

# Estimating Citizen Voting Age Population: An Experimental Product

**William Bell**

**Senior Mathematical Statistician for Small Area Estimation**

**U.S. Census Bureau**

September 18, 2020

Disclaimer: The information provided in presentation materials is for informational purposes only and may not represent the official position of the Census Bureau or the Department of Commerce. Statements made by individual presenters may not represent the agency's final position on any matter.

Data presented were approved for dissemination by the Census Bureau Disclosure Review Board (CBDRB-FY20-CED006-0031).

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# CVAP Teams

## 2020 Census Methods Internal Expert Panel

John M. Abowd (chair), William Bell, Michael Berning, J. David Brown, John L. Eltinge, Patrick J. Cantwell, Misty L. Heggeness (coordinator), Howard R. Hogan (until retirement), Jenny Hunter-Childs, Christa Jones (deputy chair), V. Thomas (Tom) Mule, Roberto Ramirez, Joseph Schafer, Victoria Velkoff

## Citizen Voting Age Population (CVAP) Technical Working Group

William Bell, J. David Brown (lead), Stephanie (Jamie) Busick, Misty L. Heggeness, Ryan Janicki, Andrew Keller, Darcy Morris, V. Thomas (Tom) Mule, Joseph Schafer, Matthew Spence, Lawrence Warren, Moises Yi

## Citizen Voting Age Population (CVAP) Implementation Team

John M. Abowd, Michael Berning, J. David Brown, Stephanie (Jamie) Busick, Michael Clark, Jaya Damineni, Karen Deaver, Michael Hawes, Liza Hill, Cynthia Davis Hollingsworth, Jane Ingold, Andrew Keller, V. Thomas (Tom) Mule, Danielle Ringstrom, Teresa Sabol, David Sheppard, Damon Smith, Steven Smith, Matthew Spence, Thomas Thornton, James Treat (chair), Epaphrodite Uwimana, James Whitehorne

# Outline of Presentation

- I. Background on CVAP, data sources, and record linkage
- II. Summary of main points of the presentation
- III. Results related to fitness for use of the data sources
- IV. Results for estimation of citizens – testing done using the 2010 Census Edited File as the frame, combining it with admin and survey data sources

# Citizen Voting Age Population (CVAP) Program

- A special tabulation of the population of U.S. citizens living in housing units and group quarters by voting age (18+), race, and ethnicity, down to census block groups, published by the Redistricting and Voting Rights Data Office, U.S. Census Bureau (RDO@CENSUS.GOV).
- Historically used for research, evaluation, and enforcement of the *Voting Rights Act*, including estimates required by Section 203 (identification of jurisdictions required to provide language support for participation in the electoral process for citizens with limited English capabilities).
- Original CVAP estimates were produced from the 2000 Census long form.
- With elimination of the long form in 2010, for the last decade CVAP has been based on [American Community Survey \(ACS\)](#) five-year estimates, updated annually.
- The post-2020 Census CVAP Special Tabulation estimates will be produced for Census tabulation blocks using 2020 Census and administrative records data, and possibly survey data sources.

# Race/ethnicity groups for the 2020 CVAP

## Not Hispanic or Latino

1. American Indian or Alaskan Native (AIAN) alone
- 2. Asian alone**
- 3. Black or African American alone**
4. Native Hawaiian or Other Pacific Islander alone
- 5. White alone**
6. Some Other Race alone
7. AIAN and White
8. Asian and White
9. Black or African American and White
10. AIAN and Black or African American
11. Remainder of two or more race responses

## **12. Hispanic or Latino**

Results here focus on the four largest race/ethnicity groups, which are in bold.

# Data sources available for CVAP

- 2020 Census
  - Census Unedited File (CUF): used in record linkage
  - Census Edited File (CEF): serves as frame for estimation of citizens
- SSA Numident
  - Applications for Social Security Numbers (SSNs) and subsequent transactions
  - Primary reference file for the Census Bureau's Person Identification Validation System (PVS) (Wagner and Layne, 2014) – used to assign Protected Identification Keys (PIKs) for record linkage
  - Information on nativity (country of birth), citizenship and noncitizen legal status
  - Covers large share of population – Nearly 90% of persons in the 2010 Census were successfully found in Numident (Rastogi and O'Hara, 2012)

SSA = Social Security Administration. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Additional Administrative Sources of Citizenship Data

- Department of State Passport Data (all U.S. passports, citizens)
- USCIS naturalizations (citizens) – all persons naturalized since 2001
  - Exception: children automatically naturalized because their parents naturalized when child did not also get a naturalization certificate
- USCIS lawful permanent residents – green card holders (noncitizens)
- ITINs – individual taxpayer identification numbers (noncitizens)

Very limited additional coverage found from the following sources:

- ADIS – Customs and Border Protection Arrivals and Departures Information System
- SEVIS – Immigration and Customs Enforcement (ICE) Office of Student and Exchange Visitor Information System
- WRAPS – State Dept. Worldwide Refugee Admissions Processing System
- Federal law enforcement records (U.S. Marshals Service, Bureau of Prisons)
- SNAP and TANF data from some states
- Driver's license files from Nebraska and South Dakota

USCIS = U.S. Customs and Immigration Service. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. SNAP = Supplemental Nutrition Assistance Program. TANF = Temporary Assistance to Needy Families. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Survey Data Sources

- ACS (American Community Survey)
- CPS (Current Population Survey)
- AHS (American Housing Survey)
- SIPP (Survey of Income and Program Participation)

# Record Linkage

- Link records from other files to records in a Reference File constructed from SSA Numident records and occurrences of ITINs. This allows assignment of Protected Identification Keys (PIKs) to the other file records, which are then used for matching of records across the various files.
  - Probabilistic record linkage
  - SSN verification, then combinations of name, address, date of birth
- EPIKs: Unduplicated unlinked records with sufficient PII, put in Enhanced Reference File (ERF), and assign PIKs to as many ERF records as possible that do not already have PIKs (while maintaining record linkage quality)
  - For simplicity, we refer to these “enhanced process PIKs” as EPIKs
  - EPIK process incorporates noncitizens without SSNs
- 2020 Census records, and other data sources, are assigned PIKs and EPIKs via linkage to the Reference File and the Enhanced Reference File
- Link administrative and survey records containing citizenship to the 2020 Census via the PIKs and EPIKs.

SSA = Social Security Administration. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Record Linkage (continued)

- EPIK linkages provide citizenship information for just 0.11% of the 2018 ACS estimated population, including both primary and secondary source linkages.
- Linkage process assigns separate quality score for each link attempt (combination of linkage variables)
  - SSN verification most reliable
  - Matching including address is more reliable than name and date of birth matching without address
- Create a single quality indicator (probability of correct linkage) from linked records using information on the link attempts and the attempt's quality score. See slides #44-45.
  - Exclude links with  $\text{Pr}(\text{link correct}) < .99$  from estimation process to minimize linkage error.

# Notes on quality of citizenship data records

- Many administrative sources require documentation of citizenship. We regard these data as highly reliable.
  - Ex. SSA Numident, passport data, USCIS naturalizations
- Data that is not current and indicates a noncitizen can be incorrect since the person may have since naturalized.
  - Currency is not of much concern for data indicating citizens
- Survey data on citizenship is subject to various errors:
  - Incorrect status reported (more frequent for true noncitizens)
  - Out-of-date reports of noncitizen status
  - Imputations for nonresponse to citizenship question on survey
- Record linkage errors can lead to errors in citizenship status for any data source.
  - Apply record linkage quality threshold to minimize linkage errors

# Summary of main points of the presentation

1. Combining the primary administrative data sources provides reliable data on citizenship for a large percentage of the population (91% in tests using 2018 ACS data or 2010 Census data)
  - Administrative sources used included SSA Numident, State Dept. passport data, USCIS lawful permanent residents and naturalizations, ITINs
  - Additional data sources beyond these provide very limited additional coverage
  - Prediction via imputation or modeling is needed for the cases not assigned citizenship status
2. Four approaches investigated for estimation of citizens:
  - Business Rules (BR) plus Hot Deck imputation from BR cases for the Census NBR (non-business rule) cases
    - Business Rules assign citizenship status to Census records based on the citizenship data sources linked to each record (assignments made for 91% of the records in the testing done, leaving 9% for imputation)
  - Business Rules plus Logistic Regression fitted to BR cases, applied to Census NBR cases
  - Business Rules plus Logistic Regression fitted to ACS NBR cases, applied to Census NBR cases
  - Latent Class model using multiple citizenship indicators (including Census BR and NBR cases, ACS data, ...)

ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. USCIS = U.S. Customs and Immigration Service. See slides #7-8 for a list of the additional data sources. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Summary of main points of the presentation (continued)

3. We compared results of the four estimation approaches, applied using the 2010 CEF (Census Edited File) and associated administrative and survey records, for the following four subsets of the CEF records:
  - BR cases (91%): very strong agreement across approaches
  - NBR-PIK: large differences across approaches, but this is a very small fraction of the population
  - NBR-SS: substantial differences across approaches for Hispanics and Nonhispanic Asian Alone
    - SS = sent to search for a PIK, NSS = not sent to search for a PIK
  - NBR-NSS: some differences seen across approaches, but generally smaller than for NBR-SS
4. An important difference between the estimation approaches is what data serve as the “training sample” used to produce a predictor of citizenship for the CEF NBR cases
  - **Hot Deck** imputation: training sample = CEF BR cases
  - **BR Logistic** regression: training sample = CEF BR cases
  - **ACS Logistic** regression: training sample = ACS NBR cases with ACS reported citizenship status
  - **Latent Class** model: no distinct subset of data used as a training sample; draws information from CEF and ACS BR and NBR cases, and other data sources

# Administrative Record Coverage of the 2018 ACS Estimated Population

Source	Percent of ACS Population (PIKs)
Numident	90.4
Citizen value for citizenship	66.9
Missing citizenship, U.S. born (citizens)	14.6
Noncitizen value for citizenship	7.8
Foreign born or uncertain country of birth, missing citizenship	1.0
U.S. Passports (citizens)	48.6
USCIS	11.5
Naturalizations (citizens)	6.6
Lawful permanent residents and refugees (noncitizens)	5.0
ITINs (noncitizens)	0.5

Notes: These percentages use ACS survey weights. The total 2018 ACS estimated population age 18 and over is 253,800,000.

# Amount of Agreement on Citizenship Status by SSA Numident, Passports, USCIS, ITINs

	% of 2018 ACS Estimated Population	
	No restrictions on record linkage	With record linkage quality restrictions
<b>Totals</b>		
Agreement	87.36	87.90
on citizens	81.46	81.96
on noncitizens	5.91	5.94
Disagreements	3.43	2.68
Missing (no linked admin records citizenship)	9.21	9.42

USCIS = U.S. Customs and Immigration Service. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. "With record linkage quality restrictions" indicates links are accepted only if the predicted  $\Pr(\text{link is correct}) \geq .99$ ; see slide #45. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Citizenship Business Rules (using 2018 ACS as the population frame)

	Rule assignment	% of 2018 ACS Population	
		No restrictions on record linkage	With record linkage quality restrictions
<b>Criteria for assigning citizen</b>			
Numident citizen	Citizen	66.94	66.94
Numident missing citizenship but U.S.-born	Citizen	14.63	14.63
U.S. passport	Citizen	3.01	2.79
USCIS naturalization certificate	Citizen	0.29	0.28

**If not U.S. citizen according to above criteria, even without record linkage quality restriction:**

Numident noncitizen	Noncitizen	5.27	5.33
ITIN	Noncitizen	0.52	0.52
USCIS lawful permanent resident or refugee	Noncitizen	0.12	0.09
ICE SEVIS record	Noncitizen	0.06	0.06
ADIS record not born in U.S.	Noncitizen	D	D
WRAPS record	Noncitizen	D	D

<b>Has PIK, but no citizenship assignment</b>	<b>Model</b>	0.05	0.37
<b>No PIK</b>	<b>Model</b>	<b>8.98</b>	<b>8.98</b>

“With record linkage quality restrictions” indicates links are accepted only if the predicted  $\Pr(\text{link is correct}) \geq .99$ ; see slide #45. “D” indicates that the number is suppressed due to disclosure restrictions. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. USCIS = U.S. Customs and Immigration Service. See slides #6-8 for definitions of the various data sources. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Citizen and Noncitizen Shares for Business Rules

<b>Business rule assignment</b>	<b>% of 2018 ACS Estimated Population</b>	
	<b>No restrictions on record linkage</b>	<b>With record linkage quality restrictions</b>
Citizen	84.88	84.64
Noncitizen	5.98	6.02
Missing	9.03	9.34

“With record linkage quality restrictions” indicates links are accepted only if the predicted  $\Pr(\text{link is correct}) \geq .99$ ; see slide #45. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

## Business Rules vs. 2008-2012 ACS Estimated Percent Citizens Sample with Both Business Rules and ACS As-Reported Citizenship Present

<b>Race/ethnicity</b>	<b>Business rules</b>	<b>ACS As-Reported</b>
<b>Total</b>	<b>93.5</b>	<b>93.4</b>
NH White Alone	98.4	98.5
NH Black Alone	95.5	95.7
Hispanic	73.3	72.4
NH Asian Alone	70.9	69.3

# Comparison of 2018 ACS As-Reported to 2018 Business Rules Citizenship

	Column Percents		
	BR Citizen	BR Noncitizen	ACS Total
<b>ACS Citizen</b>	99.29	<b>10.60</b>	
<b>ACS Noncitizen</b>	<b>0.71</b>	89.40	
<b>Benchmark Total</b>	100.00	100.00	

	Cell Percents		
<b>ACS Citizen</b>	93.11	<b>0.66</b>	93.77
<b>ACS Noncitizen</b>	<b>0.66</b>	5.56	6.23
<b>Benchmark Total</b>	93.78	6.22	

# Four approaches to determining or predicting citizenship for Census cases

- **Hot Deck:** Business Rules (BR) plus hot deck imputation (BR covers  $\approx$  91% of the data)
- **BR logistic:** BR plus logistic regression with BR data
- **ACS logistic:** BR plus logistic regression with NBR cases in the ACS sample that have a response to the ACS citizenship question
- **LC:** Latent class model

# Business Rules (BR) plus hot deck imputation (Hot Deck)

- Accept BR determinations
- For NBR cases, impute citizenship from nearest neighbor (on address list) within imputation cells defined by a cross-classification of
  - Race and detailed Hispanic origin (17 groups)
  - Whether or not the housing unit had a non-PIKed person within the unit
  - Age groups: 18 – 29, 30 – 49, 50+
- There are small numbers of resolved cases in some cells.
- Very few cases needing imputation are in units where all persons are PIKed

NBR = no business rules citizenship determination. Detailed Hispanic origin categories are Mexican, Puerto Rican, Cuban, Central American, Latin American, and Other Hispanic. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Business rules plus logistic regression with BR data (BR logistic)

- Accept BR determinations using linked data that pass a record linkage quality threshold. (See slide #44 for details.)
- Logistic regressions used to predict probabilities of citizenship for NBR cases:
  - Fit logistic regression for **BR householders** using tract indicators, CVAP race and Hispanic origin categories, and age domains (under 29, 30 to 49 and 50+) as main effects). Use this model to predict citizenship for **householders without BR citizenship**.
  - Fit logistic regression for **other household members with BR citizenship** using relationship to householder (11 categories), CVAP race and Hispanic origin categories, and age domains. This was **done separately for the cases where householders were BR citizens and for the cases where householders were BR noncitizens (two models)**.
  - Predicted citizenship probabilities for **other household members without BR citizenship** were then obtained from
$$\begin{aligned} \text{Pr(Other is Citizen)} &= \text{Pr(HH is citizen)} \times \text{Pr(Other is Citizen|HH is citizen)} \\ &+ [1 - \text{Pr(HH is citizen)}] \times \text{Pr(Other is Citizen|HH is noncitizen)} \end{aligned}$$
- A different logistic regression model was used for group quarters residents (for GQs, there is no householder)

# BR plus logistic regression with ACS data (ACS logistic)

- **Motivation:** As-reported ACS estimated citizen shares vary widely depending on
  - whether or not citizenship information can be linked to the person's survey record and,
  - if not, the reason why not, especially for race/ethnic groups that have higher noncitizen shares (Asians and Hispanics).
  - This suggests differences between the BR versus NBR data (**nonignorable missingness**).
- **Goal:** Use ACS data to address nonignorable missingness that can arise by using BR cases to develop predictions for the NBR cases.

# BR plus logistic regression with ACS data (ACS logistic, continued)

- Accept BR determinations using linked data that pass a record linkage quality threshold. (See slide #45 for details.)
- Fit logistic regression models to ACS data without BR determinations, but with ACS reported citizenship. Fit separate models to the following different groups of ACS cases:
  - NBR-PIK (no business rules but has PIK)
  - NBR-SS (no business rules and sent to PVS search for a PIK)
  - NBR-NSS (no business rules and not sent to PVS search for a PIK).
- The models use many regression variables including state indicators, age groups, race/ethnicity groups, sex, tenure, etc., plus citizenship status of householder interacted with relative vs. non-relative of householder.
- Apply fitted logistic regression model to CEF NBR cases to predict their citizenship probabilities.

NBR = no business rules citizenship determination. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Latent Class model (LC)

- Treat true citizenship status as a latent variable ( $L$ ), imperfectly measured by multiple items from various data sources (Numident, passport data, USCIS data, ITINs, Bureau of Prisons and U.S. Marshall Service law enforcement data, ACS, CPS, AHS, SIPP).
  - $L$  has three possible states: U.S.-born citizen, foreign-born citizen, and noncitizen
- Fit the latent-class model in two stages for its two parts:
  - **Measurement model** – describes relationships between  $L$  and the items that measure it.
  - **Prevalence model** – describes how the distribution of  $L$  varies over the population in relation to predictors (e.g., logistic regression).
- Carry over fitting results from Stage 1 to Stage 2 via person-level Bayes factors, with their natural interpretation as odds multipliers (for states of  $L$ ).
- Compute probability of citizenship for each person based on all available items.

USCIS = U.S. Customs and Immigration Service. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. CPS = Current Population Survey. AHS = American Housing Survey. SIPP = Survey of Income and Program Participation. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Notes on the four approaches to estimation of citizenship

- The first three accept the BR determinations (covers  $\approx 91\%$  of the data); the latent class modeling does not, but it nearly replicates the BR determinations.
  - In initial test implementations, the first three approaches used slightly different versions of the BRs. This is being harmonized.
- The first two approaches (hot deck and BR logistic) effectively assume that the NBR cases are like the BR cases, conditional on certain information (Missing at Random). However, ACS data provide evidence against this assumption.
- The third approach (ACS logistic) assumes the NBR cases found in the ACS sample are like all the other NBR cases, conditional on certain information. It also accepts the ACS citizenship responses for these cases, which include some error.
- The latent class model draws information from both CEF and ACS BR and NBR cases, and other data sources, to provide information on the NBR cases.

# 2010 CEF Percent of Cases by PIK Group

	<b>Business Rules</b>	<b>NBR-PIK</b>	<b>NBR-SS</b>	<b>NBR-NSS</b>	<b>Population (1,000s)</b>
<b>Total</b>	<b>90.9</b>	<b>0.12</b>	<b>5.8</b>	<b>3.3</b>	<b>234,600</b>
NH White Alone	93.2	0.12	3.9	2.8	157,100
NH Black Alone	88.0	0.04	6.9	5.2	27,320
Hispanic	83.2	0.18	12.8	3.8	33,350
NH Asian Alone	89.1	0.22	7.2	3.5	11,290

NBR-PIK is no business rules and has PIK. NBR-SS is no business rules and sent to PVS search. NBR-NSS is no business rules and not sent to PVS search. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from Four Approaches 2010 CEF, All Cases

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>Total</b>	<b>91.4</b>	<b>91.4</b>	<b>91.1</b>	<b>90.8</b>	<b>91.5</b>
NH White Alone	98.3	98.1	98.2	97.8	98.3
NH Black Alone	95.0	94.9	95.0	93.7	95.3
<b>Hispanic</b>	64.0	<b>64.8</b>	<b>62.5</b>	63.3	<b>65.7</b>
NH Asian Alone	67.7	67.5	67.2	68.3	67.4

The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from Four Approaches 2010 CEF, BR Cases (91% of total pop)

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>Total</b>	<b>92.6</b>	<b>92.5</b>	<b>92.6</b>	<b>92.5</b>	<b>93.0</b>
NH White Alone	98.4	98.4	98.4	98.4	98.5
NH Black	95.5	95.5	95.5	95.4	95.6
Hispanic	67.8	67.5	67.7	67.2	71.3
NH Asian Alone	69.4	68.9	69.4	69.4	69.7

Notes: The 2010-2012 ACS column uses the ACS citizenship values. BR in 2010-2012 ACS is the assignment rules used in BR + Hot Deck (using primary sources only) applied to the same 2010-2012 ACS records as in the 2010-2012 ACS column.

# Estimated Percent Citizens from Four Approaches 2010 CEF, NBR-SS Cases (5.8% of total pop)

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>Total</b>	<b>75.3</b>	<b>76.7</b>	<b>67.0</b>	<b>64.8</b>	<b>73.1</b>
NH White Alone	95.9	94.9	93.6	83.3	95.1
NH Black	91.4	89.5	88.5	70.4	90.9
Hispanic	42.0	48.6	29.0	37.2	33.4
NH Asian Alone	53.2	54.8	41.2	55.5	47.3

NBR-SS is no business rules and sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from Four Approaches 2010 CEF, NBR-SS Cases (12.8% of Hispanic pop) Hispanics

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>Hispanic</b>	<b>42.0</b>	<b>48.6</b>	<b>29.0</b>	<b>37.2</b>	<b>33.4</b>
<b>Mexican</b>	39.9	<b>48.2</b>	<b>25.6</b>	34.0	<b>30.3</b>
Puerto Rican	96.3	<b>70.0</b>	97.4	87.7	97.2
Cuban	59.9	60.4	58.3	49.7	62.8
<b>Central American</b>	28.2	<b>39.3</b>	<b>15.3</b>	31.6	<b>19.2</b>
<b>Latin American</b>	37.1	<b>47.6</b>	<b>26.5</b>	36.5	<b>33.7</b>
Other Hispanic	62.7	60.0	47.2	47.9	75.1

NBR-SS is no business rules and sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from 4 Approaches 2010 CEF, NBR-SS Cases (7.2% of NH Asian pop) NH Asian Alone

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>NH Asian Alone</b>	<b>53.2</b>	<b>54.8</b>	<b>41.2</b>	<b>55.5</b>	<b>47.3</b>
Asian Indian	47.0	49.7	36.9	53.0	40.7
Chinese	51.7	53.5	39.0	52.7	44.1
Filipino	62.9	61.6	54.5	61.0	63.3
Japanese	57.3	61.6	44.3	58.6	54.8
<b>Korean</b>	48.3	53.7	<b>29.1</b>	<b>54.8</b>	<b>35.1</b>
Vietnamese	64.5	59.7	62.7	61.8	65.3
Other Asian	53.6	54.6	39.7	55.5	47.4

NBR-SS is no business rules and sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

## Estimated Percent Citizens from Four Approaches 2010 CEF, NBR-NSS Cases (3.3% of total pop)

Race/Ethnicity	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
<b>Total</b>	<b>86.1</b>	<b>88.1</b>	<b>89.8</b>	<b>91.0</b>	<b>94.4</b>
NH White Alone	96.0	96.2	97.9	98.1	98.2
NH Black	91.7	92.2	95.1	94.9	96.6
Hispanic	54.6	63.1	61.7	68.6	69.2
NH Asian Alone	54.7	62.8	67.8	68.9	76.2

NBR-NSS is no business rules and not sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens for 2010 CEF NBR-SS Group with Alternative Applications of the ACS Logistic Model by Training Sample and Source of Citizenship Status

Race/Ethnicity	ACS logistic approach model used				BR logistic	Hot Deck	Latent Class Model
	ACS NBR-SS Sample, ACS Citizenship	ACS BR Sample, ACS Citizenship	ACS BR Sample, BR Citizenship	CEF BR Sample, BR Citizenship			
<b>Total</b>	<b>69.0</b>	<b>76.6</b>	<b>77.9</b>	<b>77.3</b>	<b>76.7</b>	<b>75.3</b>	<b>64.8</b>
NH White Alone	93.6	95.0	95.1	95.0	94.9	95.9	83.3
NH Black Alone	88.5	89.6	89.8	89.6	89.5	91.4	70.4
<b>Hispanic</b>	<b>29.0</b>	<b>48.2</b>	51.3	50.0	48.6	42.0	37.2
<b>NH Asian Alone</b>	<b>41.2</b>	<b>52.5</b>	56.5	55.4	54.8	53.2	55.5

NBR-SS is no business rules and sent to PVS search. CEF = Census Edited File. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model, not including the alternatives for the ACS logistic approach used here) are discussed on slides #20-26. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Conclusions from comparing estimation approaches using 2010 CEF data as the frame

- The four approaches yield very similar results for citizenship estimates for the total 18+ population at the national level
- Some differences can be seen in (total 18+) national estimates for Hispanics, particularly Mexicans and Central Americans
- We have examined state level estimates in which we see similar patterns in the results though, as expected, with some variations across states. These results have not yet gone through a disclosure review.

# Some larger differences can be seen when the estimates are broken out by PIK status

- BR cases: minimal differences
- NBR-PIK: very large differences, but this is a very small group
- NBR-SS: Large differences for Hispanics and for NH Asians, especially for Mexicans, Central Americans, Latin Americans, and Koreans
  - An experiment that applied the model from the ACS logistic approach in alternative ways showed that the largest contributor to differences between the estimation approaches for the NBR-SS group was whether BR cases or ACS NBR cases were used as the “training sample” for making predictions.
- NBR-NSS: Some differences seen across approaches, but generally smaller than for NBR-SS

NBR-PIK is no business rules and has PIK. NBR-SS is no business rules and sent to PVS search. NBR-NSS is no business rules and not sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Further research planned and underway

- Apply the four approaches using 2018 ACS data as the frame, along with corresponding 2018 administrative sources. See if we get similar results to those shown here from using the 2010 CEF as the frame.
- Harmonize to a common set of business rules.
- Refine the models used, drawing on results of the analyses done to date.
  - Since the ACS provides a much smaller data set than a census, there can be some limitations on model refinement for the application to the 2018 ACS, especially as it relates to detailed population subgroups and geography.
  - For the Latent Class Model, this requires certain enhancements to the modeling software.
  - Research linking of administrative files to census housing unit records (where person records could not be linked)
- Refine the record linkage quality measure.

# Disclosure Avoidance

- Data in this presentation were protected using the Disclosure Review Board's current rules for legacy data at the national level
- The 2020 CVAP data product will be protected using the 2020 Disclosure Avoidance System
  - Using a privacy-loss budget determined by the Data Stewardship Executive Policy Committee and charged to the 2020 Census
  - Using the TopDown Algorithm
  - Constrained to be fully consistent with the geographic, race and ethnicity definitions used in Table P4 of the 2020 PL94-171 redistricting data

# Questions for the committee

1. Should we use the secondary data sources and EPIK linkages given the limited additional coverage that they provide?
2. Do you have suggestions for how we decide on which data to use as the “training sample” for developing citizenship predictions for the cases not covered by the Business Rules?
  - The BR cases themselves, or some subset – Issue: evidence that the BR cases differ from the non-BR cases.
  - ACS non-BR cases with ACS as-reported citizenship – Issues: evidence of reporting error in ACS, particularly for noncitizens, plus potential for 2020 ACS to be less comparable to the 2020 Census (than was the case in 2010).
  - Use the Latent Class model, which makes use of both these data sources, and others.
  - Formulate some mathematical comparison criterion?
  - Combine results from more than one estimator – how?
3. Do you have suggestions for explaining (primarily to a technical audience) how we made this decision?
4. Do you have suggestions for ways to convey uncertainty reflecting prediction error, which is not due to sampling error, and is partly due to certain systematic errors? (Note the second question.)

Note: We plan to release a report on this work by October 31, 2020 that will indicate which estimation approach we have chosen and why, so near-term responses to these questions are appreciated.



# COMMITTEE DISCUSSION

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**

# Additional slides with more details follow



# Social Security Administration (SSA) Numident File



- Applications for Social Security Numbers (SSNs) and subsequent transactions
- Primary reference file for the Census Bureau's Person Identification Validation System (PVS) (Wagner and Layne, 2014)
- Information on nativity, citizenship and noncitizen legal status

## Strengths

- Covers large share of population – Nearly 90% of persons in the 2010 Census were successfully found in Numident (Rastogi and O'Hara, 2012)
- Numident report of citizenship can be accepted with high confidence

## Weaknesses

- Reports of non-citizenship are less reliable, because naturalizations are not always reported to SSA
- No coverage of those in the resident population without SSNs

# Record Linkage Process

- Link records from other files to records in a Reference File constructed from SSA Numident records and occurrences of ITINs. This allows assignment of Protected Identification Keys (PIKs) to the other file records, which are then used for matching of records across the various files.
  - Probabilistic record linkage
  - SSN verification, then combinations of name, address, date of birth
- Unduplicate unlinked records with sufficient PII, put in Enhanced Reference File (ERF), and assign PIKs to as many ERF records as possible that do not already have PIKs (while maintaining record linkage quality)
  - For simplicity, we refer to these “enhanced process PIKs” as EPIKs
  - EPIK process incorporates noncitizens without SSNs
- 2020 Census records are assigned PIKs and EPIKs via linkage to the Reference File and the Enhanced Reference File
- Link administrative and survey records containing citizenship to the 2020 Census via the PIKs and EPIKs.

SSA = Social Security Administration. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Record linkage quality threshold: Business rules plus logistic regression with BR data

Linkage process assigns a separate quality score for each linkage attempt (PVS module and pass)

Accept BR determinations using linked data that pass the following record linkage quality threshold.

- Quality threshold is  $\text{Pr}(\text{correct link}) \geq .99$ , where the linkage probabilities were determined by a decision tree analysis applied to links between Numident foreign-born records and other source records that indicated noncitizen. Note that if another source says noncitizen and the Numident says they were U.S.-born, this is likely to be a linkage error.
- Variables used in making the decision tree were source, PVS module and pass combined indicator, and the record linkage score. This was done separately for each state.
- Decision tree predictor developed with noncitizen records was also used to predict probabilities of correct links for citizens.

PVS = Person Identification Validation System, which is used to assign PIKs, the Protected Identification Keys. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Record linkage quality threshold: Business rules plus logistic regression with ACS data

Accept BR determinations using linked data that pass the following record linkage quality threshold.

Create single quality score from logistic regressions with noncitizen observations in the file.

- Dependent variable = 1 if linked to foreign-born Numident record, = 0 if linked to U.S.-born Numident record
- Independent variables are linkage attempt and score
- Fitted logistic regression model is applied to linked records (whether indicating citizen or noncitizen) to predict the probability that the link is correct. (Where tables indicate “record linkage quality restriction” this mean links are accepted only if their predicted probability of being correct is  $\geq .99$ .)

# Conclusions about fitness for use of data sources

- Combining the primary administrative data sources provides reliable data on citizenship for a large percentage of the population (91% as estimated using 2018 ACS data)
  - SSA Numident, State Dept. passport data, USCIS lawful permanent residents and naturalizations, ITINs (with limited additional return from the ADIS, SEVIS, and WRAPS data)
- Additional data sources (SNAP/TANF, driver's licenses, BOP, USMS, ACS, AHS, CPS, and SIPP) provide very limited additional return due to:
  - Limited population coverage of most sources (for surveys, ACS is the one exception)
  - Overlap with the primary admin sources, especially for citizens (incremental coverage is just 0.03%, as estimated using 2018 ACS population)
  - Records for noncitizens that are out-of-date
  - Record linkage problems with some sources (assessing quality of record links is important)

ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. USCIS = U.S. Customs and Immigration Service. See slides #7-8 for definitions of the additional data sources. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Shares of 2018 ACS Estimated Population by Source Citizenship Combinations

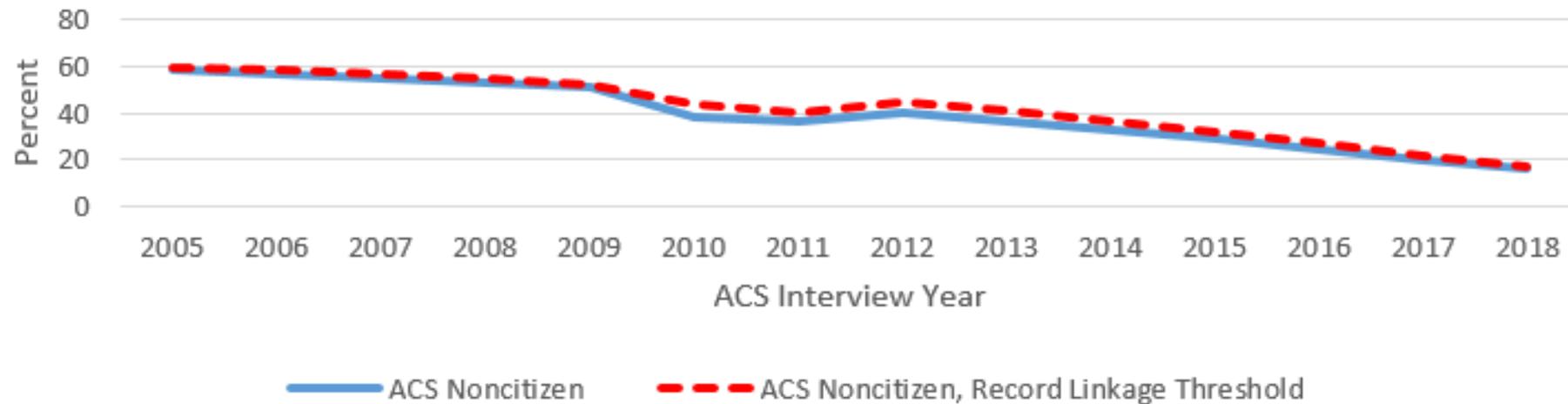
## Disagreements

Numident	U.S. Passport	USCIS	ITIN	% of 2018 ACS Estimated Population	
				No restrictions on record linkage	With record linkage quality restrictions
Noncitizen	Citizen	Citizen	Absent	1.97	1.68
Noncitizen	Absent	Citizen	Absent	0.27	0.26
Noncitizen	Citizen	Noncitizen	Absent	0.23	0.12
Noncitizen	Citizen	Absent	Absent	0.08	0.41
<b>Cumulative Total</b>				<b>2.55</b>	<b>2.47</b>
Citizen	Absent	Noncitizen	Absent	0.09	0.04
Citizen	Citizen	Noncitizen	Absent	0.67	0.16
Missing Citizenship, Foreign-Born	Citizen	Noncitizen	Absent	0.11	< 0.01
<b>Cumulative Total</b>				<b>3.42</b>	<b>2.68</b>
Missing Citizenship, U.S.-Born	Absent	Noncitizen	Absent	< 0.01	D
Missing Citizenship, U.S.-Born	Citizen	Noncitizen	Absent	< 0.01	D
Absent	Citizen	Citizen	Noncitizen	0.00	0.00
Absent	Citizen	Noncitizen	Noncitizen	0.00	0.00
Absent	Citizen	Absent	Noncitizen	0.00	0.00
<b>Total disagreements</b>				<b>3.43</b>	<b>2.68</b>

Notes: This uses the 2018 ACS voting-age sample and its sampling weights. The total number of observations is 3,989,000.

“With record linkage quality restrictions” indicates links are accepted only if the predicted Pr(link is correct)  $\geq$  .99; see slide #45. “D” indicates that the number is suppressed due to disclosure restrictions. USCIS (U.S. Customs and Immigration Service) citizens are from naturalization records; USCIS noncitizens are from lawful permanent resident, refugee, and asylee records. ITINs refers to personal tax identifiers in the range reserved for Individual Taxpayer Identification Numbers, which is public information. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Percent ACS Noncitizens That Are 2020 Business Rule Citizens, by ACS Interview Year



The number of observations is 2,156,000 without a record linkage threshold, and 1,273,000 with one.

“Record linkage threshold” indicates cases that satisfy the quality restrictions that links are accepted only if the predicted  $\Pr(\text{linkage is correct}) \geq .99$ ; see slide #45. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from Four Approaches 2010 CEF, NBR-SS Cases (5.8% of total pop) by observable characteristics

Characteristic	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
Householder Citizen	88.4	81.6	79.5	71.4	79.3
Householder Noncitizen	24.1	22.2	13.3	37.9	14.2
Difference	64.4	59.5	66.1	33.5	65.1
English Form	76.4	77.1	72.7	67.8	79.7
Non-English Form	30.0	37.8	7.6	23.4	13.8
Difference	46.5	39.9	65.1	44.4	65.9

NBR-SS is no business rules and sent to PVS search. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).

# Estimated Percent Citizens from Four Approaches 2010 CEF, NBR-SS Cases (5.8% of total pop) by observable characteristics (continued)

Characteristic	Hot Deck	BR logistic	ACS logistic	Latent Class Model	2010-2012 ACS
Non-Relative and HHLDR a Citizen	88.0	73.4	81.4	71.7	84.8
Non-Relative and HHLDR a Noncitizen	23.9	16.6	14.4	37.2	14.2
Difference	64.1	56.8	67.0	34.5	70.6

NBR-SS is no business rules and sent to PVS search. HHLDR is householder. The four estimation approaches (Hot Deck, BR logistic, ACS logistic, and Latent Class Model) are discussed on slides #20-26. CEF = Census Edited File. All original data presented in this presentation have passed Census Bureau Disclosure Review Board approval (CBDRB-FY20-CED006-0031).