Comparing 2020 U.S. Census and Administrative Record Population and Housing Estimates

J. David Brown

U.S. Census Bureau

International Total Survey Error Workshop

October 8, 2021

Any opinions and conclusions expressed herein are those of the author(s) and do not represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed (CBDRB-FY21-CES014-048 and CBDRB-FY2021-CES005-029).

United States® Census Bureau

Motivation

- Surveys suffer from declining response rates, increasing costs
- Administrative record-based statistics are cheaper, facilitating higher frequency
- What is the relative accuracy of the two methods?
- Comparison of the estimates may illuminate where errors are in one and/or other source

United States® Census Bureau

Example Administrative Record Data Sources

- Internal Revenue Service tax and information returns
- Medicare, Medicaid, and Indian Health Service health insurance
- Social Security Administration Numident
- Housing and Urban Development housing programs
- Selective Service System
- U.S. Postal Service National Change of Address file
- State welfare program and driver's license data
- Veteran's Service Group of Illinois (VSGI) commercial data

United States® Census Bureau

Administrative Record Data

- Records from January 2019-October 2020
- Include only records for which a unique person identifier can be assigned
  - Aids in unduplication, merging of demographic characteristics
- Exclude persons deceased or not yet born as of reference date
- Exclude non-U.S. residents (tourist and business visa-holders)

United States® Census Bureau

Administrative Record Data

- Use random forest model trained on 2018 American Community Survey (ACS) data to predict each person's location on reference date
- Persons with multiple addresses are assigned fractionally to each, using model probability weights normalized to sum to one
- Demographic characteristics from past Census Bureau surveys, Social Security Administration Numident, and other administrative record sources when available
- Model demographics for missing values using logistic and multinomial logit regressions

United States® Census Bureau

2020 Census Data

- Collection modes include:
  - Internet self-response (52.06% of HUs)
  - Mail self-response (11.84% of HUs)
  - Telephone self-response (1.39% of HUs)
  - Nonresponse Followup field interviews with residents (10.84% of HUs)
  - Nonresponse Followup field interviews with proxy respondents (18.21% of HUs)
  - Group quarters enumeration (0.13% of housing structures)
  - Enumeration at temporary locations
  - Administrative records (4.59% of HUs)
  - Count Imputation (0.93% of HUs)

- Publicly released P.L. 94-171 redistricting file data, excluding Puerto Rico

- TopDown Algorithm applied to Census Edited File to protect data confidentiality

- Use county-level data by age (under 18 and 18+) and race/ethnicity

United States® Census Bureau

National Population and Housing Unit Totals

	2020 Census	AR
Total Persons	331,400,000	338,800,000
Persons with Full Geography	331,400,000	330,200,000
Persons with Only County and State	0	8,293,000
Persons with Only State	0	301,000
Total Occupied Housing Units	126,800,000	119,300,000
Occupied Housing Units in Census Frame	126,800,000	114,400,000
Occupied Housing Units not in Census Frame	0	4,951,000
Total Housing Units with any AR		138,100,000

United States® Census Bureau

The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CES014-048).

National Group Quarters Population

	2020 Census	AR	AR Percent of 2020
Correctional facilities for adults	1,967,000	164,000	8.3
Juvenile facilities	88,000	25,000	28.4
Nursing facilities/Skilled-nursing facilities	1,627,000	302,000	18.6
Other institutional facilities	71,000	17,500	24.6
College/University student housing	2,792,000	223,000	8.0
Military quarters	328,000	21,000	6.4
Other noninstitutional facilities	1,365,000	515,000	37.7
Missing group quarters type	0	44,500	
Total	8,239,000	1,312,000	15.9

United States® Census Bureau

The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CES014-048).

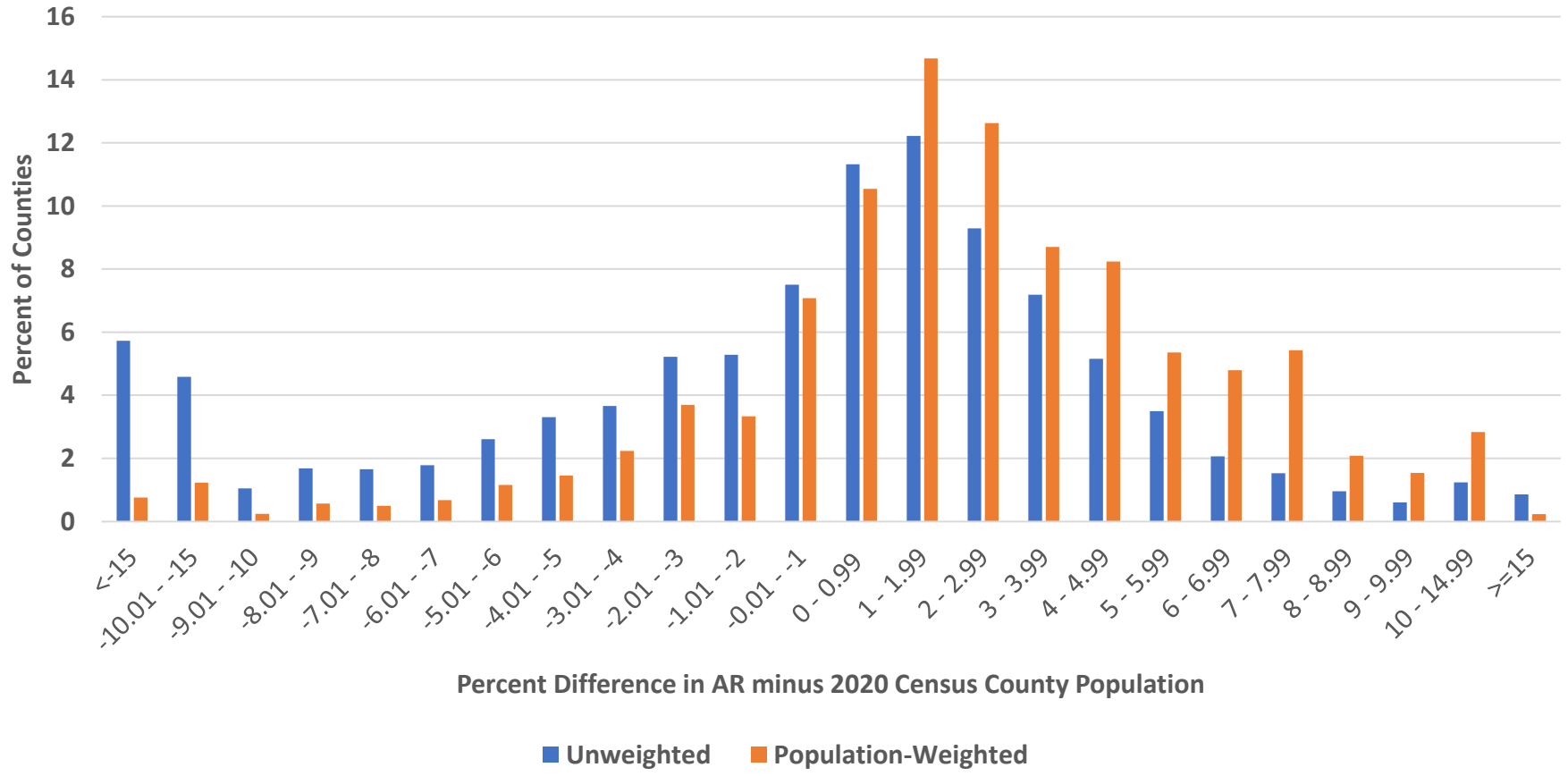
National Population by Race/Ethnicity

	2020 Census	AR	Percent of 2020 Census	Percent of AR	% Difference between AR and 2020 Census
Hispanic	62,080,000	67,910,000	18.7	20.0	9.0
NH White Alone	191,700,000	195,300,000	57.8	57.6	1.9
NH Black Alone	39,940,000	40,180,000	12.0	11.9	0.6
NH AIAN Alone	2,252,000	3,231,000	0.7	1.0	35.7
NH Asian Alone	19,620,000	15,880,000	5.9	4.7	-21.1
NH NHPI Alone	622,000	703,000	0.2	0.2	12.2
NH SOR Alone	1,690,000	6,408,000	0.5	1.9	116.5
NH Two or More Races	13,550,000	9,181,000	4.1	2.7	-38.4
Under 18 Years of Age	73,110,000	75,500,000	22.1	22.3	3.2
18 or More Years of Age	258,300,000	263,300,000	77.9	77.7	1.9

United States® Census Bureau

The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CES014-048).

Distribution of Percent Difference in AR minus 2020 Census County Population



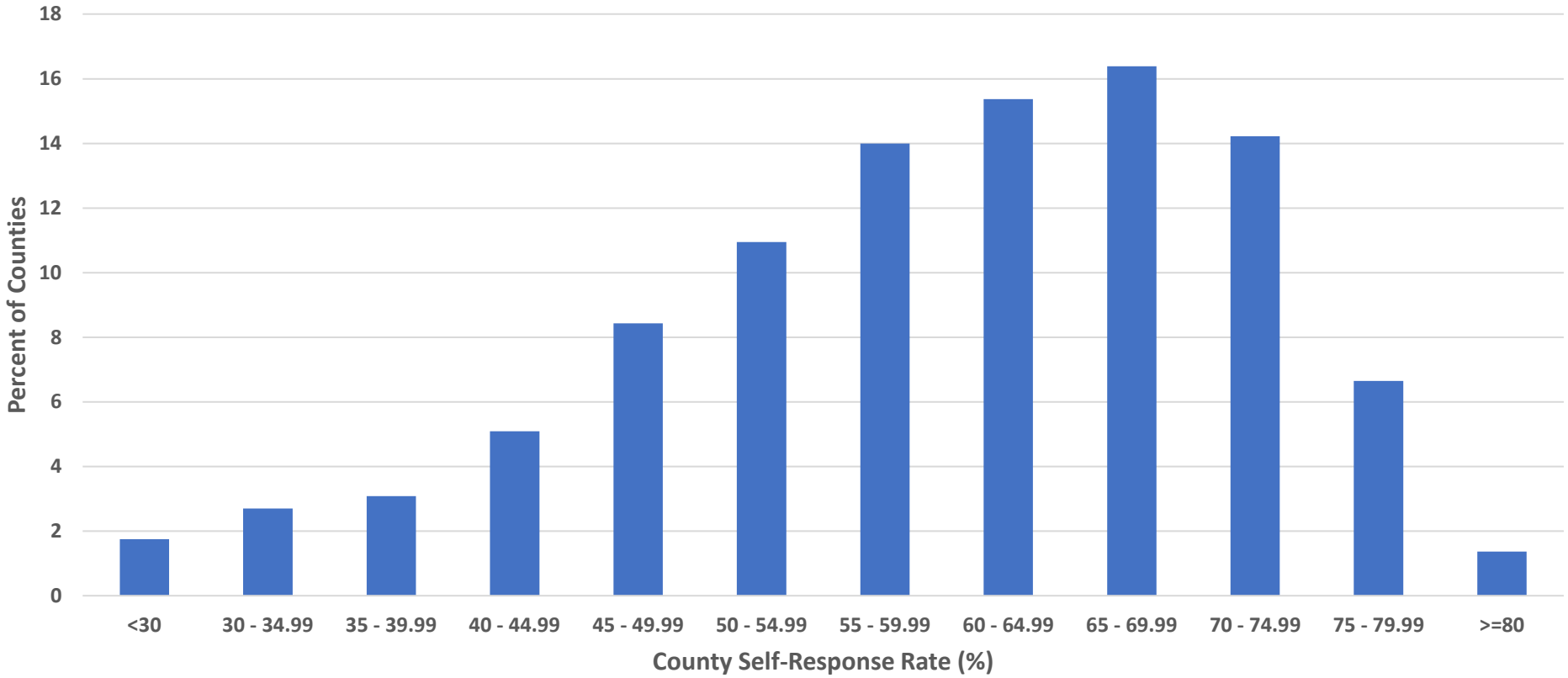
The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY21-CES014-048).

Self-Response Rate as Measure of Census Quality

- Nonresponse Followup (NRFU) fieldwork conducted several months after reference date
- Some are proxy responses and imputations
- Demographic characteristics likely to be even less accurate than person counts

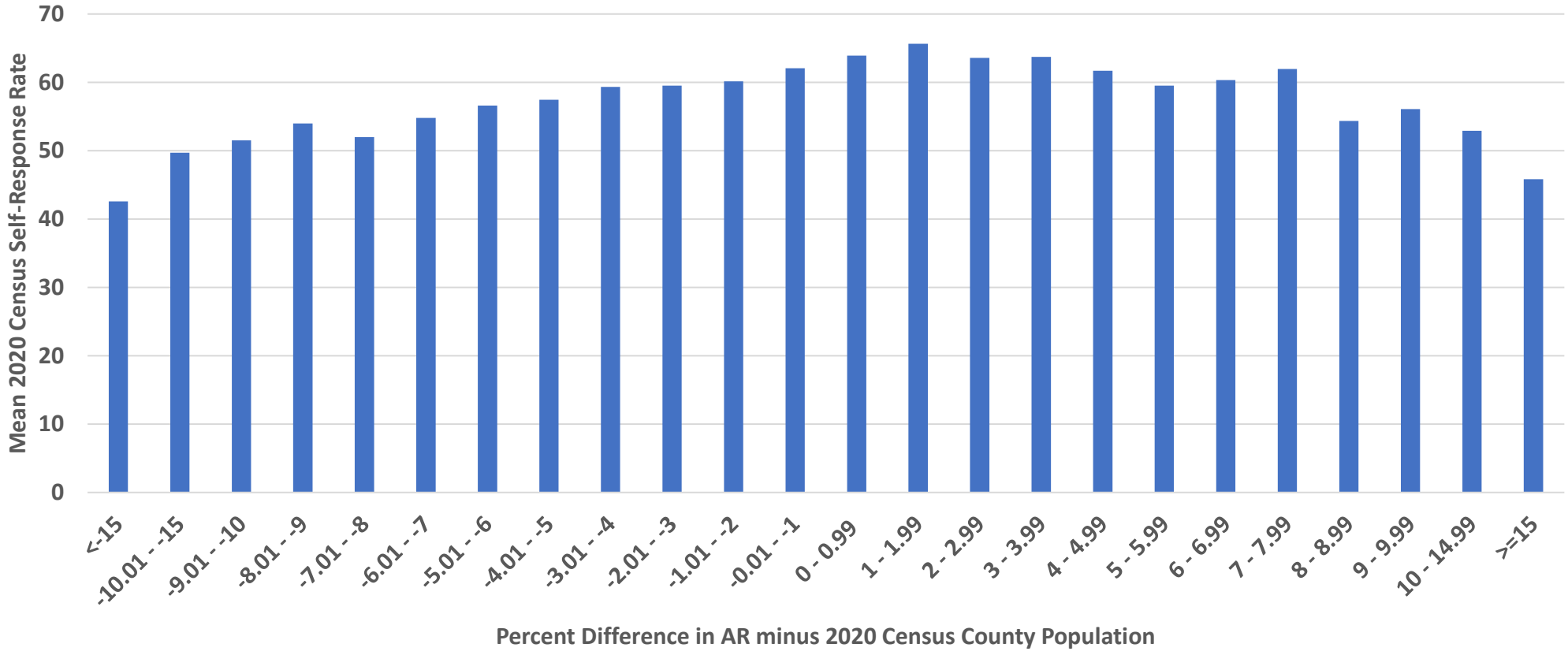
United States® Census Bureau

2020 Census County Self-Response Rate Distribution



The data presented in this figure are publicly available.

Association Between 2020 Census Self-Response and Percent Difference in AR minus 2020 Census County Population



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY21-CES014-048).

% Difference in AR minus 2020 Census County Population Regressions

- OLS regressions

- Dependent variable is percent difference between AR and 2020 Census county population for the particular group, or the absolute value of the percent difference

United States® Census Bureau

% Difference in AR minus 2020 Census County Population Regressions: Explanatory Variables

- 2020 self-response rate
- Linked without SSN or ITIN
- SSA Numident baby
- Linked through parents
- Only in Medicaid
- No housing structure geography
- Non-residential address
- No tax data
- Only in commercial data

United States® Census Bureau

% Difference in AR minus 2020 Census County Population Regressions: Explanatory Variables

- State with welfare program data
- State with driver's license data
- No vintage 2020 data
- Just one source
- Multiple counties
- Mean probability of being at address
- Recent AR for non-alive persons
- Non-U.S. resident AR
- No person identifier in commercial data

United States® Census Bureau

	Correlation with Percent Difference in County Population	Percent Difference in County Population Regression	Correlation with Absolute Value of the Percent Difference in County Population	Absolute Value of Percent Difference in County Population Regression
2020 Census Self-Response Rate	0.3455	0.01590	-0.4437	-0.02099
	0.0000	0.02472	0.0000	0.02153
Linked without SSN or ITIN	0.1783	1.528	0.2547	-0.02967
	0.0000	0.1770	0.0000	0.2079
SSA Numident Baby	0.1034	0.3279	-0.0091	0.1793
	0.0000	0.4592	0.6094	0.2368
Linked through Parents	-0.4547	-0.7977	0.5240	0.1042
	0.0000	0.3147	0.0000	0.2786
Medicaid	-0.2831	-1.784	0.3068	1.006
	0.0000	0.5693	0.0000	0.5661
No Housing Structure Geography	-0.4328	-0.5821	0.5701	0.4123
	0.0000	0.09940	0.0000	0.08669
Non-Residential Address	-0.0852	0.1480	0.2673	0.5600
	0.0000	0.3693	0.0000	0.2678
No Tax Data	-0.2542	-0.05822	0.4987	0.2391
	0.0000	0.08387	0.0000	0.06889
Only in Commercial Data	-0.3308	-0.09135	0.3478	-0.3967
	0.0000	0.1976	0.0000	0.1650

United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY2021-CES005-029).

	Correlation with Percent Difference in County Population	Percent Difference in County Population Regression	Correlation with Absolute Value of the Percent Difference in County Population	Absolute Value of Percent Difference in County Population Regression
State with Welfare Program Data	0.0468	0.7976	-0.0503	-0.1762
	0.0087	0.3078	0.0048	0.2489
State with Driver's License Data	0.1088	1.240	-0.0118	0.2275
	0.0000	0.5733	0.5073	0.4641
No Vintage 2020 Data	-0.1243	0.01185	0.3628	-0.1433
	0.0000	0.08061	0.0000	0.06585
Just One Source	-0.1956	0.1088	0.4268	0.1054
	0.0000	0.05101	0.0000	0.04250
Multiple Counties	-0.0112	-0.2583	-0.0830	0.04367
	0.5287	0.04658	0.0000	0.04005
Mean Probability of Being at Address	0.1884	0.03322	-0.2257	-0.1402
	0.0000	0.06146	0.0000	0.04465
Recent Administrative Records for Non-Alive Persons	-0.1903	0.001922	0.0798	-0.01538
	0.0000	0.05454	0.0000	0.04471
Non-U.S. Resident Administrative Records	0.0919	0.05235	0.0626	0.01767
	0.0000	0.02886	0.0004	0.03887
No Person Identifier in Commercial Data	-0.2575	0.006202	0.3653	0.03497
	0.0000	0.04771	0.0000	0.04185
R-Squared		0.3602		0.3747

United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY2021-CES005-029).

	Correlation with 2020 Census Self-Response Rate	2020 Census Self-Response Rate Regression
Linked without SSN or ITIN	-0.1748	0.1993
	0.0000	0.1841
SSA Numident Baby	0.1163	1.579
	0.0000	0.8840
Linked through Parents	-0.6330	-0.5497
	0.0000	0.2727
Medicaid	-0.3433	1.681
	0.0000	0.3964
No Housing Structure Geography	-0.6597	-0.2182
	0.0000	0.08983
Non-Residential Address	-0.3424	-1.023
	0.0000	0.3304
No Tax Data	-0.6472	-0.5901
	0.0000	0.0713
Only in Commercial Data	-0.6376	-0.6680
	0.0000	0.1770

United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY2021-CES005-029).

	Correlation with 2020 Census Self-Response Rate	2020 Census Self-Response Rate Regression
State with Welfare Program Data	0.1180	2.097
	0.0000	0.3117
State with Driver's License Data	0.0744	-1.620
	0.0000	0.5297
No Vintage 2020 Data	-0.3596	0.1208
	0.0000	0.07373
Just One Source	-0.5334	-0.06155
	0.0000	0.05130
Multiple Counties	0.1578	-0.08706
	0.0000	0.04366
Mean Probability of Being at Address	0.2079	0.08113
	0.0000	0.04359
Recent Administrative Records for Non-Alive Persons	-0.4177	-0.4074
	0.0000	0.04343
Non-U.S. Resident Administrative Records	0.0217	0.1900
	0.2237	0.05808
No Person Identifier in Commercial Data	-0.6782	-0.8869
	0.0000	0.03727
R-Squared		0.7235

United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY2021-CES005-029).

Conclusions (I)

- Higher AR population overall
- Difference under 3% in half of counties
- AR estimate is higher for Hispanics and children, lower for NH Asians
- AR weaknesses
  - Poor group quarters coverage
  - Outdated race/ethnicity information
  - Uneven coverage across counties

United States® Census Bureau

Conclusions (II)

- Wide variation in self-response rates
- Inverse-U-shaped association between self response rate and population difference
- Lower AR estimates associated with
  - Persons lacking records with housing structure-level geography
  - Persons not in tax data
  - Persons only in commercial data
  - Persons only in Medicaid
  - Persons at non-residential addresses
  - Children found only through links with parents
- If we were to exclude lower quality AR records, AR county estimates would be further from 2020 Census estimates

United States® Census Bureau

Conclusions (III)

- Persons linked without SSN located in areas where AR estimate is higher
- State data available in areas where AR coverage is higher
- AR-2020 Census gap – 2020 Census self-response rate correlation vanishes when including AR characteristics
- Self-response rate lower in counties where
  - AR lack housing structure geography
  - Non-residential addresses
  - No tax data
  - Only in commercial data
  - No person identifiers in commercial data

United States® Census Bureau

Questions for Further Study

- To what extent are race/ethnicity differences due to
  - different responses about same people
  - different edit procedures
  - different persons in AR vs. 2020 Census

- To what extent are 2020 Census GQ people found in HUs in AR?

- AR estimates higher than 2020 Census where both sources appear more reliable. Addressing AR weaknesses would not only reduce AR-Census gaps where AR estimate is lower, but lead to even higher overall AR estimate. How can we identify erroneous AR inclusions?
  - people not alive
  - non-residents of U.S.
  - incomplete unduplication

- To what extent can different location assignment for same people explain county discrepancies? Which location assignment is more accurate?

United States® Census Bureau