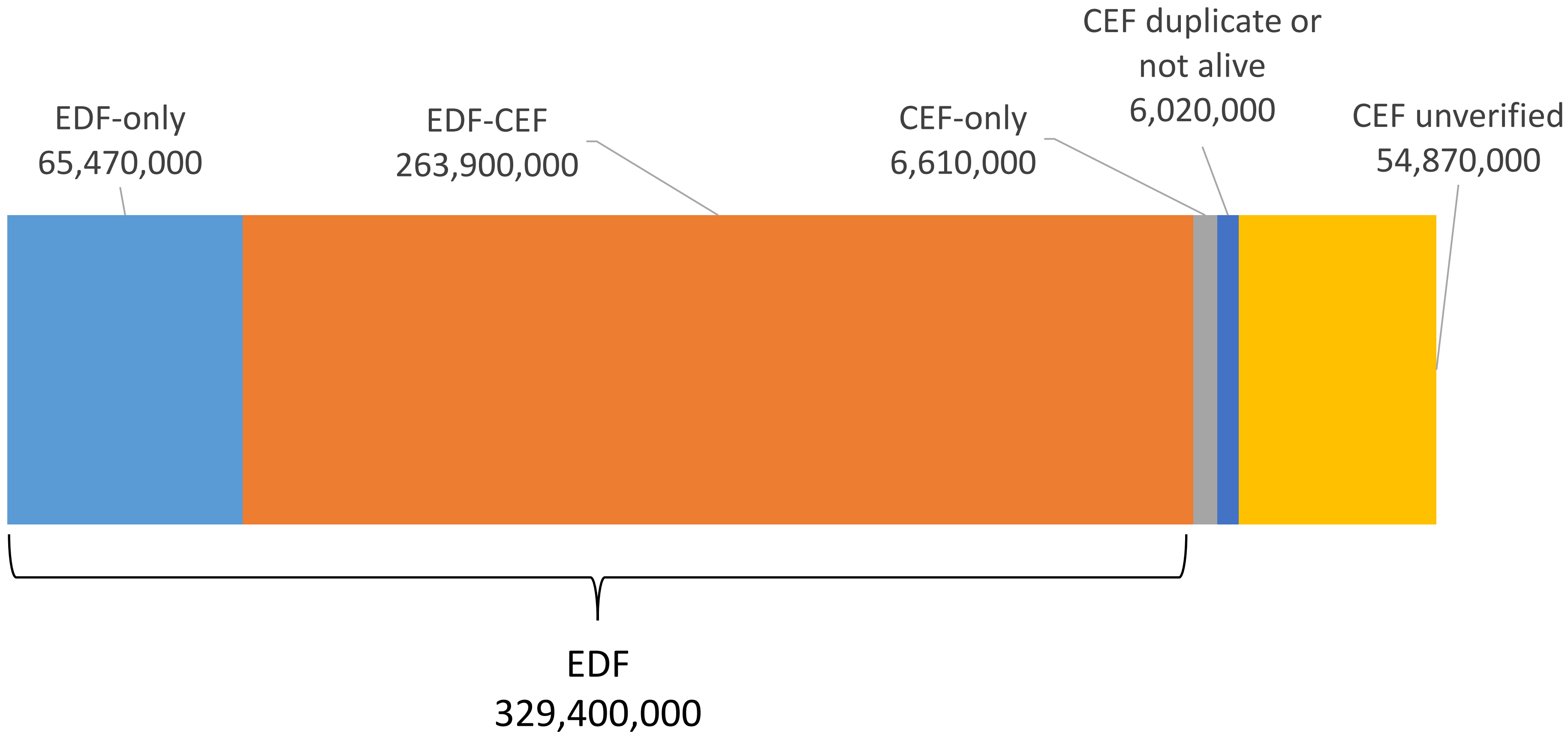
- Conduct simulation for entire 2020 Census and separately for each 2020 Census response mode
  - Self-response
  - Non-Response Follow-Up (NRFU) interview, visit 1
  - NRFU interview visits 2-6
  - NRFU proxy
  - Group quarters

People in EDF and 2020 Census Edited File (CEF)

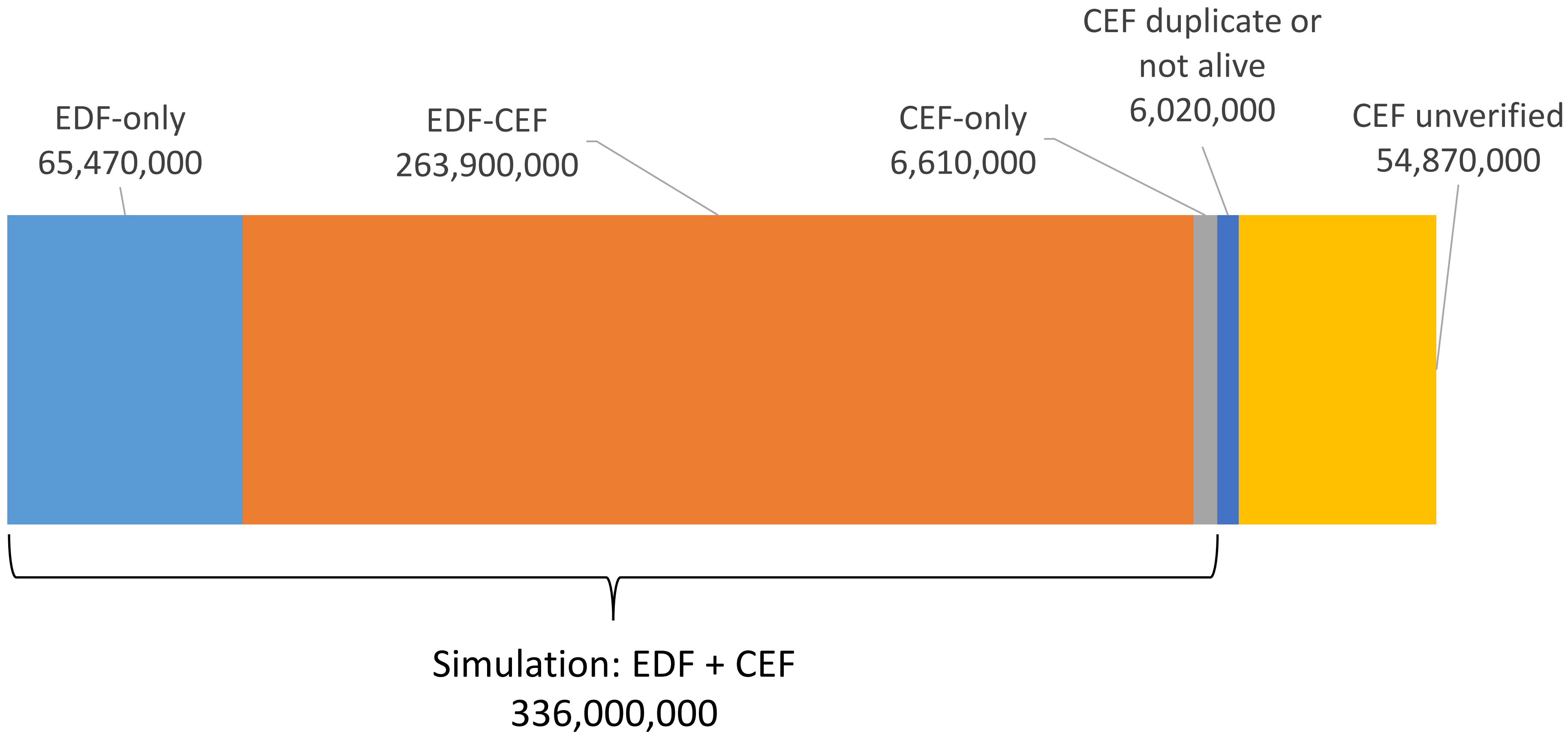


DRB Clearance Number: CBDRB-FY24-0484

People in EDF and 2020 Census Edited File (CEF)

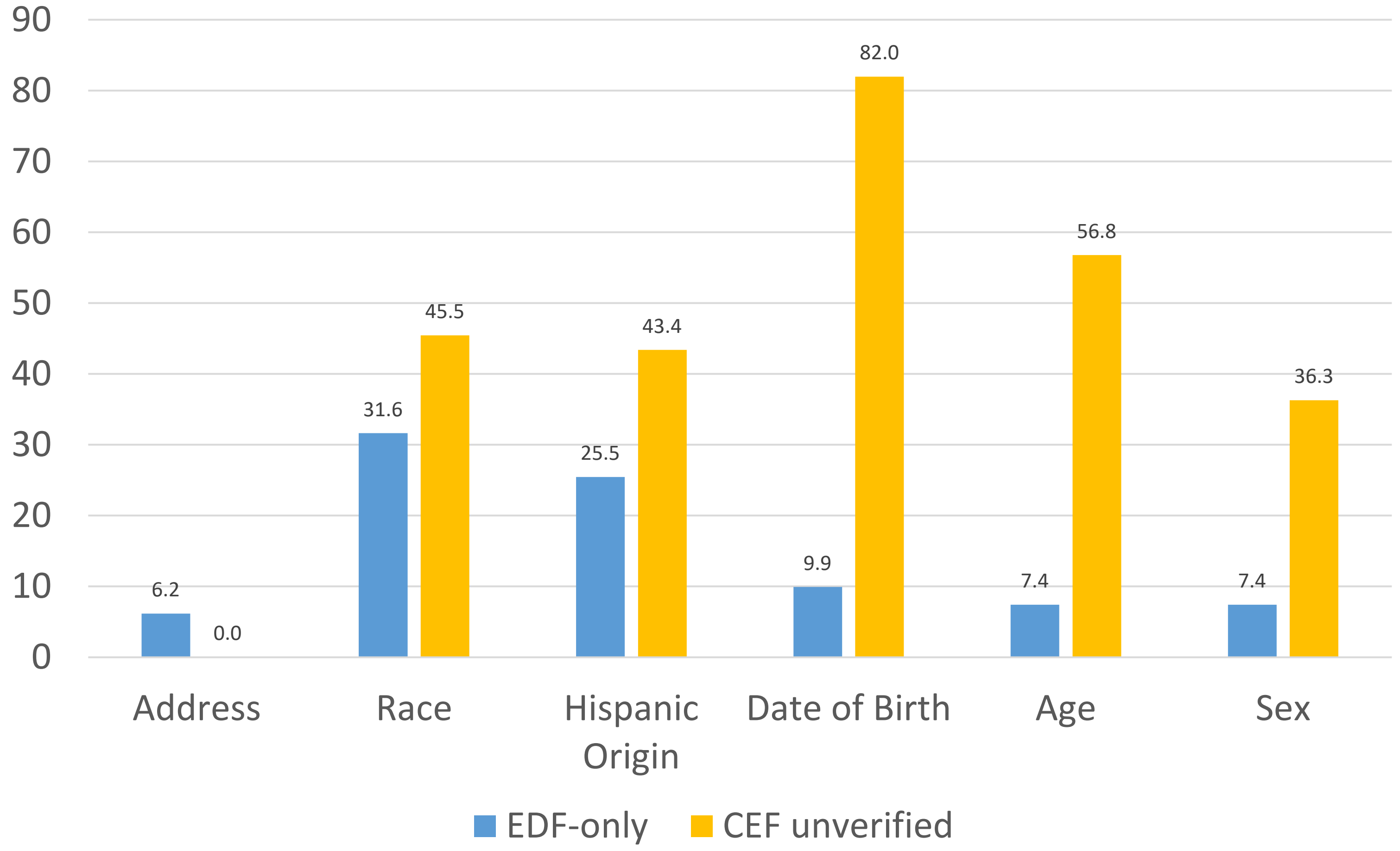


People in EDF and 2020 Census Edited File (CEF)



Percent of people missing characteristics

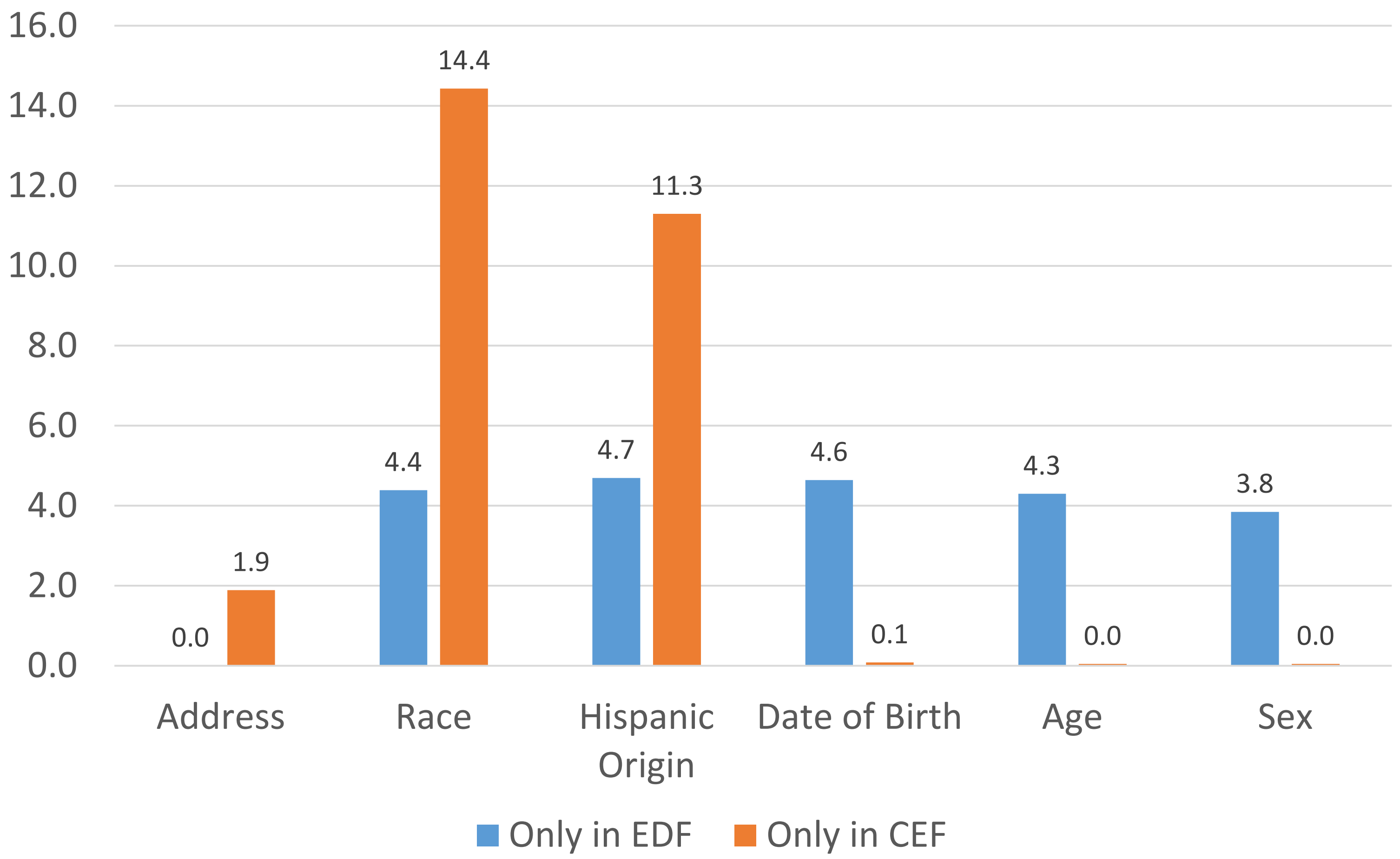
EDF-only versus CEF unverified



DRB Clearance Number: CBDRB-FY24-0484

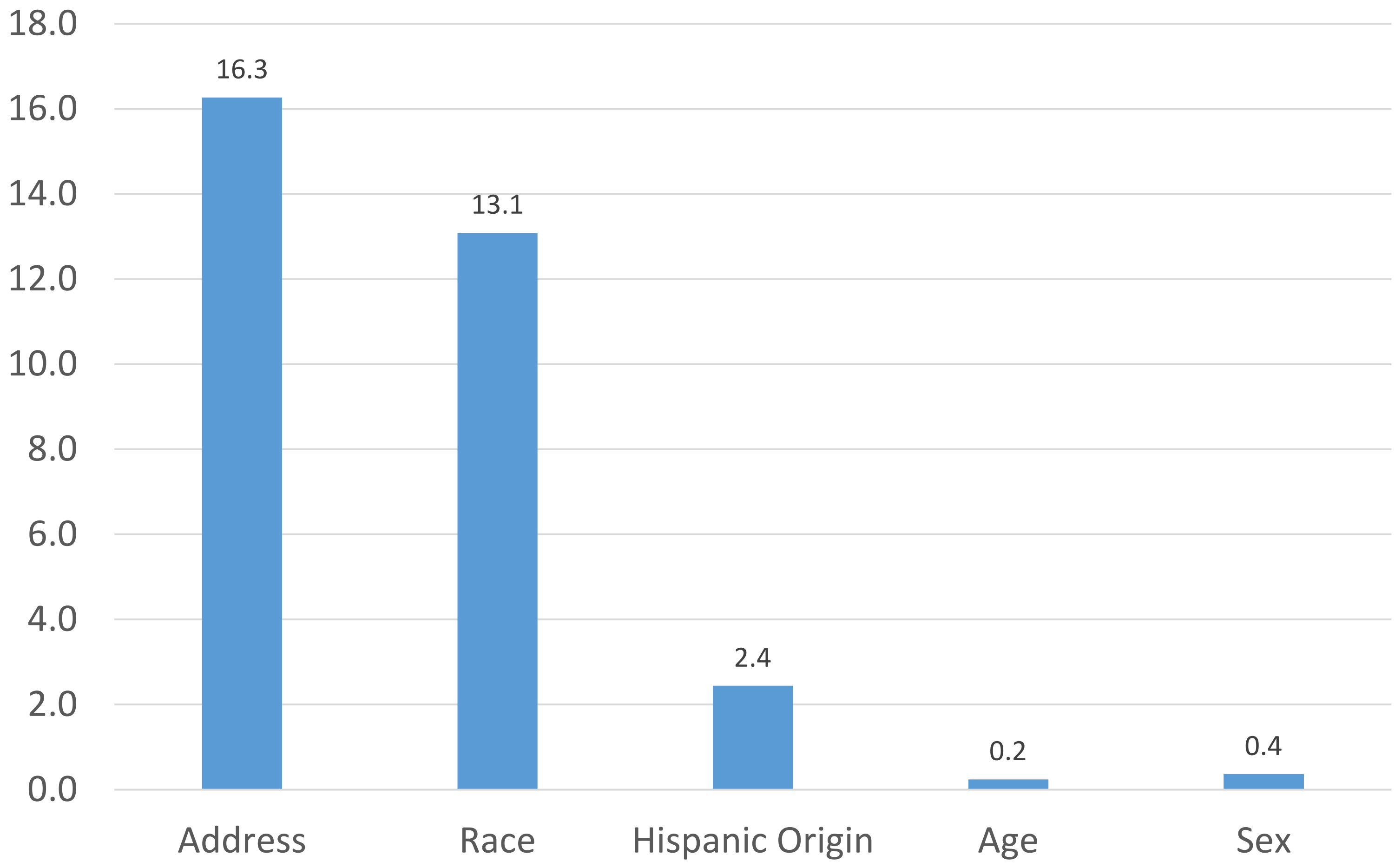
Contributions of EDF and CEF to completeness

Among people in both sources, percent with characteristic only in given source



Disagreement rates between CEF and EDF

Among people with characteristic in both sources, percent with disagreement

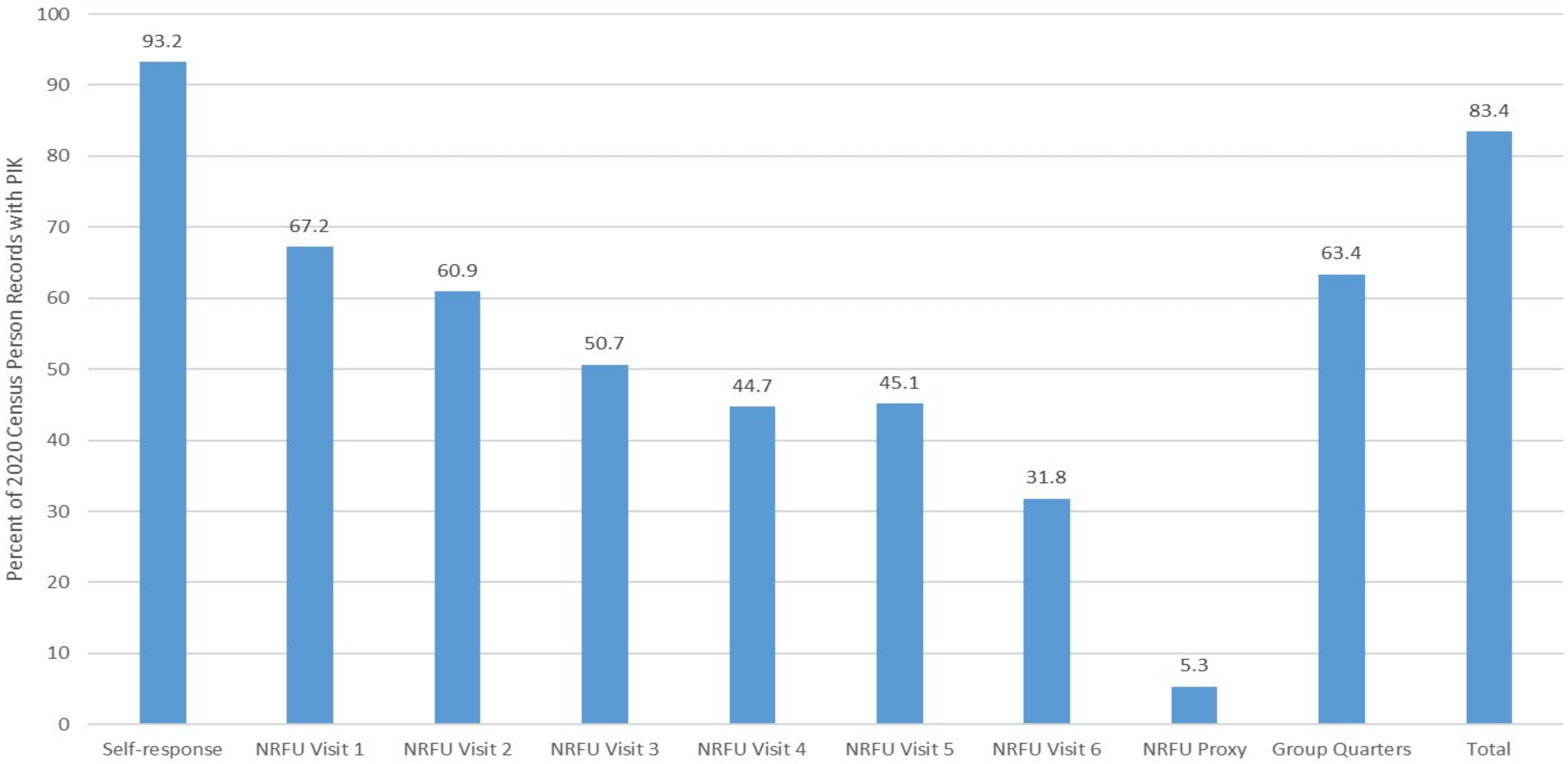


DRB Clearance Number: CBDRB-FY24-0484

Age Heaping in Population Age 23-62

Source	Whipple’s index
2020 Census (DHC)	105.9
EDF	100.6
EDF + CEF	100.6
2020 Census without PIK	130.3
Unlinked AR	100.7

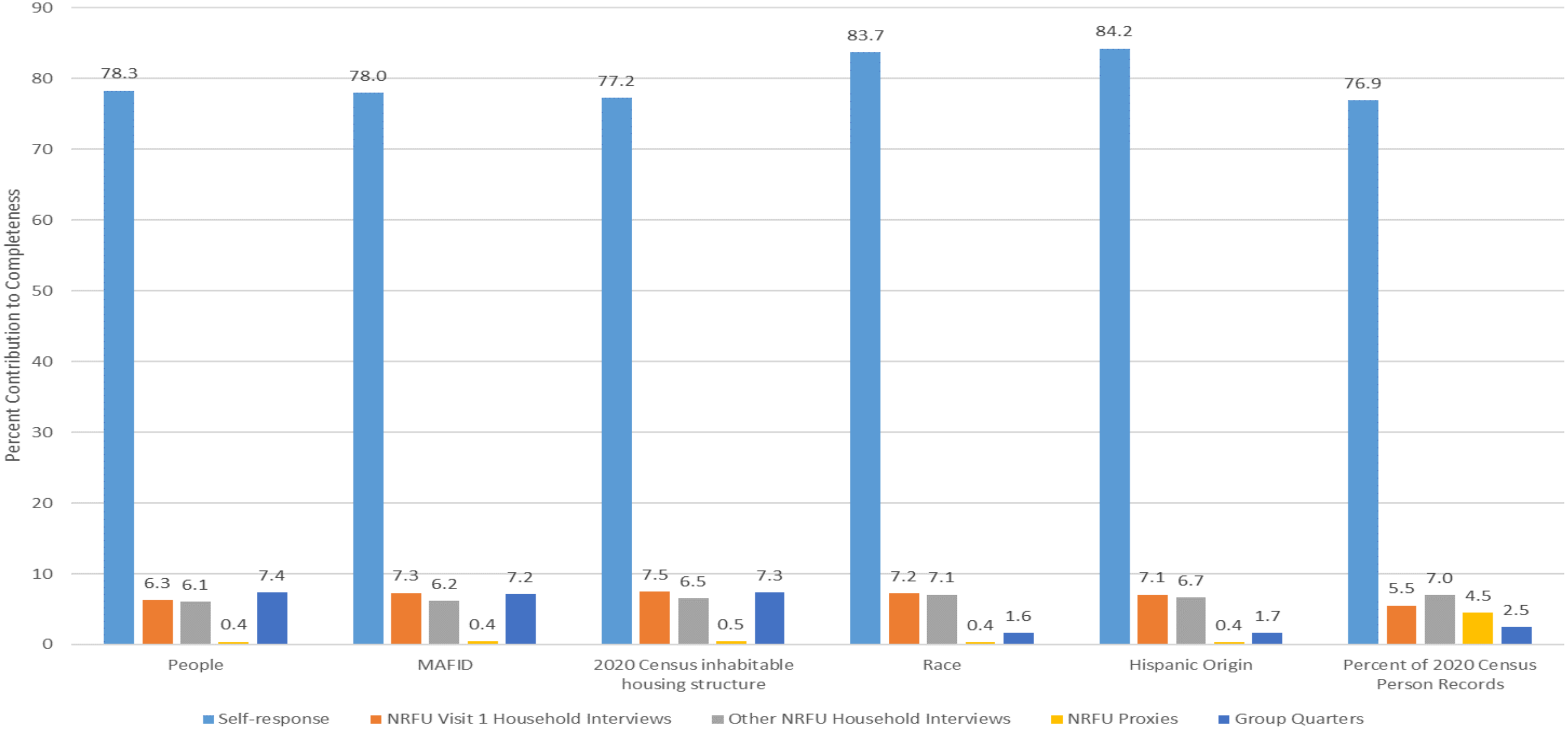
Percent of 2020 Census Person Records with PIK by Operation



DRB Clearance Number: CBDRB-FY24-0484

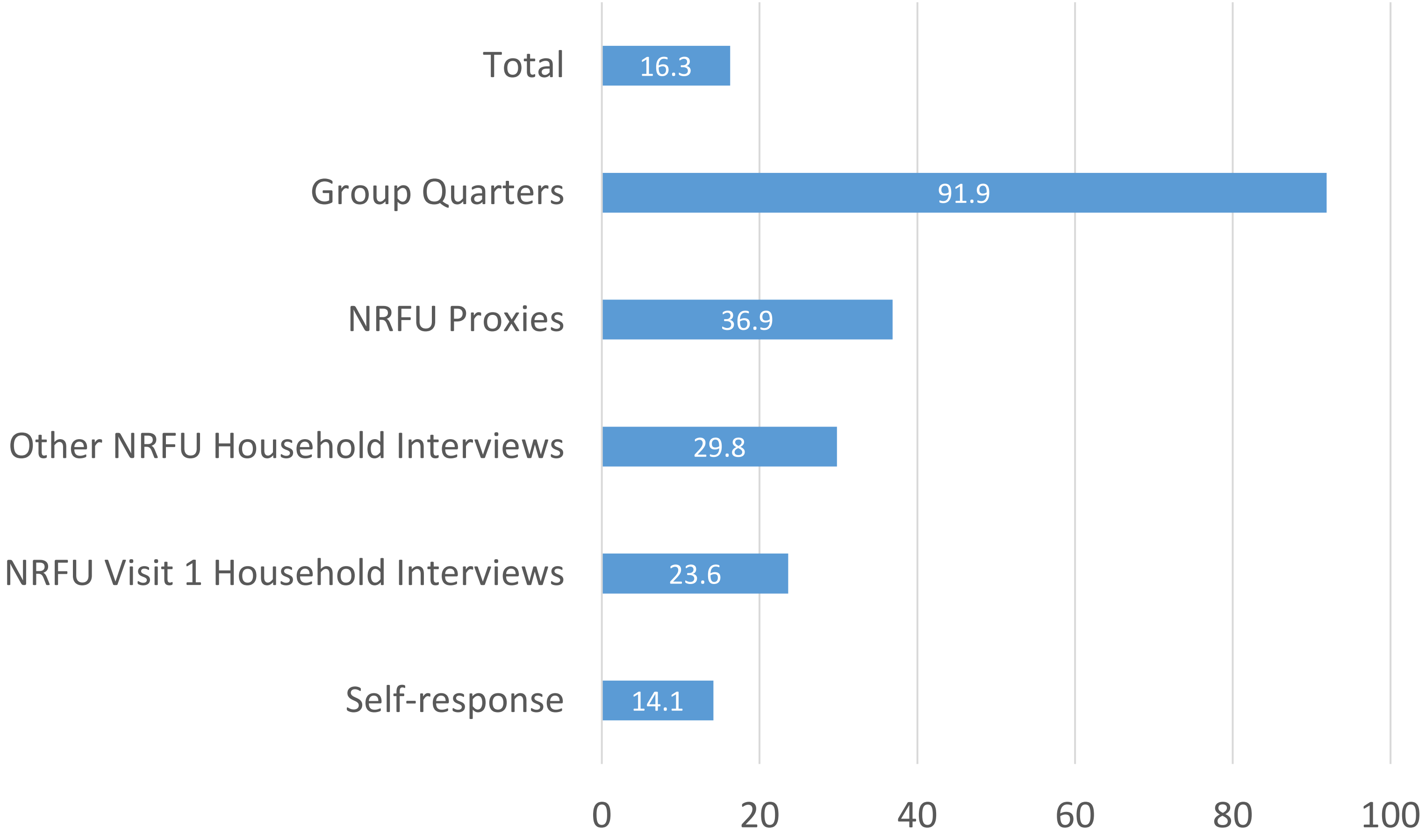
Percent Contribution to Completeness by 2020 Census Operation

Among values filled in by CEF



Address disagreement by 2020 Census response mode

Among people with addresses in both EDF and CEF, percent with disagreement



DRB Clearance Number: CBDRB-FY24-0484

Conclusions: Data integration adds value relative to each source alone

- Integrated data include more people and more complete address and demographic information than either source alone
  - EDF contributes more to date of birth, age, sex
  - CEF contributes more to address, race, Hispanic origin

- Value added of survey-style data relative to AR varies with response mode

- Could predict the amount of improvement to the AR data that survey-style collection will yield for each address

Thank you!

David Brown
j.david.brown@census.gov