

**Determination of the 2020 U.S. Citizen Voting Age Population (CVAP)
Using Administrative Records and Statistical Methodology**
Technical Report

by

**John M. Abowd
William R. Bell
J. David Brown
Michael B. Hawes
Misty L. Heggeness
Andrew D. Keller
Vincent T. Mule Jr.
Joseph L. Schafer
Matthew Spence
Lawrence Warren
Moises Yi**

U.S. Census Bureau

CES 20-33

October 30, 2020

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. 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. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K038E, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov. To subscribe to the series, please click [here](#).

Abstract

This report documents the efforts of the Census Bureau's Citizen Voting-Age Population (CVAP) Internal Expert Panel (IEP) and Technical Working Group (TWG) toward the use of multiple data sources to produce block-level statistics on the citizen voting-age population for use in enforcing the Voting Rights Act. It describes the administrative, survey, and census data sources used, and the four approaches developed for combining these data to produce CVAP estimates. It also discusses other aspects of the estimation process, including how records were linked across the multiple data sources, and the measures taken to protect the confidentiality of the data.

Keyword: citizenship, administrative records, voting-age population, big data

JEL Classification: J1, C1, C6, C8

[§] This work would not be possible without extensive support of U.S. Census leadership including: Director Steven Dillingham, Deputy Director Ron S. Jarmin, Senior Advisor Enrique Lamas, Associate Director Albert E. Fontenot, and Associate Director Victoria A. Velkoff. We are thankful to additional leaders and staff who contributed to regular meetings and provided data, suggestions, and ideas that contributed to this report. Those individuals include: Michael A. Berning, Stephanie J. Busick, Patrick J. Cantwell, Jennifer Hunter Childs, Sandra L. Clark, John L. Eltinge, Carolina Franco, Christa D. Jones, Darcy S. Morris, Roberto Ramirez, Damon R. Smith, Sara Sullivan, Evan S. Totty, James B. Treat, and Lori Zehr.

2020 Census Methods Internal Expert Panel

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

Citizen Voting Age Population (CVAP) Technical Working Group

William R. Bell, J. David Brown (lead), Stephanie Busick, Misty L. Heggeness, Ryan Janicki, Andrew D. Keller, Darcy S. Morris, Vincent T. Mule Jr., Joseph L. Schafer, Matthew Spence, Lawrence Warren, and Moises Yi

Citizen Voting Age Population (CVAP) Implementation Team

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

**Determination of the 2020 U.S. Citizen Voting Age Population (CVAP)
Using Administrative Records and Statistical Methodology**

Table of Contents

Abbreviations

Executive Summary

1. Introduction
2. Data Sources
3. Record Linkage
4. Business Rules
5. Four Approaches to Statistical Estimation of CVAP Modeled Cases using Multiple Sources
6. Hot Deck Nearest Neighbor Method
7. Householder Logistic Regression Method
8. American Community Survey Logistic Method
9. Latent Class Modeling
10. Empirical Results
11. Disclosure Avoidance
12. Recommendations

References

ABBREVIATIONS

ACS = American Community Survey

ADIS = United States Customs and Border Protection Arrivals and Departures Information System

AHS = American Housing Survey

AIAN = American Indian and Alaska Native

BOP = Federal Bureau of Prisons

BR = Business Rules

CBP= United States Customs and Border Patrol

CEF = Census Edited File

CPS = Current Population Survey

CUF = Census Unedited File

CVAP = Citizen Voting-Age Population

DOB = Date of Birth

DOC = United States Department of Commerce

DOJ = United States Department of Justice

DRB = Disclosure Review Board

DSEP = United States Census Bureau Data Stewardship Executive Policy Committee

EPIK = Enhanced Protected Identification Key

ERF = Enhanced Reference File

HU = Housing Unit

ICE = United States Immigration and Customs Enforcement

IEP = Internal Expert Panel

IMARS = United States Department of Interior Incident Management Analysis Reporting System

IRS = United States Internal Revenue Service

ITIN = Individual taxpayer identification number

LC = Latent Class

LCO = Local Census Office

LEMIS = Law Enforcement Management Information System

MDF = Microdata Detail File

NBR = No Business Rules

NCRP = National Corrections Reporting Program

NH = Non-Hispanic

NHOPI = Native Hawaiian or Other Pacific Islander

NSS = Not Sent to Search

NUMIDENT = Numerical Identification File

OMB = Office of Management and Budget

PII = Personally Identifiable Information

PIK = Protected Identification Key

PLB = Privacy-Loss Budget

PPM-PTS = Prisoner Processing and Population Management Prisoner Tracking System

PVS = Person Identification Validation System

SEVIS = Student Exchange Visitor Information System

SIPP = Survey of Income and Program Participation

SNAP = Supplemental Nutrition Assistance Program

SS = Sent to Search

SSA = Social Security Administration

SSN = Social Security Number

TANF = Temporary Assistance for Needy Families

TDA = TopDown Algorithm

TWG = Technical Working Group

USCIS = United States Customs and Immigration Services

USMS = United States Marshals Service

VRA = Voting Rights Act

WRAPS = United States Department of State Population, Refugees, and Migration Worldwide Refugee Admissions Processing System

Executive Summary

The U.S. Census Bureau's 2020 Census Methods Internal Expert Panel (IEP) was charged with recommending a method to produce Citizen Voting Age Population (CVAP) estimates at the block level by combining population data from the 2020 Census with citizenship data from various available sources, including administrative and survey sources. This is in line with (1) the Department of Commerce (DOC) Secretary Wilbur Ross's direction from March 26, 2018,¹ (2) the 2020 Census Office of Management and Budget (OMB) Paperwork Reduction Act Clearance Package of December 28, 2018,² and (3) the Presidential Executive Order of July 11, 2019 titled *Executive Order on Collecting Information about Citizenship Status in Connection with the Decennial Census*.³ In collaboration with the Census Bureau's Redistricting and Voting Rights Data Office, the IEP determined the content and format for the updated experimental CVAP data products. This defined the statistical estimand: the quantity that the Census Bureau's methods are trying to estimate.

The requirement of producing block-level CVAP estimates posed a new challenge that could not be satisfied by five-year ACS estimates as have been used for CVAP since 2011. ACS margins of error for very small geographic areas (tracts and below) are large. An analysis of the fitness-for-use of 2019 ACS CVAP estimates concluded that if the five-year estimates for the CVAP table were subjected to the ACS one-year data quality filtering rule, only 1,093 of 217,739 block-group tables could be released. Even apart from the large margins of error, many individual blocks would have no ACS sample observations.

On the other hand, the use case for block-level CVAP estimates is not geared to examining estimates for individual blocks, but rather to provide inputs to redistricting plans that can be aggregated into arbitrary geographic areas that cannot be pre-specified. Still, the availability of several large administrative data sources with information on citizenship raised the possibility of combining multiple data sources, including administrative records and surveys, to produce better estimates than could be produced solely from ACS data. This suggested the possibility of using the 2020 Census results as a population frame along with contemporaneous administrative records. Assigning or predicting citizenship status for the 2020 Census person records could be expected to yield substantial improvements to ACS CVAP estimates due to (i) a potentially enormous reduction of sampling error (if administrative records citizenship indicators could be assigned to a large share of the census records), and (ii) the potentially greater currency and detail of the 2020 Census counts and contemporaneous administrative records compared to the 2015-2019 ACS data.

¹ See the Administrative Record for the citizenship question litigation at <https://www2.census.gov/foia/records/citizenship-records/ar-final-filed-all-docs--certification-index-documents--060818.pdf>. See Bates numbers 1313-1320.

² "Accordingly, the Secretary has directed the Census Bureau to proceed with the 2020 Census without a citizenship question on the questionnaire, and rather to produce Citizenship Voting Age Population (CVAP) information prior to April 1, 2021 that states may use in redistricting." For more information, see OMB PRA 2020 Census Supporting Statement A (full revised final), submitted July 3, 2019, approved July 12, 2019 (<https://www.reginfo.gov/public/do/DownloadDocument?objectID=88197702>).

³ For more information, see: <https://www.whitehouse.gov/presidential-actions/executive-order-collecting-information-citizenship-status-connection-decennial-census/>.

The IEP met on a regular basis from July 2018 to the present, reviewing the efforts of a 2020 CVAP Technical Working Group, which was developed to exhaust all viable options for CVAP production at the block level with the 2020 Census and administrative data. The working group explored four alternative approaches for using multisource data in the production of CVAP statistics. Three of these approaches started with “business rules” for using the citizenship data sources to assign citizenship data to census records. Two experiments, one using 2010 Census data and the other using 2018 American Community Survey (ACS) data, combining these data with corresponding administrative and survey sources appropriate for the two years, found the business rules (BR) could reliably assign citizenship to just over 90% of the population, leaving just under 10% of cases for whom citizenship status required statistical estimation.

The three approaches pursued to augment BRs with statistical estimation were (i) Hot Deck method that imputes citizenship status of the non-BR (NBR) cases using donors from the BR cases, (ii) BR logistic method that predicts probabilities of citizenship status for the NBR cases using logistic regression models fitted to the BR cases, and (iii) ACS logistic method that predicts probabilities of citizenship for the NBR cases using logistic regression models fitted to ACS records that could not be given BR citizenship assignments, but that did have citizenship reported to ACS. By developing predictors of citizenship probabilities for the census NBR cases based on data from the ACS NBR cases, the ACS Logistic approach seeks to address potential bias that could arise for the Hot Deck and BR Logistic approaches should their assumption that the BR cases are like the NBR cases fail. This is a type of non-ignorable missing data problem.

The working group also explored a fourth approach, latent class (LC) modeling, that uses a multivariate model to combine information from multiple citizenship data sources to produce predicted probabilities of citizenship for all person records. Despite not using explicit business rules, the LC modeling produced citizenship estimates for the BR cases that were very close to those from the BR assignments, providing strong confirmation for the BRs. While the LC modeling has some advantages compared to the other three approaches, certain effects found in the logistic regression modeling for detailed population subgroups could not be fully replicated in the LC model without enhancements to the model that require innovative enhancements to the computer software. While intensive work has been done on these enhancements, they are not complete as of this writing, and this work is ongoing.

Summary of Results on Fitness for Use of the Citizenship Data Sources

Primary administrative data sources on citizenship obtained by the CVAP implementation team for use by the CVAP Technical Working Group included the following:

- Social Security Administration (SSA) Numerical Identification File (NUMIDENT)
- U.S. State Department passport data
- U.S. Citizenship and Immigration Services (USCIS) naturalizations and lawful permanent residents data
- Individual Taxpayer Identification Numbers (ITINs)
- U.S. Customs and Border Protection Arrivals and Departures Information System (ADIS) data

- U.S. Immigration and Customs Enforcement (ICE) Student Exchange Visitor Information System (SEVIS) data
- U.S State Department Worldwide Refugee Admissions Processing System (WRAPS) data
- U.S. Department of Interior Incident Management Analysis Reporting System (IMARS) and Law Enforcement Management Information System (LEMIS) data

Secondary administrative data sources obtained included Federal Bureau of Prisons (BOP) data, U.S. Marshals Service (USMS) data, Bureau of Justice Statistics (BJS) - National Corrections Reporting Program (NCRP) data, driver's license data from a few states, and Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) files for eight states. Census Bureau surveys that provided citizenship data included the ACS, the Current Population Survey (CPS), the American Housing Survey (AHS), and the Survey of Income and Program Participation (SIPP). Records from these data sources for the appropriate years were linked to the 2010 Census records and the 2018 ACS records via the assignment of Protected Identification Keys (PIKs) from the Person Identification Validation System (PVS).⁴ These linkages permitted assignment of citizenship status available from a citizenship data source record to a 2010 Census record or 2018 ACS record. This process required that we pay attention to distinguishing high-quality assignments of citizenship status from low-quality assignments that could arise from record linkage error or from incorrect or out-of-date records in the citizenship data source.

Population coverage of the citizenship data sources was analyzed using the 2018 ACS as the frame. This revealed the following results:

- Overall coverage of the citizenship data sources was estimated at about 91%, accounting for the considerable overlap among the sources and eliminating citizenship indicators of questionable quality (e.g., out of date records) or with poor record links. This leaves about 9% of the records requiring statistical estimation of citizenship.
- The primary administrative sources showed strong consistency, with disagreements for only 3.4% of the 2018 ACS population. The disagreements were almost all explainable, generally arising when an indicator of citizenship from one record conflicted with an outdated indicator of noncitizen from another record. A common example was an ACS record that linked to both a NUMIDENT noncitizen record and to a more recent U.S. passport or USCIS naturalization certificate.
- BRs constructed from the data sources favored (i) sources with the greatest coverage, (ii) sources for which citizenship status is determined based on person documentation supplied (in contrast to, say, survey reports), and (iii) citizen indications over noncitizen indications, due to the possibility of noncitizens having naturalized to citizens without their noncitizen record being updated. The first steps in the BRs are shown in Table ES.

⁴ Additional PIKs, referred to as EPIKs, were assigned using an enhanced reference file in an attempt to link more administrative and survey records to the 2010 Census and 2018 ACS records. The increase in coverage of the 2018 ACS estimated population from using EPIK citizenship information was only about 0.11%. See Section 3.2.

Table ES Business Rule Citizenship Assignments from Administrative Record Primary Sources

Primary criteria for assigning as citizen	Rule assignment	Percent of 2018 ACS Population
NUMIDENT citizen	Citizen	72.43
NUMIDENT missing citizenship and U.S.-born	Citizen	9.15
U.S. passport	Citizen	3.01
USCIS naturalization certificate	Citizen	0.29
If not U.S. citizen according to any of the above criteria:		
NUMIDENT noncitizen	Noncitizen	5.27
ITIN	Noncitizen	0.52
⋮	⋮	⋮

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

The six rules shown in Table ES cover 90.7% of the population, with 87% coming from the NUMIDENT. Each of the additional data sources not shown above provide additional coverage for less than 0.1% of the population, and only a few provide more than 0.01%.

- Comparing BR determinations for 2018 ACS sample cases with ACS estimates obtained from the sample cases with both BR determinations and ACS citizenship responses reveals important facts about both. The ACS estimated percent of citizens for these BR cases was 93.69%, compared to 93.56% of these cases determined as citizens by the BRs. The closeness of these estimates masks ACS misreporting error, however. Of the BR determined citizens, ACS had just 0.71% misreported as noncitizens. For the BR determined noncitizens, however, ACS had 12.21% misreported as citizens. The much larger number of citizens compared to noncitizens in the population produced very near cancellation of these asymmetric reporting errors. One cannot assume such fortuitous cancellation of reporting errors will occur for every population subgroup, though, whether defined geographically or by demographic characteristics.

Summary of Results Comparing Citizenship Estimates from the Alternative Approaches

Citizenship estimation for the three approaches using BRs started by tabulating citizen counts from the cases with BR determinations, and then added to these results the predicted number of citizens (by imputation or by logistic regression modeling) from the NBR cases, which comprised about 9% of the population. For this purpose, the NBR cases were broken down into three groups, yielding a breakdown of the total population into four groups:

- BR: cases that received a PIK and citizenship assigned by the BRs (90.8% of the 2010 Census Edited File (CEF) population)
- NBR-PIK: cases that received a PIK but for which no BR citizenship was determined (0.125% of the 2010 CEF population)
- NBR-SS: cases with sufficient personally identifiable information (PII) that they were sent to search for a PIK, though none was obtained, and hence no BR citizenship could be assigned (5.8% of the 2010 CEF population)

- NBR-NSS: cases without sufficient PII so they were not sent to search for a PIK and obtained no BR citizenship (3.3% of the 2010 CEF population)

These four groups have very different citizen percentages, and different amounts of available information (NBR-NSS has the least), making it appropriate to consider them separately. Summarizing the main results using the 2010 CEF as the frame:

- Though the three BR-plus approaches used slightly different versions of the BRs, the resulting BR citizen percentages were very nearly the same; all were 92.5% or 92.6% citizens. The LC approach also produced 92.5% citizens, despite not explicitly using BRs. The estimate for BR cases from 2010-2012 ACS data was close at 93.0%.
- Estimates for the NBR-PIK group varied widely across the four approaches, but the group is so small that these differences have no appreciable effect on estimates for the full population.
- Estimates for the NBR-SS group varied substantially across the four approaches for Hispanics and non-Hispanic (NH) Asians; estimates for NH Whites and NH Blacks varied much less. For Hispanics, the ACS logistic approach produced the lowest estimate of 30.0% citizens, which was not far from the 2010-2012 ACS direct estimate of 33.4%. The BR logistic approach gave the highest estimate of citizens at 48.6%, and the hot deck approach was not far away at 42.0%. These higher estimates for the two approaches that develop their predictors for the NBR-SS group using the BR data as a training sample are consistent with the non-ignorable missing data concern. Estimates for NH Asians in the NBR-SS group showed a similar pattern to that for Hispanics, though the levels of the estimates were higher (e.g., 41.2% for ACS logistic).
- An experiment applied the ACS NBR-SS logistic model in several alternative ways defined by whether the ACS NBR-SS cases or BR cases were used as the training sample, and whether ACS reported citizenship or BR citizenship assignments were used as the dependent variable in the prediction. This exercise found that what made the largest difference in the results, setting substantially apart the ACS logistic estimates for Hispanics and NH Asians, was whether the ACS NBR cases or BR cases were used as the training sample. This suggested that the ACS logistic approach's use of ACS NBR cases was addressing some of the non-ignorable missing data issue.
- Estimates for the NBR-NSS group varied across the approaches, though not as much as those for the NBR-SS group, and with a different pattern. The hot deck approach had the lowest estimates for the four largest race/ethnicity groups, particularly for Hispanics and NH Asians (at about 54.6%). The ACS logistic was second lowest for Hispanics (61.7%), while the BR logistic was second lowest for NH Asians (62.8%).

Estimates using the 2018 ACS as the frame⁵ showed slightly higher citizen percentages overall, with similar patterns to those from the 2010 CEF frame for BR cases and for NBR-SS cases for NH Whites and NH Blacks, with little difference across approaches. For NBR-SS Hispanics the ACS logistic estimates were lowest among the three BR approaches, though the direct ACS estimate was even lower. Results for NH Asians did not vary much across the approaches. Results for the NBR-PIK and NBR-NSS groups may not be comparable to those from estimates using the

⁵ Results with the 2018 ACS as the frame were not obtained for the LC approach.

2010 CEF frame because with the reduced size of the 2018 ACS frame compared to the 2010 CEF, these groups were very small.

Recommendations

Based on the Citizen Voting-Age Population (CVAP) Technical Working Group's evaluations of the data sources, empirical results from the four estimation approaches, and CVAP production considerations, the U.S. Census Bureau's 2020 Census Methods Internal Expert Panel (IEP) makes the following recommendations.

1. The IEP believes that the BRs used for citizenship assignment, which differed slightly across the approaches, can provide accurate citizenship estimates for the census cases that can be reliably linked to the administrative and survey data sources. In the experiments done, differences in these results across the three approaches were minor. The IEP thus recommends proceeding by developing a single harmonized set of BRs as follows:
 - a. Persons are classified as citizens if they are citizens in the Social Security Administration (SSA) NUMIDENT file, have a U.S. passport or USCIS naturalization certificate, or do not have SSA NUMIDENT citizenship but are U.S.-born in those data.
 - b. Persons lacking that information are deemed noncitizens if noncitizens in the SSA NUMIDENT, SEVIS, WRAPS, IMARS, LEMIS, BOP, USMS, driver's license data, NCRP, ACS, AHS, CPS, or SIPP; have a nine-digit taxpayer ID number in the ITIN range; noncitizens in USCIS data with better record linkage quality; or noncitizens in ADIS with better linkage quality and more recent vintage.
 - c. If none of the above apply, then persons are treated as citizens if they are citizens in ADIS, BOP, USMS, driver's licenses, SNAP/TANF, NCRP, ACS, AHS, CPS, or SIPP.

Statistical estimation will be required to estimate citizenship for the cases not covered by the BRs.

2. The American Community Survey (ACS) logistic method is the preferred method for the production of the 2020 CVAP experimental data products, subject to the caveat listed in 2.b. below.
 - a. The IEP believes that this method best addresses the non-ignorable missing data issue that arises when BR cases of linkable citizenship information are used to develop predictors of citizenship probabilities for the NBR cases. By training models on ACS records that also lack linkable citizenship information, but have as-reported ACS citizenship responses, the ACS logistic method helps address this issue, especially for those cases with sufficient PII to be sent to search for a PIK.
 - b. The evidence about non-ignorable missingness is less strong for the NBR cases with insufficient PII to be sent to search for a PIK, and the IEP recommends that the CVAP Technical Working Group perform further study of these cases. Another reason for further study is the possibility that the size of this group in the 2020 Census could be larger than was the case in the 2010 Census (when it was about 3.3% of the population). The IEP recommends that the CVAP Technical Working

Group investigate enhancements to the use of logistic regression with either the BR cases or ACS cases and perform further evaluations of the results. A final recommendation on treatment of these cases will be made following this additional investigation.

- c. For the cases that received a PIK, but for which no citizenship status could be assigned, the estimates differed across the alternative approaches. However, the IEP recognizes that this is a very small group of records, with little impact on the overall estimates, and with no clear reason to expect significant growth of this group in 2020. Therefore, the IEP recommends that this group be combined with one of the other two NBR groups, based on an assessment of evidence of non-ignorability in this small population.
3. The IEP recognizes that LC modeling is a promising approach for producing CVAP estimates. However, given the need to enhance the software to accommodate LC models with the desired detail, and the fact that this enhanced software is still under development, the IEP recommends that the LC approach be used for evaluation, via comparisons made to the results from the recommended approaches, and not for the CVAP production at this time. The LC approach should also be examined for its ability to produce uncertainty measures (standard errors) for citizenship estimates.
 4. Any newly received citizenship data sources not covered by the tests in this report should be evaluated for use based on the same methods applied to the sources that are included in this report.
 5. The Census Bureau should continue to enhance and develop improved record linkage for the production of official statistics using multisource data, including the production of enhanced CVAP statistics.
 - a. The PVS reference files should be expanded to include records in government sources that have sufficient PII, but have not received a PIK when attempting to link to the current production PVS reference files. This facilitates linkage for individuals without SSNs or nine-digit taxpayer IDs in the ITIN range.
 - b. Record linkage quality measures derived from PVS module, pass, and score information should be used when evaluating a linked records' fitness for use.

1. Introduction⁶

This report documents the efforts of the Census Bureau’s Citizen Voting-Age Population (CVAP) Internal Expert Panel (IEP) and Technical Working Group (TWG) towards the use of multiple data sources to produce block-level statistics on the citizen voting-age population for use in enforcing the Voting Rights Act. It describes the administrative, survey, and census data sources used, and the four approaches developed for combining these data to produce CVAP estimates. It also discusses other aspects of the estimation process, including how records were linked across the multiple data sources, and the measures taken to protect the confidentiality of the data. It begins with a brief background discussion.

In December 2017, the U.S. Department of Justice (DOJ) sent a letter to the Census Bureau’s Deputy Director and Chief Operating Officer, Ron S. Jarmin, who at the time was performing the nonexclusive duties and functions of the Census Bureau Director, requesting the addition of a citizenship question to the 2020 Census enumeration form for the development of block-level CVAP statistics for use in enforcing the Voting Rights Act. In response to this request, John Abowd, the Census Bureau’s Chief Scientist, sent a January 19, 2018 memo to Department of Commerce (DOC) Secretary Wilbur Ross reviewing three alternatives for producing block-level CVAP statistics: “(A) no change in data collection⁷, (B) adding a citizenship question to the 2020 Census, and (C) obtaining citizenship status from administrative records for the whole 2020 Census population.” At the Secretary’s request, the Census Bureau then performed a preliminary analysis of combining Alternatives B and C into a new Alternative D that would use both citizenship responses from the census and administrative records on citizenship.⁸ On March 26, 2018, Secretary Ross instructed the Census Bureau to address the DOJ request by both adding the citizenship question to the 2020 Census and obtaining additional federal and state administrative records.

To inform the decision on how best to produce block-level CVAP statistics, Dr. Abowd directed Census Bureau staff to analyze the quality of various citizenship data sources, including household surveys and administrative records, and to study the effect of adding a sensitive question like citizenship to a Census enumeration form. Results of these investigations are reported in Brown et al. (2019a, b). Preliminary results from the investigations informed the January 19 and March 1, 2018 memos to the Secretary of Commerce.

Following the Secretary’s March 26, 2018 decision, various court cases were brought forward seeking to block the addition of a citizenship question to the 2020 Census. These cases worked

⁶ The data in the main body of this document are released under the following disclosure review numbers: CBDRB-FY21-CED002-B0001 and CBDRB-FY21-CED005-0001.

⁷ Under this option, the CVAP statistics would have been produced as in recent years using data from the Census Bureau’s American Community Survey.

⁸ The memo describing this analysis, dated March 1, 2018 from John M. Abowd to Secretary Ross through Ron S. Jarmin, was never published by the Census Bureau but was released as part of the administrative record in *New York et al. v. Department of Commerce* 2018 <https://www2.census.gov/foia/records/citizenship-records/ar-final-filed-all-docs--certification-index-documents--060818.pdf>. See Bates numbers 1308-1312.

their way through the courts, with a consolidated case (Department of Commerce et al. v. New York et al.) reaching the U.S. Supreme Court. On June 27, 2019, the Supreme Court issued a ruling holding, in agreement with the District Court’s ruling, that “the decision to reinstate a citizenship question cannot adequately be explained in terms of DOJ’s request for improved citizenship data to better enforce the VRA,” and the case was remanded. In view of the pressing production schedule for the 2020 Census, the Attorney General and the Secretary of Commerce concluded that the only practical alternative was to remove the citizenship question from the census. President Donald Trump subsequently issued an executive order on July 11, 2019 instructing all executive departments and agencies to provide to the Census Bureau access to any of their administrative records that could be useful for producing the CVAP estimates, to the maximum extent permissible under law. The order also directed that, consistent with law, efforts be strengthened to obtain state administrative records concerning citizenship.

In fact, the Census Bureau had, for some time, been working with other government agencies to establish agreements that would permit and then achieve the transfer to Census of administrative records data sources containing information on citizenship, consistent with the proposed alternatives C and D discussed above. The decision not to include the citizenship question on the census form increased the importance of this effort, which was then assisted by the executive order.

In September 2018, the CVAP Technical Working Group (TWG) was formed under the direction of Dr. Abowd and the Census Bureau’s 2020 Census Methods Internal Expert Panel (IEP). The CVAP Technical Working Group’s charge was to develop recommendations for the production of official statistics related to citizenship status and the citizen voting-age population using multi-source data. The TWG examined the available data sources on citizenship and identified others that could potentially be acquired. It also developed two general approaches that could be used to combine multiple data sources to produce the CVAP estimates. One approach, latent class modeling, uses a statistical model that assumes an unobserved true status (citizen or noncitizen) for every individual in the census and treats the data sources modeled as providing indicators of the true status for each person that are subject to errors. After fitting the model to the data, it can be used to form predictions of the true citizenship status of individuals, which resolves those cases for which multiple data sources supply conflicting information. These results can then be tabulated to produce the desired citizenship statistics.

The other general approaches used “business rules” to directly assign citizenship status from the data sources to individuals in the census, under rules established to choose a preferred assignment for cases where multiple data sources are in conflict. The TWG examined, for illustration, two versions of the business rules approach: one that favored survey responses over administrative sources, and one that favored administrative sources over survey responses. A feasibility study then applied these business rules approaches, and the latent class modeling approach, to data for 2010 from the ACS, the Social Security NUMIDENT file, and a file of Individual Taxpayer Identification Numbers (ITINs). The ACS data served as a stand-in for census data, since the 2010 Census did not include a long form that collected detailed information including citizenship. (It should be noted that while these efforts developed and illustrated the two approaches for

combining census and administrative records data on citizenship, these approaches can also be applied to the case with no census citizenship question.) These efforts, including results of the feasibility study, are documented in the group's July 2, 2019 internal report (CVAP Working Group 2019).

The purposes of this report are to: describe the expanded set of administrative data sources now being used; discuss three improved versions of the business rules approach that have been investigated, and an enhanced version of the latent class model; and illustrate these approaches on data from earlier years. Given the Supreme Court ruling that led to removal of the citizenship question from the 2020 Census, these approaches are presented and illustrated in the setting of a census with no citizenship question.

Section 2 discusses the data sources being used: administrative records sources, survey data sources, and census data. Section 3 discusses the record linkage techniques that will be used for combining data sources. Section 4 discusses the business rules used for the citizenship data sources, including their coverages of the population, and presents results comparing the sources that reflect their strengths and limitations. Section 5 briefly discusses four approaches (discussed in more detail in Sections 6 – 9) that we developed to go beyond business rules to produce CVAP estimates for the full population. Sections 6 – 8 discuss improved versions of the business rules approach, and Section 9 discusses the enhanced latent class modeling approach. Section 10 describes empirical results illustrating the four approaches on data for 2010 (making use of 2010 Census population data) and 2018 (with ACS standing in for the census). Section 11 describes how the estimation results will be tabulated to produce the required CVAP estimates within disclosure avoidance procedures that will be applied to guarantee the confidentiality of the records from all the data sources used. Finally, Section 12 provides recommendations for an appropriate method to generate enhanced CVAP statistics that comply with the July 11, 2019 Executive Order.⁹

⁹ For more details on the Executive Order, see: <https://www.whitehouse.gov/presidential-actions/executive-order-collecting-information-citizenship-status-connection-decennial-census/>.

2. Data Description

This section describes all the citizenship data sources used in this report.¹⁰ Descriptions are given of the type of data included in each source and a general overview of how they contribute to our knowledge of citizenship for individuals within the core frames of the data used to replicate statistics on citizenship in this report (primarily the 2010 Census Edited File and the 2018 American Community Survey). More attention is given to describing the core input sources of data used in our analysis and, where appropriate, we explain the limitations of each dataset.

In this section, we divide the administrative data into primary and secondary sources. Primary sources provide proof of citizenship status that other sources may use as evidence, while the secondary sources are derivative information. We use exactly the same criteria when testing how to use each primary, secondary, and survey source.

2.1 Administrative Data – Primary Sources

2.1.1. Social Security Administration

The Social Security Administration (SSA) Numerical Identification File (NUMIDENT) is a record of individual applications for Social Security cards and certain subsequent transactions for those individuals. Examples of data elements on a NUMIDENT record include name, date and place of birth, parents' names, and date of death. Unique, life-long Social Security numbers (SSNs) are assigned to individuals based on these applications. In addition, a full record of all changes to the account information (such as change of name) is also maintained. To obtain an SSN, the applicant must provide documented identifying information to SSA. Through the “enumeration at birth” program,¹¹ children born in the United States are issued an SSN upon birth. All U.S. citizens are eligible to have an SSN, which is required to work, to receive Social Security benefits and to receive other federal-government-administered social services. The NUMIDENT provides the most comprehensive coverage of U.S.-born persons. NUMIDENT coverage of noncitizens is less complete, as generally only those authorized by the Department of Homeland Security to work are eligible for an SSN.

SSA began requiring documentary evidence of citizenship for some categories of persons in 1974, and all applicants were required to provide it starting in 1978.¹² SSA began recording citizenship information in the NUMIDENT starting in May 1981. For persons who applied for an SSN prior to May 1981 and who have not subsequently updated their SSN record, the only available citizenship-related information is place of birth and an associated foreign country of birth indicator for those not born in the United States.

Occasionally the country of birth code is entered in the state of birth field by mistake in SSN applications. The foreign country of birth indicator is calculated from the country of birth field.

¹⁰ We describe every source currently available to the Census Bureau. All sources include 2018 data, which is the base year used in testing in Section 4.

¹¹ For more information, see <https://secure.ssa.gov/poms.nsf/lnx/0110205505>.

¹² See <https://secure.ssa.gov/apps10/poms.nsf/lnx/0110210001> for a timeline of these requirements, and <https://secure.ssa.gov/apps10/poms.nsf/lnx/0110210500> and its links for a description of acceptable documentary evidence.

Since the country of birth is missing for these mistaken entries, the foreign country of birth indicator is blank (i.e., indicating not foreign-born). The NUMIDENT collapses the country and state of birth fields into a single variable, and the foreign country of birth indicator. The foreign country of birth indicator is occasionally miscoded as blank for persons born in countries with two-letter codes that are the same as a U.S. state (e.g., “CA” is California and Canada, “IN” is Indiana and India, and “MO” is Missouri and Morocco). To address this issue, among records with a blank foreign country indicator, we calculate the share of persons who are ever NUMIDENT noncitizens within each city and state/country combination, among records which have a blank foreign country indicator. If the foreign country indicator were mistake-free, we would expect very few persons who are ever NUMIDENT noncitizens to have a blank foreign country indicator.¹³ We create a new foreign-born indicator that is missing if the ever-NUMIDENT-noncitizen share for that person’s city and state/country combination is five percent or higher, and otherwise is one if the NUMIDENT foreign-born indicator is present and zero if it is not present. This recode effectively sets to missing the foreign-born status of persons in the combinations of city and state/country codes where the foreign-born coding ambiguity is salient.

SSA is not automatically notified when previously noncitizen SSN holders become naturalized citizens, so naturalizations may be captured with a delay or not at all. Consistent with this, Brown et al. (2019a) show that NUMIDENT citizenship changes usually occur sometime after the year of naturalization reported in the 2016 ACS (described in Section 2.2), and 38 percent of NUMIDENT noncitizens are reported as citizens in the ACS. To change citizenship status on an individual’s SSN card, naturalized citizens must apply for a new card, showing proof of the naturalization (U.S. passport or certificate of naturalization).¹⁴ Naturalized citizens wishing to work have an incentive to apply for a new card showing their U.S. citizenship, because noncitizen work permits expire, and the NUMIDENT is used in combination with U.S. Citizenship and Immigration Services (USCIS) data in the E-Verify program that confirms that job applicants are eligible to work.

2.1.2. U.S. Department of State Passport

The U.S. passport data, provided by the Department of State, cover all U.S. passports issued between 1978 and April 1, 2020). According to the USCIS, U.S. passports provide the most definitive proof of citizenship.¹⁵ Though all citizens are eligible for a passport, many do not have one. The absence of a passport is thus weak evidence that a person is not a citizen (see Section 9).

2.1.3. U.S. Citizenship and Immigration Services

USCIS naturalizations and lawful permanent residents’ data include all available electronic records for approved applications for naturalization and lawful permanent residence, as well as asylum and refugee data, through April 1, 2020. Records for individuals protected by 8 U.S.C. § 1367 (victims of trafficking, criminal activity, or benefit under the Violence against Women Act) are

¹³ U.S.-born children whose parents are foreign diplomats are not automatically given U.S. citizenship, but this is a tiny fraction of all U.S.-born persons.

¹⁴ For more information, see <https://www.ssa.gov/ssnumber/ss5doc.htm>.

¹⁵ See <https://www.uscis.gov/us-citizenship/proof-us-citizenship-and-identification-when-applying-a-job>.

excluded. Most naturalized persons acquire a USCIS naturalization certificate, with the exception of children under 18 who are automatically naturalized when a parent is naturalized. A child's derived naturalization could be documented either through a naturalization certificate or a passport, and the latter is less expensive. Since we have both USCIS and passport data, the data cover both alternatives.

2.1.4. Individual Taxpayer Identification Numbers

We created indicators for records with a nine-digit personal tax identifier in the range reserved for ITINs, which is public information.¹⁶ ITINs are issued to persons who need to interact with the Internal Revenue Service, but who are not eligible to have a Social Security Number (SSN). Since all citizens are eligible to have an SSN, ITIN-holders must be noncitizens during the time they use the ITIN. The IRS requires that ITIN applicants provide documentation of their connection to a foreign country, e.g., via a foreign passport or U.S. visa.¹⁷ ITINs that haven't been included in a federal tax return for three consecutive years expire. The ITIN can then be reassigned to another person. The record linkage system (described in Section 3) does not attempt to follow people when they transition from having an ITIN to having an SSN. Linkage of a 2020 Census record to the expired ITIN for a person who has been naturalized instead of the subsequently-issued SSN would introduce error in the citizenship status for that person.

2.1.5. U.S. Customs and Border Protection Arrivals and Departures Information System

The Customs and Border Protection Arrivals and Departures Information System (ADIS) data contain arrival and departure transactions at U.S. border crossings between January 1, 2013 and June 2020. The records cover persons with nonimmigrant visas who are lawfully in the United States within the terms of their admission, as well as in-country visa overstays. Those protected by 8 U.S.C. § 1367—applicants and/or recipients of T nonimmigrant status (victims of trafficking), U nonimmigrant status (victims of criminal activity) or of benefits under the Violence against Women Act—are excluded.¹⁸ The U.S. address is self-reported, and the data are not updated as the individual moves within the U.S.

2.1.6. U.S. Immigration and Customs Enforcement Office of Student and Exchange Visitor Program

The Immigration and Customs Enforcement (ICE) Office of Student and Exchange Visitor Program provides Student Exchange Visitor Information System (SEVIS) data from 2013 to April 1, 2020. The data cover student and exchange visitor visas, as well as those of their dependents. The data are not updated as individuals move and change immigration status.

2.1.7. Department of State Population, Refugees, and Migration

The Department of State Population, Refugees, and Migration provides Worldwide Refugee Admissions Processing System (WRAPS) data, which contains all records available for

¹⁶ In subsequent sections of this report we refer to these indicator variables as ITINs.

¹⁷ See <https://www.irs.gov/instructions/iw7>.

¹⁸ Those excluded are applicants and/or recipients of T nonimmigrant status (victims of trafficking), U nonimmigrant status (victims of criminal activity) or of benefits under the Violence against Women Act.

individuals arriving through the U.S. Refugee Admissions Program between January 1, 2013 and July 31, 2020. The address field is based on where the person lived 90 days after arrival in the country. The records are not updated as the person moves and changes immigration status.

2.1.8. Department of Interior

The Department of Interior Incident Management Analysis Reporting System (IMARS) and Law Enforcement Management Information System (LEMIS) data cover persons arrested by U.S. Immigration and Customs Enforcement (ICE) on Department of Interior land through September 2019. The data contain the date of arrest. ICE believes the persons are noncitizens, as these arrests are based in part on immigration violations.

2.2 Administrative Data – Secondary Sources

2.2.1. Federal Bureau of Prisons

The Federal Bureau of Prisons (BOP) data contain records of inmates in federal prisons from 1980 to April 1, 2020. The data include the date the inmate was last in federal prison. The BOP receives its citizenship status information from ICE.

2.2.2. U.S. Marshals Service

The U.S. Marshals Service (USMS) data include records for federal prisoners in the custody of USMS. The data come from the USMS Prisoner Processing and Population Management/Prisoner Tracking System (PPM-PTS), and they cover 2010 to April 1, 2020.

2.2.3. National Corrections Reporting Program

The National Corrections Reporting Program (NCRP) data from the Bureau of Justice Statistics in the U.S. Department of Justice contain records for prisoners in state custody in several states in 2018.

2.2.4. State Driver's Licenses

The Nebraska Department of Motor Vehicles' Driver's License administrative data contains the most recent driver's license, identification card, or junior driver's license for each individual. The file used for this report contains data through March 2020. The Department of Motor Vehicles verifies applicants' citizenship status. The citizenship variable is either "Y" for citizen or blank. The blanks include persons verified as noncitizens, as well as those for whom citizenship status was not verified. It is thus possible that some citizens could have blank values for citizenship.

The South Dakota Department of Motor Vehicle file contains driver's licenses and state ID's issued or updated between January 1, 2018 and May 1, 2020. Proof of citizenship and immigration status is required in the application process.

2.2.5. State Public Assistance Programs

Colorado, Idaho, Kentucky, Mississippi, New York, South Carolina, South Dakota, and Wyoming Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF) files contain a citizenship variable. We also use program denial codes from

Connecticut, Hawaii, Indiana, Maryland, and Oregon. Some of the reasons for denial are due to not meeting immigration status requirements. State agencies administering these programs verify citizenship and immigration status. How strictly immigration status rules are enforced may vary across agencies. Some of these sources are updated through 2019, and others are not.

2.3 Census Bureau Household Survey Sources

The Census Bureau currently conducts four surveys that ask citizenship questions, including the American Community Survey (ACS), the Current Population Survey (CPS), the American Housing Survey (AHS), and the Survey of Income and Program Participation (SIPP). The universe for citizenship questions on these surveys is all persons living in the household. The ACS, CPS, SIPP, and AHS distinguish between citizens born in the United States, those born in U.S. territories, those born abroad to U.S. citizen parents, and those of foreign nativity but naturalized. Additionally, the SIPP asks about more nuanced naturalizations, including becoming a citizen through one's own or a spouse's military service or via adoption by U.S. citizen parents.

Survey responses are not verified, and they are current as of the time of the interview, which in some cases took place many years ago.

2.4 Census Unedited File (CUF) and Census Edited File (CEF)

Responses from the 2020 Census will provide the definitive list of persons and their characteristics such as location, age, race, and Hispanic origin that are cross-tabulated with citizenship to create the CVAP tables. Two files from the decennial response processing operations will be used: the Census Unedited File (CUF) and the Census Edited File (CEF).

The CUF is available approximately two months before the CEF. Both files contain the finalized list of persons with unique person identifiers, but the demographic variables on the CUF include missing values due to item non-response and count imputation, and they are not edited for consistency.

The CEF contains edited demographic data and has no missing values. The edit and characteristic imputation procedure that creates the CEF from the CUF adjusts individual-level demographic information for consistency and fills in missing values either from administrative records sources or from nearest-neighbor hot deck imputation. The CEF is the final person-level file before disclosure avoidance is applied and general decennial tabulation occurs.

3. Record Linkage

Record linkage is central to producing statistics using multiple data sources as inputs. It plays a particularly large role in the work conducted for this report because we expect to link the 2020 Census to multiple administrative record and survey data sources. In this section, we describe the Census Bureau’s production process for assigning unique person identifiers to data records. We then describe how we enhanced this process to expand the universe of persons receiving a unique person identifier. These enhancements enable us to link administrative record data on citizenship to a larger share of the U.S. population. We discuss errors that can arise in the record linkage process, and how they may produce errors in estimates of citizenship. Finally, we construct a record linkage quality proxy that informs the modeling and business rules in subsequent sections.¹⁹

3.1 The Current Record Linkage System

Protected Identification Keys (PIKs) are assigned by the Census Bureau’s Person Identification Validation System (PVS).²⁰ The PVS discussed in this report is the current production version, which is the same version that is used for administrative record linkage in the 2020 Census production environment for the CUF and CEF. PIKs are anonymous unique person identifiers that are temporally invariant just like Social Security Numbers (SSN). SSNs are replaced by PIKs in files that initially contain SSNs when received by the Census Bureau, since access to files containing SSNs is limited to a small staff that specializes in maintaining the record linkage system. This process facilitates linking person records across files while protecting individual PII from distribution within the Census Bureau.

The PVS uses probabilistic record linkage (Fellegi and Sunter, 1969) to match data from an incoming file (e.g., a survey or administrative record file) to reference files containing data from SSA enhanced with address data obtained from other federal administrative record files.²¹ Reference files contain all variants of a person’s name, date of birth, and sex, as well as current and recent addresses. The standard PVS methodology consists of an initial edit procedure to clean and standardize the linking fields (name, date of birth, sex, and address), followed by a cascading matching process involving several modules that are described below including: Verification, GeoSearch, NameSearch, DOB (date of birth) Search, and Household composition. Records failing to link within a module proceed to the next module in sequential fashion.

Because it is not feasible to compare all records from a given input file to all records in the reference file, comparisons are restricted to records that agree on certain characteristics, a process called *blocking*. *Blocking* works as follows. The data are split into blocks based on exact matches of certain fields or parts of a field. Then, probabilistic record linkage is performed within each block. A typical blocking strategy gives rise to a series of ‘passes’ within each module. It starts with a restrictive pass where the records have to agree on a very constrained set of characteristics

¹⁹ In this report business rules are a set of criteria specifying how to classify a person’s citizenship status for each combination of citizenship information across sources.

²⁰ For details see Wagner and Layne (2014).

²¹ Individual Taxpayer Identification Numbers (ITINs) and the names and addresses associated with them in federal administrative records are also included; however, the Census Bureau does not have access to the ITIN application data. See section 2 for a description of ITINs.

(e.g., address including apartment number), then broadens the blocking universe (e.g., to street or 5-digit ZIP code) for subsequent passes. Below we describe each of the PVS modules.

Verification Module: If the input file has SSNs, the verification module checks for an exact SSN match to the reference files and verifies that the name and date of birth elements sufficiently agree. If they do, the SSN is considered verified and PVS assigns the corresponding PIK to the person record on the input file.

GeoSearch Module: Records not assigned a PIK in the verification module are sent to the GeoSearch module. This module blocks on various levels of address and zip code information and attempts to find matches, typically based on name, date of birth, and sex.

Name Search Module: This module uses only the name and date of birth fields in the search process, and it includes all possible combinations of alternate name and dates of birth for a given SSN.

Date of Birth (DOB) Search Module: In this module, the reference files are blocked based on month and day of birth prior to matching attempts. This module looks through the reference files for the records that fail the previous modules, using name, sex, and date of birth data.

Household Composition Search Module: When an incoming record fails to find a match in the reference files through the preceding modules, it proceeds to the Household Composition module. This module requires at least one person in the household of the unmatched person to have received a PIK. It then creates a universe of unmatched records with historical name, date of birth, sex, and address data from households whose members with PIKs were observed in the past. The module attempts to find a match in this universe based on name and date of birth information.

3.2 Enhancements to the Record Linkage System

As reported by Brown, et al. (2019a), a PIK could not be assigned to 9.0 percent of persons in the 2010 Census. This included 3.3 percent of census records not sent to PVS search due to insufficient personally identifiable information (PII).²² Since we cannot link these records to other data sources due to insufficient PII, we must use statistical models to predict their citizenship status, as discussed in Sections 6 – 9. The other 5.7 percent of census records did not receive PIKs because they could not be found in the reference files or had links to multiple reference file records. Changes to the record linkage system can potentially facilitate the linkage of some census records not found in the NUMIDENT, nor among ITINs, to other data sources that could provide their citizenship status.

To do this, we created an Enhanced Reference File (ERF) by assembling records that have sufficient PII to be sent to search, but are not found in the NUMIDENT or ITINs.²³ The records

²² Records that do not contain at least name and age are not sent to PVS search. Proxy responses and count imputations are leading causes of insufficient PII. PIK rates were just 33.8 percent for proxy responses in the 2010 Census (Rastogi and O'Hara, 2012) and were zero by definition for count imputations.

²³ Records not receiving a PIK because they match multiple reference file records are not included.

are taken from several sources.²⁴ Once assembled, the records were unduplicated. Each cluster of records that appear to represent the same person were assigned a unique Enhanced Protected Identification Key (EPIK). All addresses, name variations, and date of birth variations associated with each EPIK have been assembled in the ERF.

We created EPIK crosswalks for each of the files making up the ERF, as well as the 2010 CUF and 2018 ACS. Construction of these crosswalks involved taking records from each source file that have sufficient PII to be sent to search and are not found in the NUMIDENT or ITINs. Attempts were made to link the records to the ERF using a series of modules like those used in the current PVS system described above. These crosswalks will facilitate the linkage of administrative record citizenship data to the 2020 CUF for some of the persons not receiving PIKs in either the administrative data or the CUF.²⁵

The share of the 2018 ACS without a PIK but with EPIK citizenship information is 0.108 percent. Table 3.1 shows that over half of these records come from SEVIS, and most of the rest come from USCIS, ADIS, and passport data. The low share contributed by EPIKs could be due in part to outdated address information for records not receiving PIKs. This segment of the population is likely to be highly mobile within the U.S.

3.3 Record Linkage Error

The record linkage system will occasionally link a record for one person in a source to a different person in the same or another source. This can happen, for example, when different people have the same or very similar names and dates of birth. Mismatches can also occur when a parent and child have the same name and address.

A linkage error may or may not introduce errors into the citizenship estimates. We show two examples where the linkage error causes citizenship estimation error, followed by two examples where it does not. Person A is in the 2020 CUF, and her true status is noncitizen. We note the effects of the linkage errors on determination of citizenship status by business rules (such as are discussed in Sections 4 – 8) and by latent class models (which are discussed in Section 9).

Example 1: Record linkage error across citizenship sources. A green card for person A is incorrectly linked to a U.S. passport for person B, and both are linked to person A's CUF record. Without the link between the green card and U.S. passport, business rules classify person A as a noncitizen, and the latent class model produces a low probability of person A being a citizen. With

²⁴ The sources include U.S. Customs and Immigration Services (USCIS) naturalizations and lawful permanent residents, Customs and Border Protection Arrivals and Departures Information System (ADIS), State Department passports and Worldwide Refugee Admissions Processing System (WRAPS), Bureau of Prisons, Bureau of Justice Statistics National Corrections Reporting Program (NCPR), U.S. Marshals Service, Department of Interior Incident Management Analysis Reporting System (IMARS), Medicare and Medicaid enrollment, Department of Housing and Urban Development Longitudinal File, Indian Health Service enrollment, Selective Service System, Veterans Affairs, and Nebraska driver's licenses. We did not include IRS data, because its primary linkage variable is the taxpayer identification number (SSN or ITIN), and those with ITINs or verified SSNs are already included in the PVS reference files. IRS records do not contain date of birth, and many of the records contain only the first four letters of the last name. We are currently researching whether to include state SNAP, TANF, or WIC files in the ERF.

²⁵ We will be unable to link in person-level citizenship information to CUF records lacking a PIK or EPIK. Sections 5 – 8 discuss statistical models that can be used to predict their citizenship.

the link, business rules classify person A as a citizen, and the latent class model produces a greater citizen probability.

Example 2: Record linkage error between a citizenship source and the 2020 CUF. The U.S. passport record for Person B is linked to person A's record in the 2020 CUF. The business rule classifies person A as a citizen, and the latent class model produces a high citizen probability. Without this link, person A's citizen probability is estimated by the model without her administrative record citizenship information, and the citizen probability is lower than it would be with a link to person B's passport record.

In both Examples 1 and 2, person A is misclassified as a citizen with business rules, and the modeled citizen probability is higher with the incorrect link.

Example 3: Person A's green card record is incorrectly linked to a noncitizen Bureau of Prisons record for person C. The Bureau of Prisons noncitizen record doesn't affect person A's business rules classification (it is noncitizen with or without it), and it further lowers the modeled citizen probability (moving it in the direction of the correct status).

Example 4: A noncitizen driver's license record for person C is incorrectly linked to person A in the 2020 CUF. As in example 3, the linked driver's license record lowers the modeled citizen probability and doesn't affect person A's business rules classification.

The fact that the variables used for record linkage are correlated with citizenship status may result in smaller linkage error effects on citizenship estimates than random linkage errors would cause. Age, sex, location, race, and ethnicity are all strong predictors of citizen or noncitizen status in the latent class modeling discussed in Section 9. Though race and ethnicity are not variables used in the record linkage, names are used, and names are also correlated with race and ethnicity.²⁶ The correlation with citizenship status means that administrative records incorrectly linked among themselves will tend to indicate the same citizenship status. Thus, incorrectly linked administrative records to CUF records will tend to indicate the same citizenship status as the CUF person's true status more frequently than would be the case with random linkage errors.

3.4 Record Linkage Quality Proxy

Layne, Wagner, and Rothhaas (2014) have estimated aggregated false match rates for the current production PVS record linkage system. It would be ideal to have probabilities of correct linkage at the record-to-record level for the purpose of estimating the effect of linkage error on citizenship estimation error. A record linkage research team at the Census Bureau is currently exploring alternative linkage methods that could provide such probabilities. This work will not be completed in time for implementation in the 2020 CVAP tabulations, however.

Absent estimated probabilities of correct record linkage, we construct an alternative model for record linkage quality. This linkage quality indicator can be used to moderate source information in the models, giving records with higher expected linkage quality more influence. It can also inform selection of records for the business rules.

²⁶ Names will be used when imputing race and ethnicity based on the internal specifications of the 2020 CEF.

For the record linkage quality model, we create an outcome measure for each PIK-source combination, because record linkage error is specific to the PIK-source combinations, not to the PIK in general. Layne, Wagner, and Rothhaas (2014) show that false match rates vary considerably across modules and passes in the PVS. PVS scores measure the confidence of the link within a module-pass, but they are not comparable across module-passes.²⁷ There is thus no single proxy for record linkage error in the PVS. If we were to use both module-pass indicators (of which there are 18 in the examples below) and PVS scores in the model for each PIK-source combination, we would have to include a large number of regressors for each PIK-source outcome.

For most sources we construct an agreement indicator equal to one if a noncitizen in the source is foreign-born in the NUMIDENT, and zero if the noncitizen is U.S.-born. Passport data cover only citizens, but the data also include country of birth, so the agreement indicator for passports is equal to one if the NUMIDENT foreign-born indicator and the constructed passport foreign-born indicator agree, and it is zero otherwise. We use agreement on foreign-born status rather than citizenship status because, unlike citizenship, foreign-born status is time-invariant, so discrepancies are not due to misaligned timing across sources. Also, though citizenship misreporting may occur due to confidentiality concerns, it is hard to imagine a U.S.-born survey respondent reporting that they are a noncitizen for this reason.

We estimate logistic regressions with the agreement indicator as the dependent variable and the PVS score and module-pass indicators as independent variables. Regressions are run separately by source. The model prediction is our proposed proxy for record linkage error in the PIK-source. This proxy distills the module-pass and score information into a single variable.

Table 3.2 displays foreign- versus U.S.-birth agreement rates between USCIS and NUMIDENT data for USCIS noncitizens by module-pass combinations, in the order in which they are attempted in PVS. Agreement rates vary noticeably across module-passes. SSN verification cases have a very high agreement rate. Later passes generally have lower agreement rates, which is consistent with disagreement being associated with record linkage error. The patterns are not linear, however, suggesting that a quality measure that simply ranks module-passes may not be ideal. These patterns are consistent with the findings of Layne, Wagner, and Rothhaas (2014).

The results of logistic regressions of the foreign- versus U.S.-birth agreement indicator for noncitizens in the source on the PVS score and module-pass indicators are shown in Table 3.3. The PVS score is significantly positively associated with agreement, with the exceptions of LEMIS, Nebraska driver's licenses, and SNAP/TANF. The marginal effect is quite large in the ADIS data, suggesting that PVS information can help identify weaker linkages in that source. The proxy may not be useful in identifying erroneous linkages in the sources with negative PVS score coefficients.

Table 3.4 shows how foreign- versus U.S.-birth agreement rates vary between the top and bottom deciles of the distribution of the predicted foreign-born agreement probabilities generated from the

²⁷ Identifiers used in the linkage process are assigned a “distance” measure for the degree of difference between them in the records being compared. The identifiers are assigned weights, and the total weighted comparison yields the PVS score.

models in Table 3.3. As intended, the agreement rates are higher in the top than the bottom decile, but the distance between them varies considerably across sources. The difference is minimal for passports and Nebraska driver's licenses, but is more than 72 percentage points for ADIS. This suggests that record linkage quality is an issue for some sources, but not others.

3.5 An Alternative Record Linkage Quality Proxy

We also used an alternative approach to assessing record linkage quality in conjunction with the business rules used for the estimation approach described in Section 7. This method developed decision trees to generate predicted probabilities of correct record linkage quality as described below. Decision trees were generated separately for each state.

This processing started with person records in the frame (2010 CEF) where a PIK was assigned. For these records, the processing checked each non-NUMIDENT source to see if it qualified for inclusion in the record linkage quality analysis. If the non-NUMIDENT source indicated a noncitizen status and the NUMIDENT foreign born agreed, then the CEF person/source pair received a '1'. If the non-NUMIDENT source indicated their noncitizen status and the NUMIDENT foreign born disagreed, then this CEF person/source pair received a '0'. A similar process was used for foreign-born status comparisons of passports to assess agreement or disagreement for that source. If a person had three sources that qualified, then this person would contribute three records to the subsequent decision tree analysis.

For each state, a decision tree analysis was run. The analysis variable was the record linkage quality indicator (1/0) described in the paragraph above. The predictor variables used in estimating the decision tree were data source, PVS module and pass combined indicator, and the PVS score. Each state was run separately with all the data sources included in each run. The estimated tree used a maximum depth of eight and required minimum final node sizes (known as "leafsizes") to be 20 persons or more. The result of the tree was that each source, with its PVS module, pass, and PVS score, was assigned to one of the final nodes.²⁸

For each 2010 Census person with a PIK that is linked to non-NUMIDENT sources, the predicted probability of correct record linkage was assigned to each non-NUMIDENT source for that person. The assignment was based on which node corresponded to the person's source, PVS module, pass, and score. Consider, for example, a 2010 Census person who had a PIK that linked to both USCIS and passport sources. Suppose also that the PIK that was assigned to the USCIS record got a PVS score that was greater than the PVS score assigned by the passport PVS processing. To assess the record linkage quality of the USCIS PIK, this procedure found the node that contained the USCIS source with its higher PVS score, and used that predicted probability, say a . To assess the record linkage quality of the passport PIK, the procedure found the node that contained the passport source and used that predicted probability, c . When making the rules, the predicted probability of accurate record linkage needed to exceed a certain cutoff value, say b . Suppose that, for this record,

²⁸ The SAS HPSPLIT procedure was used for this work.

$c < b < a$. Then, for this person, the USCIS information would be used (since $a > b$), but the passport information would not be used (since $c < b$).

For CEF persons without PIKs who linked to a source based on EPIK assignments, record linkage quality determinations were also made. This was done, essentially, by applying the decision tree developed for the records that received a PIK to those records with an EPIK. This assigned probabilities of correct linkage to the EPIK linkages that were then used as described for the PIK linkages.

Table 3.1 EPIK Coverage in 2018 ACS by Source

	Percent of EPIKs
U.S. Passports	11.56
USCIS Naturalizations and LPRs	17.81
ADIS	16.67
SEVIS	55.81
BOP	0.94
USMS	1.63

Notes: These are for persons age 18 and over in the 2018 ACS, using ACS person weights. The percentages for WRAPS and IMARS are too small to be released by the Disclosure Review Board (DRB). The percent of the 2018 ACS with EPIK citizenship information is 0.108. The numbers do not add to 100 percent, because some EPIKs have citizenship information from multiple sources. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 3.2 Foreign-Born Agreement Rates between NUMIDENT and USCIS

Module	Pass	Percent Agreement	Observations
Verification		99.70	22,110,000
GeoSearch	1	99.73	2,487,000
	2	100.00	1,700
	3	99.68	152,000
	4	D	80
	5	99.84	316,000
	6	93.33	750
	7	99.71	115,000
	8	D	1,200
	9	99.80	1,210,000
Name Search	1	96.41	11,130,000
	2	94.05	745,000
	3	72.41	2,900
	4	62.30	96,000
DOB Search	1	84.51	24,500
	2	85.98	15,000
	3	82.50	4,000
	4	94.56	67,500
Total		98.52	38,480,000

Notes: A USCIS record disagrees with the NUMIDENT if the NUMIDENT says the person is U.S.-born. "D" indicates the cell is suppressed due to disclosure avoidance rules. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 3.3 Logistic Regressions for Foreign-Born Agreement between NUMIDENT and Source

Variable	Coefficient	Marginal Effect
	U.S. Passports	
PVS Score	0.2280	0.0008348
	(0.0004365)	(0.0000019)
N	175,000,000	
	USCIS	
PVS Score	0.1652	0.002286
	(0.0005695)	(0.00000827)
N	38,480,000	
	ADIS	
PVS Score	0.2049	0.02924
	(0.0002994)	(0.0000401)
N	13,320,000	
	SEVIS	
PVS Score	0.3430	0.003741
	(0.002695)	(0.0000321)
N	2,294,000	
	WRAPS	
PVS Score	0.3440	0.0000922
	(0.02515)	(0.0000105)
N	398,000	
	BOP	
PVS Score	0.2448	0.009728
	(0.002086)	(0.0000801)
N	1,127,000	
	USMS	
PVS Score	0.5436	0.01059
	(0.004723)	(0.0000496)
N	474,000	
	IMARS	
PVS Score	0.07849	0.01675
	(0.07052)	(0.01448)
N	80	
	LEMIS	
PVS Score	-0.05799	-0.01099
	(0.01644)	(0.003014)
N	500	
	Nebraska Driver's Licenses	
PVS Score	-0.4976	-0.03268
	(0.01900)	(0.001218)
N	59,000	

Table 3.3 Continued

Variable	Coefficient	Marginal Effect
	South Dakota Driver's Licenses	
PVS Score	0.01553	0.0001547
	(0.01710)	(0.001706)
N	11,000	
	SNAP/TANF	
PVS Score	-0.0412	-0.002747
	(0.0003)	(0.0000195)
N	1,160,000	
	2005-2018 ACS	
PVS Score	0.05457	0.0005269
	(0.0001923)	(0.00000198)
N	52,250,000	
	2001-2017 AHS	
PVS Score	0.0613	0.001938
	(0.005958)	(0.0001919)
N	38,000	
	2005-2018 CPS	
PVS Score	0.05262	0.001515
	(0.004158)	(0.0001224)
N	68,000	
	2004-2017 SIPP	
PVS Score	0.08386	0.004447
	(0.01091)	(0.0005898)
N	10,000	
	2018 NCRP	
PVS Score	0.1277	0.01200
	(0.01872)	(0.001748)
N	12,000	

Notes: Module-pass controls are included. Robust standard errors are in parentheses. The standard errors for the marginal effects are calculated via the delta-method. The samples are noncitizen records in the source. The dependent variable equals one if the 2020 quarter 1 NUMIDENT indicates the person is foreign-born, and zero if U.S.-born. The one exception is U.S. passport records; a passport record disagrees with the NUMIDENT if the passport record indicates the person is foreign-born and the NUMIDENT says U.S.-born, or vice versa. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 3.4 Foreign-Born Agreement Rates with NUMIDENT by Predicted Record Linkage Quality (RLQ)

Source	Percent Agreement RLQ \geq 90 th Percentile	Percent Agreement RLQ \leq 10 th Percentile
U.S. Passports	99.86	99.05
USCIS	99.72	92.63
ADIS	99.80	26.98
SEVIS	99.98	86.38
WRAPS	D	99.81
BOP	97.37	62.33
USMS	99.59	49.43
IMARS	D	D
LEMIS	50.00	D
Nebraska Driver's Licenses	99.69	99.53
South Dakota Driver's Licenses	100.00	92.31
SNAP/TANF	95.09	58.25
2005-2018 ACS	96.29	85.52
2001-2017 AHS	100.00	73.91
2005-2018 CPS	99.20	70.59
2004-2017 SIPP	100.00	75.00
2018 NCRP	93.10	58.33

Notes: Agreement is equal to one if the noncitizen record from the source is linked to a foreign-born NUMIDENT record, and zero if linked to a U.S.-born NUMIDENT record. "D" indicates that the number is suppressed due to disclosure avoidance restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

4. Business Rules

Three of the four methods for estimating citizenship status analyzed in this report use business rules to create citizenship values for person records linked to one or more administrative or survey sources providing information on citizenship. This section considers how to construct the business rules.

4.1 Source Coverage

To analyze data source coverage of the population, we use the 2018 ACS for persons age 18 and over as the frame. By using a recent survey, we exclude records for persons who have died or moved out of the country prior to 2018. This allows us to use all of our data sources, which an earlier population frame such as the 2010 Census Edited File (CEF) does not permit. Table 4.1 lists the sources and their shares of the voting age population, as represented by the 2018 ACS. Throughout the report, we use the most recent source citizenship information available at the time of the population frame: April 1, 2010 for the 2010 CEF, and the interview date for the ACS.

The NUMIDENT provides by far the greatest coverage, at 90.4 percent of the population (see Table 4.1). NUMIDENT coverage of U.S.-born persons should be nearly comprehensive for the reasons documented in Section 2.1.1; however, some U.S.-born persons in the CEF and ACS will lack a NUMIDENT record due to record linkages errors. In addition, because NUMIDENT citizenship data are sometimes missing, 10.2 percent of the NUMIDENT records linked to the 2018 ACS lack a value for the citizenship variable. We pay attention to whether the foreign-born indicator is sufficiently reliable to classify U.S.-born persons as citizens when the citizenship variable is missing.²⁹ Citizenship status is unclear for the one percent who have missing citizenship and are foreign-born or lack a foreign-born indicator.

Passports are held by 48.6 percent of the population, as-reported citizenship in ACS surveys prior to 2018 cover 14.1 percent, U.S. Customs and Immigration Services (USCIS) data cover 11.5 percent, and Customs and Border Patrol (CBP) Arrivals and Departures Information System (ADIS) data cover 1.7 percent (See Table 4.1). None of the other datasets cover more than one percent of the population. Though Individual Taxpayer Identification Numbers (ITINs) represent just 0.4 percent of the population, this is a non-trivial share of the persons not covered by the NUMIDENT, and also of the total number of noncitizens (since ITINs reflect noncitizen status).

Given the discussions in Sections 2 and 3, we note the following points about data quality that guide the analyses given here. First, citizen indicators are generally more reliable than noncitizen indicators, because a noncitizen indicator from a source could have been correct when initially recorded, but was outdated if the person subsequently was naturalized. In contrast, persons changing from U.S. citizens to noncitizens are extremely rare, and most such cases occur for people living outside the U.S., who are not part of the CVAP reference population. Also, many administrative sources require persons to submit proof of U.S. citizenship for their inclusion in the records as citizens. Survey sources recording responses to a citizenship question are subject to

²⁹ U.S.-born children whose parents are foreign diplomats are not automatically given U.S. citizenship, but this is a tiny fraction of all U.S.-born persons.

misreporting error, which we will see is more frequent for true noncitizens than for true citizens. Survey data will also include imputations of citizenship status for nonresponse to the citizenship question, which are less reliable than reported status. Finally, as discussed in Section 3, record linkage errors can lead to errors by falsely linking a noncitizen person record to the census record of a true citizen or a citizen person record to the census record of a true noncitizen.

4.2 Evaluating Sources for Business Rules

We consider whether 2018 ACS citizenship responses are of sufficient quality to serve as a potential guide for determining whether and how to use each available source-citizenship combination in the business rules. To provide a comparator to evaluate the accuracy of ACS citizenship responses, we first construct a set of business rules for assigning citizenship status to census records using the primary sources described in Section 2, which should be the highest-quality sources given their strict verification. The degree of agreement among these sources gives an indication of their quality. Table 4.2 compares citizenship information in the NUMIDENT, passports, USCIS, and ITINs for each combination of values across sources, using the 2018 ACS as the population frame. Note that even if source information is high-quality, the linked information could be incorrect due to linkage error. We provide a second set of results dropping passport and USCIS records at or below the 25th percentile for the record linkage quality indicator discussed in Section 3.

Overall (last line of Table 4.2), only 3.4 percent of the population is in a discrepant category, 2.6 percentage points of which are NUMIDENT noncitizens with a passport and/or a USCIS naturalization certificate. These appear to be persons who have not notified SSA about their naturalization. The other major discrepancies involve passports for USCIS noncitizen records, which can occur for persons naturalized as children when their parents are naturalized. When imposing a record linkage threshold, the overall discrepant share falls to 2.7 percent.

USCIS data should not contain U.S.-born persons. Indeed, almost no U.S.-born NUMIDENT persons with missing citizenship are linked to a USCIS record, and that share drops by more than half when imposing a record linkage quality threshold. U.S.-born NUMIDENT persons are often linked to passports, however. These patterns suggest that U.S. birth in the NUMIDENT is a reliable indicator of being a citizen, and that USCIS data cover only foreign-born persons, as agency documentation indicates.

Passport and USCIS data contribute citizenship information for 84 percent of foreign-born NUMIDENT persons with missing citizenship. This share falls to 40 percent when a record linkage quality threshold is used.

Virtually no ITINs overlap with the other sources, as expected, since persons in the other sources should be eligible for SSNs.

Table 4.3 shows the discrepancy rates for ADIS, Student Exchange Visitor Information System (SEVIS), and Worldwide Refugee Admissions Processing System (WRAPS) noncitizen records compared to the NUMIDENT, passports, and USCIS, where a record is discrepant if NUMIDENT,

passports, or USCIS indicates the person is a citizen.³⁰ SEVIS and WRAPS noncitizen records nearly always link only to other noncitizen records, whereas ADIS records link to citizen records about half the time. Imposing a record linkage quality threshold reduces the discrepancies for each source, but the discrepancy rate remains very high for ADIS.

We combine the primary data sources using the following set of business rules. First, we classify a person as a citizen if they are a NUMIDENT citizen, a U.S.-born NUMIDENT person with missing citizenship, hold a passport, or have a USCIS naturalization certificate. Noncitizens are those not classified as citizens and are either a NUMIDENT noncitizen, ITIN, USCIS lawful permanent resident or refugee, ADIS temporary visa holder, SEVIS temporary visa holder, or WRAPS refugee. The rules set the citizenship value to missing for foreign-born persons with missing NUMIDENT citizenship and no citizenship information from any of the other sources. This set of rules assumes that being a citizen in just one of these citizen sources is sufficient evidence that the person is a citizen, even if contradicted by another source's noncitizen value.³¹ As shown in Tables 4.2 and 4.3, the discrepancies among these sources are minimal. When discrepancies do exist, there is good reason to trust the citizen value more than the noncitizen value (e.g., a naturalization certificate over a NUMIDENT noncitizen value or a passport over a USCIS lawful permanent resident record). We do not use ADIS data here due to their high disagreement with the other sources.

Tables 4.4.A – C show agreement rates between the business rules and 2018 ACS citizenship among ACS persons age 18 and over. Table 4.4.A includes only ACS records with as-reported citizenship. The citizen share in the ACS and the benchmark differ by just 0.13 percentage points. The disagreement rates conditional on being a business rules citizen or on being an ACS citizen are very small (0.71 percent and 0.84 percent, respectively), while they are much larger when conditioning on being a business rule or ACS noncitizen (12.2 percent and 10.5 percent, respectively). Edited or imputed ACS citizenship data produce much larger discrepancies with the business rules (Tables 4.4.B and C, respectively). The ACS imputations also produce a more than two percentage point higher citizen share than the business rules. These results suggest that the as-reported ACS citizenship responses are reasonably good, but the edits and imputes are less reliable.

When developing our recommended set of business rules, we started with rules using all the administrative and survey records with citizenship information that can be linked to the 2018 ACS. A person is classified as a citizen if any source indicates the person is a citizen. A person is classified as a noncitizen if no source indicates the person is a citizen and at least one source says noncitizen. A person is classified as missing if and only if there are no citizenship or noncitizen sources available for that person.

³⁰ We focus on noncitizens here, because the SEVIS and WRAPS records are entirely for noncitizens, and those for ADIS are nearly so.

³¹ The use of citizen data to override noncitizen data implicitly ignores loss of citizenship, which is a rare event. According to U.S. Treasury statistics, 42,341 persons renounced their U.S. citizenship between January 1998 and June 2020. See Goodin (2020) who states that most of these people already live abroad, which means they are unlikely to be in the resident U.S. population as defined for the 2020 Census.

We use as-reported 2018 ACS citizenship responses to guide the refinement of the business rules. As documented in Table 4.4 there is a modest amount of misclassification in the ACS self-reports, particularly for the ACS response of noncitizen. There is also potential for misclassification when different administrative records give conflicting answers for citizenship—citizen on one and noncitizen on the other. Our analysis of whether a particular combination of administrative record source and the citizenship value it implies improves or compounds the misclassification problem is based on the exercise summarized in Table 4.5. We consider all possible pairs of administrative record source and citizenship value on that record source. We remove each pair from the business rules, re-run the rules, then examine the resulting output. The pair is then restored to the administrative data universe, the next pair is deleted, and its output is examined. We repeat this process until all pairs have been examined. There are three possible outcomes for every record in the population frame: (1) its business rule citizenship value is unchanged; (2) its citizenship value flips from citizen to noncitizen or vice versa; or (3) its citizenship value becomes missing. In the first case, the marginal effect of the pair is zero for that person—the remaining administrative records gave the same business rule result. In the second case, the administrative record-citizenship value pair was the defining datum in the business rules, and its absence flips the business rule. In the third case, the pair was the only administrative data for the person, and in its absence the citizenship value is missing.

In the first case, there is nothing further to analyze. If every pair generated the no-change marginal outcome, we would say that the administrative data had two-source confirmation of every classification. It is the second and third cases that require an objective function to measure the marginal effect on the quality of the citizenship data from the particular administrative record-citizenship value pair. We use an objective function based on the reported citizenship value in the ACS.

For the second case, we compare the percent citizen (or noncitizen) among persons whose business rule classification changed when the marginal administrative record-citizenship value pair was deleted to the percent citizen (or noncitizen) in the ACS for the same people. If the pre-deletion citizen (or noncitizen) BR percentage is closer to the ACS, we recommend retaining the pair in the BR universe. If the post-deletion citizen (or noncitizen) BR percentage is closer to the ACS, we recommend deleting the pair from the BR universe. If the pre-deletion and post-deletion BR percentages are both close to the ACS percentage, we recommend retaining the administrative record-citizenship value pair in the BR universe. This is not the same as treating the ACS as ground truth. Instead, we are arguing that an outcome far from the ACS average for the affected persons is more likely to be BR misclassification. This is also why, when the differences are small compared to the ACS, we recommend retaining the pair.

In the third case, instead of comparing the citizen (or noncitizen) BR percentages before and after the deletion to the ACS percentage, we substitute the prediction from our BR-based prediction model because the person has missing BR citizenship status after the deletion. Once again, if the before-deletion citizen (or noncitizen) percentage is closer to the ACS percentage than the model prediction, we recommend keeping the pair in the BR universe. If the model prediction is closer, we recommend deleting the pair from the BR universe. When the ACS doesn't provide clear

guidance for a particular source-citizenship combination (the before and after percentages are both close to the ACS percentage), we recommend keeping the pair in the business rules because the ACS has response error, and there is no guarantee that the citizenship model trained on the ACS data will perform the same in the 2020 Census as in the 2018 ACS.

Now consider Table 4.5 in detail. As noted above, there are two possible effects from dropping a source-citizenship combination. One is to change a person's classification from citizen to noncitizen. This occurs if a citizen source is dropped, there are no other sources indicating the person is a citizen, and at least one source indicates the person is a noncitizen.³² In such cases, we calculate the citizen share of as-reported ACS responses (see the second column, “% ACS Citizens for Changed Values,” in Table 4.5). If the ACS citizen share is under 50 percent, then dropping the citizen source would increase agreement between the business rules and ACS citizenship. The second possible effect is that the person would no longer have business rules citizenship if the source being dropped is the only one available. For these cases, we compare the business rules citizen share if the source were kept, as well as the model-estimated share, to the as-reported 2018 ACS citizen share (see the last three columns of Table 4.5). If the model-estimated share is closer to the ACS share than the business rules share, then omitting the source-citizenship combination would raise business rules-ACS agreement.³³

Besides omitting all values from a source-citizenship combination, we also consider dropping records at the 25th percentile or below of the record linkage quality distribution, as well as the 25th percentile or below (i.e., older records) of the record vintage distribution.³⁴ Older-vintage records are more likely to contain out-of-date information. To illustrate this, we calculated the percent citizens according to the 2018 business rules used in Table 4.4 among persons with ACS noncitizen responses in prior years. Figure 4.1 shows that two-thirds of the 2005 ACS noncitizens are business rules citizens in 2018. The business-rule citizen share gradually declines as the ACS interview year gets closer to 2018, falling to about 20 percent in 2017. Vintage is less likely to matter for citizen responses, since U.S. resident citizens almost never switch to noncitizen status, as noted above.

Regarding the effects of omitting a source on switches from citizen to noncitizen classification, several citizen sources have low 2018 ACS citizen shares, including ADIS, Bureau of Prisons (BOP), U.S. Marshall Service (USMS), Nebraska Driver's Licenses, Supplemental Nutritional Assistance Program (SNAP), 2005-2017 ACS, American Housing Survey (AHS), Current Population Survey (CPS), Survey of Income and Program Participation (SIPP), South Dakota Driver's Licenses, and National Corrections Reporting Program (NCRP) data. Combinations for which model estimates are clearly closer to the ACS than the business rules include USCIS noncitizens with record linkage quality in the 25th percentile or below, ADIS noncitizens with

³² Since citizen values are given preference in these rules, dropping a noncitizen value would not cause the classification to change from noncitizen to citizen.

³³ If we were to omit a significant number of records with citizenship information from the business rules, we could consider including these sources as model predictors. The number we recommend excluding is very small, however, so we would not be able to produce reliable coefficients for such indicators.

³⁴ The 25th percentile threshold works well to divide higher and lower-quality records in the sources for which we recommend applying a threshold. Given the small sample sizes in these tests, we are unable to determine threshold values with great precision for each source.

vintage or record linkage quality in the 25th percentile or below, SNAP noncitizens, and NCRP noncitizens. BOP citizens, SNAP citizens, and ACS citizens have business rule and model estimates that are very similar to each other. Though the model is slightly closer to the ACS for those source-citizenship combinations, we recommend keeping them in the business rules.

Table 4.5 shows the as-reported 2018 ACS citizen share for cases where omitting a source's citizen value results in a switch to a noncitizen classification. We display this separately for the recommended deletions listed above, as well as for all the other potential deletions of this type, which we recommend retaining. Only 27 percent are citizens among the recommended deletions, so switching the classification to noncitizen brings the BR estimates closer to the ACS. In contrast, 96 percent are citizens among other potential omissions, so switching would enlarge the gap with the ACS; hence, we recommend retaining those pairs.

Among cases with just one citizenship source, the model estimate is 21.1 percentage points $((0.723 - 0.467) - (0.467 - 0))$ closer to the ACS than the business rule for the recommended deletions listed above. Among recommended retentions, the model estimate is 0.11 percentage points $(0.997 - 0.986)$ further away from the ACS for business rule citizen values and 21.8 percentage points $((0.416 - 0.099) - (0.099 - 0.000))$ further away for business rule noncitizen values; hence the recommendation to retain these pairs.

Our recommended business rules are shown in Table 4.6. A person is classified as a citizen if they have information suggesting they are a citizen in the NUMIDENT, passports, and/or USCIS. Persons without evidence of being a citizen in the above sources and who have a noncitizen value in at least one of the primary or secondary noncitizen sources listed in the table are classified as noncitizens. Persons with no information suggesting citizenship status in either of the above groups and who have citizen values from ADIS or the listed secondary citizen sources are classified as citizens. A source's position within the source category (primary citizen sources versus primary and secondary noncitizen sources versus secondary citizen sources (plus ADIS)) doesn't matter when constructing the rules. The order matters only when calculating the source-citizenship combination's share of the population in the table. Only the NUMIDENT, USCIS, ITIN, SEVIS, and ADIS data cover more than 0.01 percent of the population in the table. The business rules, in aggregate, don't cover 9.12 percent of the population, which require modeling.

Table 4.7 compares several candidate variants of the business rules we considered with as-reported 2018 ACS responses. The "All Records, Citizen Values Prioritized" is the set we started with in Table 4.5 before pruning. "Primary, then Secondary" first uses citizen values from primary sources, then noncitizen values from primary sources, and finally secondary source citizen and noncitizen values as long as they don't disagree with each other. "Primary, then Secondary" is the set of rules used by the ACS logistic approach for the 2010 CEF test. "Primary Only" uses primary citizen values, then primary noncitizen values. The Hot Deck approach uses the "Primary Only" rules. "Decision Tree Thresholds" uses a decision tree machine learning technique to predict record linkage quality (see Section 3). The method imposes record linkage quality thresholds for each source. This is used in the Business Rules logistic regression approach described in

Section 7.³⁵ “Primary, RLQ Restrictions” is the set of rules used in Table 4.4. “Strict Pruning Rules” is the same as the recommended rules, except that BOP citizen, SNAP citizen, and ACS citizen values are not used at all.

We see from Table 4.7 that the differences across the estimates are quite small, so the choice of business rules doesn’t affect the overall citizenship estimates much. The “Primary, RLQ Restrictions” rules have the highest disagreement, the lowest correlation, and furthest citizen share from the ACS. This approach is the most aggressive at dropping lower-quality records, but it doesn’t take into account the fact that lower-quality records can still produce closer results to the ACS than the model can. The “Decision Tree Thresholds” and “All Records, Citizen Values Prioritized” are also further away from the ACS than the others. Note that “Decision Tree Thresholds” is also bit more aggressive in imposing record linkage quality thresholds. The “All Records, Citizen Values Prioritized” doesn’t consider source quality, linkage quality, or vintage at all. The recommended rules and “Strict Pruning Rules” have the lowest disagreement, highest correlations, and have citizen shares within just 0.01 percent of the ACS.

Finally, we examine whether it is better to use the model estimates when the quality of the PIK or EPIK link to the 2018 ACS is low. Table 4.8 compares the 2018 ACS citizen share to the estimates when using different combinations of business rules and the model. The modeled observations are for records from the lower tail of the record linkage quality distribution, where the threshold differs in each row of the table. The citizen share is further away from the ACS when more citizenship values are modeled. There is little difference between dropping business rules for the bottom 10 percent or 5 percent of the distribution versus using only business rules. This suggests that business rules brought in with lower-quality links produce estimates at least as close to the ACS as the model does.

As the model is improved, it may become optimal to omit more source-citizenship combinations from the business rules, as well as use the model in place of the business rules for records with lower-quality PIK and EPIK links to the population frame.

³⁵ The results for this approach in Table 4.7 use the same model as the others in the table. The results are qualitatively similar when using the Business Rules logistic model described in Section 6.

Table 4.1 Administrative Record Coverage of 2018 ACS

Source	Percent of 2018 ACS
NUMIDENT	90.40
Of which: Citizen value for citizenship	72.43
Noncitizen value for citizenship	7.81
U.S.-born, missing citizenship	9.15
Foreign-born or uncertain country of birth, missing citizenship	1.02
U.S. Passports	48.58
USCIS	11.52
Of which: Citizens	6.55
Noncitizens	4.97
ITINs	0.52
CBP ADIS	1.66
Of which: Citizens	0.01
Noncitizens	1.66
ICE SEVIS (Noncitizens)	0.30
WRAPS (Noncitizens)	0.08
Dept. of Interior IMARS	D
Dept. of Interior LEMIS	0.00
Bureau of Prisons	0.25
Of which: Citizens	0.21
Noncitizens	0.03
U.S. Marshals Service	0.14
Of which: Citizens	0.11
Noncitizens	0.03
NCRP	0.04
Of which: Citizens	0.04
Noncitizens	<0.01
Nebraska Driver's Licenses	0.33
Of which: Citizens	0.32
Noncitizens	0.01
South Dakota Driver's Licenses	0.03
Of which: Citizens	0.03
Noncitizens	<0.01
SNAP/TANF	2.85
Of which: Citizens	2.64
Noncitizens	0.21
2005-2017 ACS	14.07
Of which: Citizens	13.48
Noncitizens	0.59

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.1 Administrative Record Coverage of 2018 ACS Continued

Source	Percent of 2018 ACS
2001, 2003, 2004, 2005, 2007, 2009, 2011, 2013, 2015, and 2017 AHS	0.24
Of which: Citizens	0.23
Noncitizens	0.01
2005-2017 CPS	0.45
Of which: Citizens	0.42
Noncitizens	0.03
2004-2017 SIPP	0.08
Of which: Citizens	0.07
Noncitizens	<0.01

Notes: These percentages use ACS survey weights. The 2018 ACS estimate of the population age 18 and over is 253,800,000. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.2 Percent of ACS by Source Citizenship Combinations

NUMIDENT	U.S. Passport	USCIS	ITIN	Percent of 2018 ACS	Percent of 2018 ACS, >25% RLQ
	Citizenship Agreement				
Citizen	Absent	Absent	Absent	31.77	42.40
Citizen	Absent	Citizen	Absent	0.36	0.39
Citizen	Citizen	Absent	Absent	35.90	26.70
Citizen	Citizen	Citizen	Absent	3.63	2.65
Noncitizen	Absent	Absent	Absent	1.54	1.93
Noncitizen	Absent	Noncitizen	Absent	3.73	3.46
Missing Citizenship, U.S.-Born	Absent	Absent	Absent	3.78	6.02
Missing Citizenship, U.S.-Born	Absent	Citizen	Absent	<0.01	<0.01
Missing Citizenship, U.S.-Born	Citizen	Absent	Absent	5.35	3.13
Missing Citizenship, U.S.-Born	Citizen	Citizen	Absent	0.01	<0.01
Missing Citizenship, Foreign-Born or Missing	Absent	Citizen	Absent	0.03	0.03
Missing Citizenship, Foreign-Born or Missing	Absent	Noncitizen	Absent	0.10	0.08
Missing Citizenship, Foreign-Born or Missing	Citizen	Absent	Absent	0.33	0.21
Missing Citizenship, Foreign-Born or Missing	Citizen	Citizen	Absent	0.29	0.08
Absent	Absent	Noncitizen	Noncitizen	D	<0.01
Absent	Absent	Absent	Noncitizen	0.51	0.52
Absent	Absent	Citizen	Absent	<0.01	<0.01
Absent	Absent	Noncitizen	Absent	0.01	0.01
Absent	Citizen	Absent	Absent	0.01	0.01
Absent	Citizen	Citizen	Absent	<0.01	<0.01

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.2 Shares of ACS Population by Source Citizenship Combinations Continued

NUMIDENT	U.S. Passport	USCIS	ITIN	Percent of 2018 ACS	Percent of 2018 ACS, >25% RLQ
Citizenship Disagreement					
Citizen	Absent	Noncitizen	Absent	0.09	0.07
Citizen	Citizen	Noncitizen	Absent	0.67	0.21
Noncitizen	Absent	Citizen	Absent	0.27	0.28
Noncitizen	Citizen	Absent	Absent	0.08	0.28
Noncitizen	Citizen	Citizen	Absent	1.97	1.71
Noncitizen	Citizen	Noncitizen	Absent	0.23	0.14
Missing Citizenship, U.S.-Born	Absent	Noncitizen	Absent	<0.01	<0.01
Missing Citizenship, U.S.-Born	Citizen	Noncitizen	Absent	D	D
Missing Citizenship, Foreign-Born or Missing	Citizen	Noncitizen	Absent	0.11	<0.01
Absent	Absent	Citizen	Noncitizen	0.00	0.00
Absent	Citizen	Citizen	Noncitizen	0.00	0.00
Absent	Citizen	Noncitizen	Noncitizen	0.00	0.00
Absent	Citizen	Absent	Noncitizen	0.00	0.00
Absent	Citizen	Noncitizen	Absent	D	D

No Citizenship Classification					
Missing Citizenship, Foreign-Born or Missing	Absent	Absent	Absent	0.16	0.61
Absent	Absent	Absent	Absent	9.05	9.05
Total				100.00	100.00
Citizenship Disagreement Share				3.43	2.71

Notes: This uses the 2018 ACS sample and its sampling weights. The total number of unweighted observations is 3,989,000. RLQ is record linkage probability (see Section 3 for details). The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.3 ADIS, SEVIS, and WRAPS Disagreement Rates with Other Primary Sources

	Percent Discrepant	Percent Discrepant, Record Linkage Quality > 25 th Percentile
ADIS	50.38	36.26
SEVIS	5.10	1.81
WRAPS	2.97	2.22

Notes: This uses the 2018 ACS sample age 18 and over and its sampling weights. The number of observations for ADIS is 58,000 and 42,500 without and with the record linkage quality threshold; for SEVIS it is 10,000 and 9,200, and for WRAPS it is 1,900 and 1,400, respectively. A record from the source in the row is discrepant if its citizenship status is different from at least one of the other primary sources (NUMIDENT, passports, USCIS, or ITINs), among those with non-missing values. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.4.A Comparison of 2018 ACS to 2018 BR Citizenship, ACS As-Reported

	Cell Percents		
	BR Citizen	BR Noncitizen	ACS Total
ACS Citizen	92.90	0.79	93.69
ACS Noncitizen	0.66	5.65	6.31
Benchmark Total	93.56	6.44	

	Column Percents		
	BR Citizen	BR Noncitizen	ACS Total
ACS Citizen	99.29	12.21	
ACS Noncitizen	0.71	87.79	
Benchmark Total	100.00	100.00	

	Row Percents		
	BR Citizen	BR Noncitizen	ACS Total
ACS Citizen	99.16	0.84	100.00
ACS Noncitizen	10.46	89.54	100.00

Notes: The number of observations is 3,441,000. These percentages use ACS person weights. The sample is all persons age 18 and over with 2018 business rules citizenship and as-reported ACS citizenship. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.4.B Comparison of 2018 ACS to 2018 BR Citizenship, ACS Edited

	Cell Percents		
	BR Citizen	BR Noncitizen	ACS Total
ACS Citizen	83.80	6.23	90.03
ACS Noncitizen	6.87	3.10	9.97
Benchmark Total	90.67	9.33	

	Column Percents		
ACS Citizen	92.42	66.79	
ACS Noncitizen	7.58	33.21	
Benchmark Total	100.00	100.00	

	Row Percents		
ACS Citizen	93.08	6.92	100.00
ACS Noncitizen	68.94	31.06	100.00

Notes: The number of observations is 13,000. These percentages use ACS person weights. The sample is all persons age 18 and over with 2018 business rules citizenship and edited ACS citizenship. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.4.C Comparison of 2018 ACS to 2018 BR Citizenship, ACS Imputed

	Cell Percents		
	BR Citizen	BR Noncitizen	ACS Total
ACS Citizen	83.49	7.51	91.00
ACS Noncitizen	5.12	3.87	9.00
Benchmark Total	88.61	11.39	

	Column Percents		
ACS Citizen	94.22	65.97	
ACS Noncitizen	5.78	34.03	
Benchmark Total	100.00	100.00	

	Row Percents		
ACS Citizen	92.04	7.96	100.00
ACS Noncitizen	57.72	42.28	100.00

Notes: The number of observations is 191,000. These percentages use ACS person weights. The sample is all persons age 18 and over with 2018 business rules citizenship and edited ACS citizenship. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.5 Analysis of the Effects on Business Rules Resulting from the Deletion of a Single Citizen Value-Administrative Record Pair Based on Changes in the Business Rule Citizenship Determination as Compared to the As-Reported 2018 ACS Citizenship Status

	Source Deletions Resulting in Switches from Citizen to Noncitizen		Source Deletions Resulting in No BR Citizenship Determination			
	% ACS Citizens for Changed Values	Number of Changed Values	Obs. Affected by the Source Deletion	Previous BR for Deleted Source Obs.	Model Prediction for Deleted Source Obs.	As-Reported ACS for Deleted Source Obs.
Citizen Value-Administrative Record Pairs Recommend for Deletion from Business Rules	30.00	2,000	750	0.000	0.723	0.467
Citizen Value-Administrative Record Pairs Recommended for Retention in Business Rules	96.22	141,000	1,013,000	1.000	0.986	0.997
Noncitizen Value-Administrative Record Pairs Recommended for Retention in Business Rules	NA	NA	42,500	0.000	0.416	0.099

Notes: In the second column, the changed values are from citizen to noncitizen. The percent agreement is the percent noncitizens among as-reported 2018 ACS responses for these persons. These switches occur when omitting an administrative record citizen value, and the only other sources have noncitizen values. Model estimates are from the BR model trained on 2013-2017 ACS data (see Section 8). These results are unweighted. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.6 Citizenship Business Rules

Criteria for assigning as citizen	Rule assignment	Percent of 2018 ACS
NUMIDENT citizen	Citizen	72.43
NUMIDENT missing citizenship and U.S.-born	Citizen	9.15
U.S. passport	Citizen	3.01
USCIS naturalization certificate	Citizen	0.29
If not U.S. citizen according to any of the above criteria:		
NUMIDENT noncitizen	Noncitizen	5.27
ITIN	Noncitizen	0.52
USCIS lawful permanent resident, refugee, or asylee RLQ > 25 th percentile (0.9749)	Noncitizen	0.09
ADIS noncitizen records vintage > 7/12/2015, RLQ > 25 th percentile (0.3472)	Noncitizen	0.02
ICE SEVIS record	Noncitizen	0.06
WRAPS record	Noncitizen	D
IMARS record	Noncitizen	D
LEMIS record	Noncitizen	0.00
BOP noncitizen	Noncitizen	<0.01
USMS noncitizen	Noncitizen	<0.01
Nebraska Driver's License noncitizen	Noncitizen	<0.01
South Dakota Driver's License noncitizen	Noncitizen	0.00
SNAP/TANF noncitizen	Noncitizen	<0.01
NCRP noncitizen	Noncitizen	0.00
2005-2017 ACS noncitizen	Noncitizen	<0.01
AHS noncitizen	Noncitizen	<0.01
CPS noncitizen	Noncitizen	<0.01
SIPP noncitizen	Noncitizen	0.00
If none of the above criteria apply:		
ADIS citizen	Citizen	0.00
BOP citizen	Citizen	<0.01
USMS citizen	Citizen	<0.01
Nebraska Driver's License citizen	Citizen	<0.01
South Dakota Driver's License citizen	Citizen	<0.01
SNAP/TANF citizen	Citizen	<0.01
NCRP citizen	Citizen	<0.01
2005-2017 ACS citizen	Citizen	0.02
AHS citizen	Citizen	<0.01
CPS citizen	Citizen	<0.01
SIPP citizen	Citizen	<0.01
No Business Rules assignment	Model	9.12
Total		100.00

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.7 Comparison of Business Rules and As-Reported 2018 ACS Responses

	Percent Citizens	Percent Disagreement with As-Reported 2018 ACS Citizenship	Correlation with As-Reported 2018 ACS Citizenship	Number of Observations with Business Rules
	Weighted			
All Records, Citizen Values Prioritized	93.75	1.375	0.8833	217,800,000
Primary, then Secondary	93.68	1.336	0.8868	217,800,000
Primary Only	93.69	1.335	0.8865	217,700,000
Decision Tree Thresholds	93.74	1.329	0.8826	217,400,000
Primary, RLQ Thresholds	93.56	1.447	0.8772	216,600,000
Strict Pruning Rules	93.69	1.325	0.8873	217,700,000
Recommended Rules	93.69	1.325	0.8873	217,700,000
As-Reported 2018 ACS	93.68	NA	1.000	
	Unweighted			
All Records, Citizen Values Prioritized	95.54	0.9511	0.8892	3,462,000
Primary, then Secondary	95.49	0.9250	0.8924	3,461,000
Primary Only	95.50	0.9242	0.8917	3,459,000
Decision Tree Thresholds	95.55	0.9176	0.8862	3,455,000
Primary, RLQ Thresholds	95.38	1.025	0.8802	3,441,000
Strict Pruning Rules	95.50	0.9140	0.8930	3,459,000
Recommended Rules	95.50	0.9144	0.8930	3,461,000
As-Reported 2018 ACS	95.47	NA	1.000	

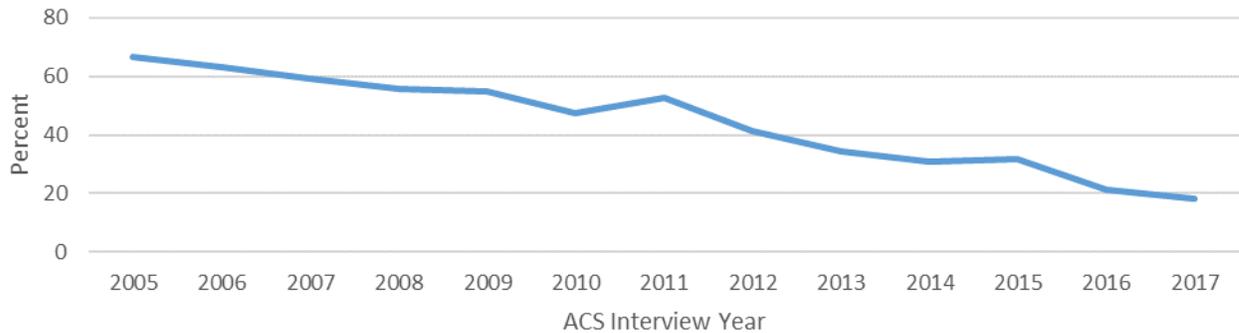
Note: The percent disagreement statistics use the number of observations in the last column. For the percent citizens and correlations statistics, the number of observations is 217,800,000 using ACS person weights and 3,462,000 without weights. Model estimates from the BR model trained on 2013-2017 ACS data (see Section 8) are used for cases with no business rules citizenship due to blanking out a source-citizenship combination. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 4.8 Citizen Shares with Different Thresholds of Business Rules-ACS Linkage Probability

Threshold	Percent Citizens
> 25% RLQ	93.50
> 20% RLQ	93.50
> 15% RLQ	93.50
> 10% RLQ	93.70
> 5% RLQ	93.72
All Business Rules	93.70
As-Reported 2018 ACS	93.69

Notes: The number of observations is 3,461,000. These percentages use ACS person weights. The sample is all persons age 18 and over with 2018 business rules citizenship without a threshold and as-reported 2018 ACS citizenship. The values where the business rules are dropped due to the record linkage probability threshold come from BR model estimates using the 2013-2017 ACS as a training sample (see Section 8). The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 4.1. Percent ACS Noncitizens That Are 2018 Benchmark Citizens, By ACS Interview Year



Notes: The number of observations is 21,000. No survey weights are used. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

5. Four Approaches to Statistical Estimation of CVAP Modeled Cases using Multiple Sources

The business rules (BR) discussed in Section 4 provide a comprehensive start towards producing estimates of citizenship for the full voting-age population by assigning citizenship to 91 percent of cases in the test with 2010 Census data, as well as in the test with 2018 ACS data. This leaves the task of predicting citizenship for the remaining 9 percent of cases. (These percentages could be different for the 2020 Census.) One approach to addressing this task would be to use the cases with observed citizenship (i.e., the BR cases) in the same manner that other Census Bureau survey data products do: as donors for imputing citizenship for cases where it is not observed (i.e., the non-BR cases) based on a feasible ignorable missing data mechanism that conditions on frame and response data for the particular survey.³⁶ Such imputation models are usually based on a procedure known as hot deck imputation (Little and Rubin, 2002, page 60). Rather than using hot deck imputation of citizenship from BR cases, a logical generalization is to fit a statistical model to the BR cases and use that model to predict citizenship status for the non-BR (NBR) cases. A further generalization is to fit a statistical model to one or more of the additional data sources and use that model to predict citizenship status for the NBR cases.³⁷

The Technical Working Group (TWG) pursued four approaches to get beyond BRs with statistical estimation to cover the NBR cases. These were:

- Hot Deck – impute citizenship status of the NBR cases using donors from the BR cases,
- BR logistic – predict probabilities of citizenship status for the NBR cases using logistic regression models fitted to the BR cases,
- ACS logistic – predict probabilities of citizenship for the NBR cases using logistic regression models fitted to ACS records that could not be given BR citizenship assignments, but that did have citizenship reported to ACS.
- Latent Class (LC) modeling, which does not use BRs, but instead uses a multivariate model to combine information from multiple citizenship data sources to produce predicted probabilities of citizenship for all person records.

The Hot Deck approach examined here is very similar to the standard imputation scheme that will be used for missing person characteristics in the 2020 Census, but with imputation cells defined specifically for the purpose of imputing citizenship status. The other three approaches provide more flexibility in regard to how the available information is used to form predictions of citizenship status since they are not constrained by the formation of imputation cells, but can

³⁶ A missing data mechanism is ignorable for likelihood-based inference if the probability density (or mass) function of missingness depends only on the observable response and frame data in the survey; the joint probability density (or mass) function of missingness and the observed response data can be factored into the probability density (or mass) of missingness conditional on the observed response and frame data times the probability density (or mass) of the observed response data conditional on the frame data; and the two conditional probability distributions depend on distinct, independent parameters. (Little and Rubin, 2002, page 119).

³⁷ We call such models (feasible) non-ignorable missing data mechanisms. They are also called selection-bias models. (Little and Rubin, 2002, chapter 15).

incorporate covariates in a prediction model with or without including their interaction terms. The motivation for the ACS logistic approach is that, by developing predictors of citizenship probabilities for the census NBR cases based on data from the ACS NBR cases, it seeks to address potential bias that could arise for the Hot Deck and BR Logistic approaches should their assumption that the BR cases are like the NBR cases fail to hold. This would reflect a type of feasible non-ignorable missing data mechanism. Finally, the LC model does not use explicit BRs, but forms predictions for all the cases. The LC model is a more general form that affords additional flexibility in regard to how the citizenship data sources are used to form the predictors of citizenship status.

The predictions of citizenship status from the BRs are in the form of designating person records as citizens or noncitizens, which creates a citizenship variable that equals 1 for citizens and 0 for noncitizens. Summing this variable gives the estimated number of citizens among the BR cases. The Hot Deck approach extends the citizenship variable by assigning 0s and 1s to the NBR cases by hot deck imputation. Summing the citizenship variable over all cases provides the estimate of total citizens. The BR logistic and ACS logistic approaches also assign 0 or 1 to the BR cases, but not to the NBR cases. Instead, the logistic regression models predict probabilities of citizenship that are between 0 and 1. Taking the 0s and 1s from their BR assignments as also representing predictions of citizenship probabilities, the predictions of total citizens from these approaches are obtained by summing over the probabilities of citizenship for all cases, which is an expected value calculation. The LC model provides predicted probabilities of citizenship for all cases, using no BRs, and similarly predicts total citizens by summing over all these predicted probabilities.

Sections 6 – 9 that follow describe the four estimation approaches and provide some results specific to their application. Section 10 then compares results of estimated citizen percentages from the four approaches. Results are provided for two tests: one using the 2010 Census data as the population frame and one using 2018 ACS data as the population frame. Using the 2010 Census data as the frame has the advantage that, as a census, it is more comparable to the 2020 Census (which will provide the frame for the 2020 CVAP estimates) than is the ACS. In addition, it provides much more geographically detailed data, and the data does not require sample weighting to make it representative of the full population. Using the 2018 ACS has the advantage that the corresponding 2018 administrative records files used are more recent, and so are more comparable to the 2020 administrative records files that will be used in producing the 2020 CVAP estimates.

Estimation results presented in Sections 6 – 10 are given for various subgroups of the total population, notably for the 12 race/ethnicity groups. Another important breakdown separates results for the cases that received BR citizenship versus those that did not. A closely related breakdown is to separate the cases that received a PIK, and so could potentially be linked to other data sources with citizenship information, versus those without a PIK. Combining these two concepts led to a four group breakdown found useful both in developing estimation models and in examining estimation results. These four groups are:

- BR cases: those cases that were assigned citizenship status according to the BRs.
- NBR-PIK: cases that received a PIK and so could potentially be linked to citizenship data sources, but either no links were found or, if link(s) were found, they did not determine the person's citizenship status.
- NBR-SS (sent to search): cases with sufficient information reported so an attempt was made to assign them a PIK, though none was obtained so BR citizenship could not be assigned.
- NBR-NSS (not sent to search): cases without sufficient information reported to attempt to assign them a PIK; hence no BR citizenship could be assigned.

As will be seen in Section 10, while the first three estimation approaches used slightly different versions of the BRs, they produced very similar results. The LC modeling also produced very similar results for the BR cases. Thus, with 91 percent of the cases assigned BR citizenship in our tests, material differences in the estimation approaches affect only 9 percent of the frame that comprises the three NBR groups. However, the NBR-PIK group is very small and so has inconsequential effects on the citizenship estimates for the total population aggregated across the four groups. Therefore, much of the focus in developing the estimation approaches and in comparing their results has been on the NBR-SS and NBR-NSS groups. Results will be presented for each of the four groups, as well as for the total voting-age population aggregating across the four groups. Results for the total are dominated by results for the BR cases since these represent 91 percent of the total voting age population.

The fact that the LC approach produced estimates for the BR cases that were very close to those from the BR assignments of the other three approaches provides strong confirmation for the BRs and is reassuring for the LC approach. However, while the LC modeling has some advantages compared to the other three approaches, certain effects found in the logistic regression modeling for detailed population subgroups could not be fully replicated in the LC model without enhancements to the model that required enhancements to the computer software for fitting it. While intensive work has been done on the software enhancements, they are not complete as of this writing, and this work is ongoing.

Research on developing and studying these alternative estimation approaches has so far focused on their point estimates (predictions) of citizenship and not on providing corresponding measures of their statistical uncertainty. Thus, no measures of uncertainty accompany the empirical results presented here. We intend to explore measures of statistical uncertainty for the approaches and would note that the LC approach provides these as a matter of course.

6. Business Rules Plus Nearest Neighbor Imputation Methodology

This section documents a deterministic approach to producing the Citizen Voting Age Population (CVAP) estimates using a hot deck nearest neighbor imputation methodology. This approach is split into two stages. The first stage uses business rules to assign citizenship for people with a PIK and a valid citizenship value on a selected set of administrative record sources.³⁸ The second stage uses a nearest neighbor approach to impute a yes or no citizenship value for the remaining cases. The remaining cases are split into two classes, within which the imputation is done separately: 1) individuals for whom a PIK was assigned but for whom no citizenship value was obtained from the administrative record sources, and 2) individuals for whom a PIK could not be assigned, so no administrative record source can be used. We first apply this method to voting-age persons in the 2010 CEF.

Business rules are applied in a hierarchical manner to the administrative records in the listed order shown in Table 6.1 and cycled through two times. The first iteration checks whether any source indicates citizenship. If so, the individual is assigned as a citizen. The second iteration checks whether any source indicates noncitizen for the cases not assigned citizen in the first iteration. If so, the individual is assigned as a noncitizen. Table 6.1 includes the assigned citizenship value, the number of persons assigned by that rule, and its share of the voting-age population distribution. Note that the Census NUMIDENT accounts for over 85% of the overall distribution from Rules 1, 2, and 5. The remaining 21.5 million unresolved cases must be imputed.

Table 6.1: Business Rules Used

Rule	Description	Citizen – (Y/N)	N	Percent
1	2010 Census NUMIDENT – citizen indication	Y	143,100,000	61.0%
2	2010 Census NUMIDENT – noncitizen indication, no indication of foreign birth	Y	46,480,000	19.8%
3	Passport Indication	Y	6,833,000	2.9%
4	USCIS file – citizen indication	Y	932,000	0.4%
5	2010 Census NUMIDENT – noncitizen indication	N	11,750,000	5.0%
6	USCIS file – noncitizen indication	N	437,000	0.2%
7	Presence on ITIN file	N	3,529,000	1.5%
	Imputation		21,500,000	9.2%
	Total		234,600,000	100.0%

Note: The population frame is persons age 18 and over in the 2010 CEF. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

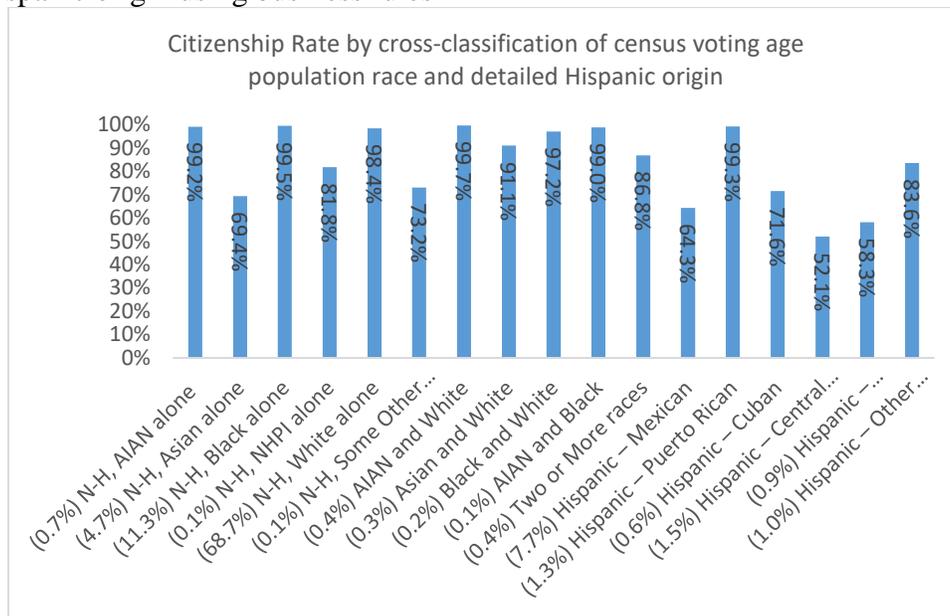
After the business rules have been exhausted, there are 21.5 million cases (9.2%) in need of imputation. In step two, the goal is to assign a citizenship value from someone in a local geography sharing the same characteristics. With respect to the geography, within each state the file is sorted by Local Census Office (LCO), census tract, census block, and sequence number within the census

³⁸ The first stage uses business rules based on the following administrative record sources: Census NUMIDENT, State Department Passport data, Department of Homeland Security U.S. Customs and Immigration Services (USCIS), and Individual Taxpayer Identification Number (ITIN) roster.

block. With respect to characteristics, we partition the universe by three variables. First, we split the universe by the cross-classification of census voting-age population, race and ethnicity. This creates seventeen disjoint groups. The single race groups include non-Hispanic American Indian and Alaska Native (AIAN) alone, Asian alone, Black or African American alone,³⁹ Native Hawaiian or Other Pacific Islander (NHOPI) alone, White alone, and some Other Race alone. Multiple race groups used include non-Hispanic AIAN and White, Asian and White, Black and White, AIAN and Black, and Two or More races. The remaining groups are detailed Hispanic origin categories: Mexican, Puerto Rican, Cuban, Central American/Dominican Republic, Latin/South American, and Other Hispanic.

We include detailed Hispanic origin as a further breakdown of Hispanic or Latino because of the differential citizenship rates among the Hispanic subgroups.⁴⁰ Figure 6.1 shows the citizenship rates among persons where business rules could be applied by cross-classification of census voting-age population, race and detailed Hispanic origin. Notice that Puerto Rican Hispanics have a 99% citizenship rate when applying the business rules as opposed to a 52% citizenship rate for Central American Hispanics.

Figure 6.1: Citizenship rates by cross-classification of census voting age population race and detailed Hispanic origin using business rules



Source: 2010 Census Edited File linked to multiple administrative records

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

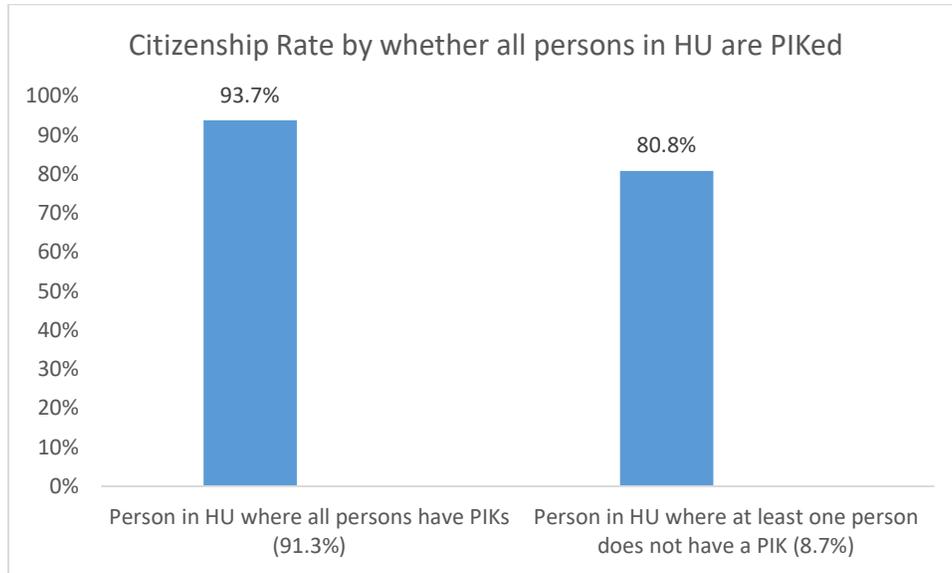
Second, we split the universe by whether or not the housing unit (HU) has a person without a Protected Identification Key (PIK) within the unit. These are mixed households in which persons with resolved citizenship that link to administrative records live with persons not linking to administrative records. All persons in the household are then given the same value indicating that there is at least one person in the household that does not link to administrative data. The resolved

³⁹ The terms “Black” and “Black or African American” are used interchangeably in this report.

⁴⁰ The terms “Hispanic” and “Hispanic or Latino” are used interchangeably in this report.

persons can then serve as donors in imputation cells for persons who do not link to administrative records. Figure 6.2 shows the citizenship rates using business rules among the persons in housing units (HUs) where all persons are PIKed versus persons in HUs where at least one person is not PIKed. Notice that the citizenship rate when using business rules is about thirteen percentage points higher for persons in units where all persons were PIKed compared to those where least one person was not PIKed.

Figure 6.2: Citizenship rates using business rules by whether all persons in housing unit (HU) are PIKed

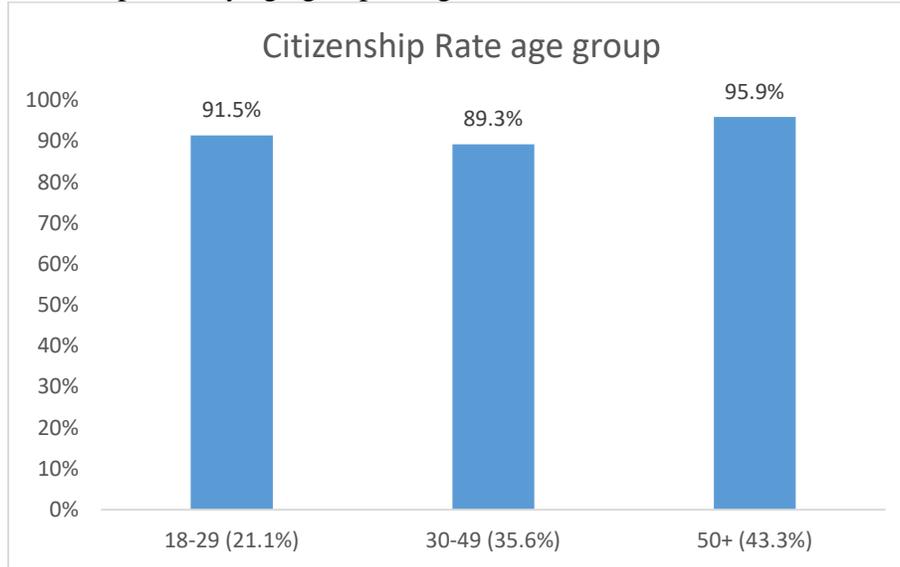


Source: 2010 Census Edited File linked to multiple administrative records

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Last, we split the universe by three age groups: 18-29, 30-49, and 50+. Figure 6.3 shows the citizenship rates among the age groups where citizenship could be applied by business rules.

Figure 6.3: Citizenship rates by age group using business rules



Source: 2010 Census Edited File linked to multiple administrative records

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Using the three variables discussed above produces 102 imputation cells, which were developed separately for each state. We searched for donors conditional on the imputation cell. Thus, for each unresolved person, the nearest resolved person preceding it on the sorted file, and within the imputation cell, is the donor. The search proceeds in a circular manner. Hence, if there is not a donor preceding the unresolved person in the same tract and imputation cell, the search continues at the bottom of the list within the same tract and cell. If there is not a donor within the tract, the search extends to the persons in the same LCO in the cell in a similar manner, starting with the preceding tract in the same LCO. If there is not a donor within the same LCO, the search extends to the persons within the same state and cell. There is no limit on the number of times the same donor can be used.

Modifications for Testing with 2018 ACS Data

The same method was applied to the 2018 ACS one-year data. We built a collapsing routine into the technique. The collapsing routine used the twelve-group census voting-age population race as the cell to search for donors, when the 102 cell method did not have any donors. This allowed searching across states for donors. Results are reported in Section 10.

7. Business Rules plus Householder Logistic Regression Approach

This approach has two components. The first is a business rules-based assignment of citizenship status for each person for which this is possible. For the applications discussed in Section 10, this approach used a set of business rules (discussed below) that differ from those discussed in Sections 4 and 6, though it could be used with any such alternative rules. These business rules also used a measure of record linkage quality that is discussed in Section 3, and that differs from the measure used by the approach discussed in Section 8. Comparison of results for the different versions of the business rules are given in Section 10.

The second component of this method predicts probabilities of citizenship for the cases not covered by the business rules. This is done using a set of logistic regression models that are fitted to the cases whose citizenship status was assigned by the business rules from the first component. These logistic regression models are then used to predict citizenship probabilities for the cases not assigned citizenship status by the business rules. For the household population, this approach uses one logistic regression model to predict the citizenship probabilities of householders whose citizenship status was not assigned by the business rules. It then uses two additional logistic regression models to predict the citizenship probabilities of the remaining people in households. One logistic regression model applies to cases where the householder is a citizen, and another to cases where the householder is a noncitizen. Model fitting results showed differences across the three models, suggesting better predictions would be obtained from the three-model approach than by using a single model. A separate, fourth, logistic regression model applies to the group quarters population, for which there are no householders.

The analysis done with the 2010 CEF omitted certain administrative sources because their PVS record linkage module, passes, and scores were not available to be used in the record linkage error modeling.⁴¹ Analyses done with the 2018 ACS data used a more expanded set of administrative data sources. In both cases, the tabulations done to produce estimates of citizens from the business rules assignments and model predictions of citizenship probabilities are over person records with an edited and imputed age greater than or equal to 18. Since the modeling conditions on the citizenship of the householder, householders age less than 18 were included in the modeling, but excluded from the citizen tabulations.

⁴¹ For analysis with the 2010 CEF data, the following administrative record sources were used: Social Security Administration 2018 Census NUMIDENT, United States Citizenship and Immigration Services (USCIS) Citizenship and Immigration Data, State Department Passport, Internal Revenue Service Individual Tax Identification Number (ITIN), Bureau of Prisons (BOP) 2019 Master Prison File, United States Marshal Service (USMS) 2010-2019 Received and Custody Files, Department of Interior 2019 Incident Management Analysis and Reporting System (IMARS), American Community Survey (ACS), American Housing Survey (AHS), Current Population Survey (CPS), and the Survey of Income and Program Participation (SIPP). The sources that were not used because the PVS record linkage module, passes and scores were not available to be used in the record linkage error modeling were: State of Nebraska Department of Motor Vehicles, Supplemental Nutrition Assistance Program (SNAP) for Idaho, Mississippi, Wyoming and New York, Temporary Assistance for Needy Families (TANF) for Idaho and Mississippi.

7.1 Business Rule-Based Assignment of Citizenship Status

This method assigns a rule-based citizenship status accounting for potential record linkage error for each source in the hierarchical rules listed in Table 7.1.⁴² The rules used for USCIS, BOP and USMS can be applied to other sources using EPIKs when available or can be applied to sources not used in this analysis, since they did not have data available until 2013.

To link a PIK or an EPIK to a source used in the assignments below, the record linkage quality index needed to meet a minimum threshold value of c . This threshold value can be adjusted. For the analysis with the 2010 CEF data, we set $c = .99$.

The first step in the business rules was to determine, whenever possible, a citizenship status for each record in each source. These preliminary determinations did not determine the final citizenship status for those individuals with information on citizenship from multiple sources that did not agree. The initial citizenship determinations by source are given in Table 7.1.

Table 7.1. Initial Citizenship Assignments by Data Source (Business Rules)

NUMIDENT	<p>If the NUMIDENT citizen variable indicates the person is a citizen, or if the person is native born, then the person is a citizen.</p> <p>If the NUMIDENT citizen variable indicates the person is not a citizen, or if the NUMIDENT indicates the person is foreign-born, then the person is a noncitizen.</p>
USCIS, Bureau of Prisons, U.S. Marshalls Services, U.S. Passports, IMARS, ACS, AHS, CPS, SIPP	If the source indicates the person is a citizen (a noncitizen) and the record linkage quality of the PIK or EPIK assignment is above the cutoff, then the person is a citizen (a noncitizen).
Individual Taxpayer Identification Number	If the 2010 Census Edited File (CEF) record has a PIK in the ITIN range, then the person is a noncitizen.

Based on the assignment of citizenship status to individual sources for each person, an overall citizenship status was assigned to each person using the following hierarchy:

⁴² Note that only USCIS Passport, Bureau of Prisons and USMS have EPIKs, so those are the only rules that incorporate EPIK record linkage error.

- If any of the sources indicated that the person was a citizen then the person was classified as a citizen,
- Otherwise, if any of the sources indicated that the person was not a citizen then the person was classified as a noncitizen, and
- The remaining people had unresolved citizenship status.

7.2 Logistic Regression Models to Predict Probabilities of Citizenship for Individuals with Unresolved Citizenship Status from the Business Rules

For this second component, we ran a set of logistic regression models separately for each state, to predict citizenship probabilities for the unresolved cases for the household and group quarter populations.

For the household population, logistic regressions were implemented for the unresolved household population in three steps:

1. We estimated logistic regression models for the householders with citizenship assigned from the business rules using tabulated geographic tract, race, ethnicity, and age domains (under 29, 30 to 49 and 50+) as the main effects. Then, for householder records without citizenship assigned from the business rules, a predicted probability is assigned from this model. The model fitting and prediction includes records for householders under age 18 since some of these are missing assigned citizenship status. Predicted probabilities of citizenship are needed for under-18 householders with other persons in the household to calculate predicted citizenship probabilities for these other persons – see below.
2. We estimated logistic regression models for the other household members where the householder is a citizen. Model fitting results show that the other members of the household (spouse, children, grandchildren, etc.) have different citizenship probabilities if the householder is a citizen as compared to a noncitizen. This model uses 11 collapsed relationship to householder categories, race, ethnicity, and age domains. Thus, the other household members are either assigned citizenship status by the business rules or receive, from this fitted model, a predicted probability of their being a citizen given that the householder is a citizen.
3. We estimated similar logistic regressions for the other household members, where the householder is a noncitizen. The same independent variables are used as for the model where the householder is a citizen. This model predicts the probabilities of the other household members being a citizen given that the householder is a noncitizen for the other household members who are not assigned citizenship status by the business rules.

Table 7.2 shows citizenship rates estimated from the 2010 CEF business rules cases for other members of the household according to whether the householder is a citizen or not. These results motivate the conditional structure of the model above with the separate logistic regression models used at Steps 2 and 3. Note from Table 7.2 that when the householder is a citizen, all but father/mother, parent-in-law, and other relatives have citizenship rates above 90 percent. When the householder is not a citizen, the results show lower citizenship percentages and more variation. Grandchildren are citizens at a rate of 75 percent, and biological, adopted, or stepchildren have

citizenship percentages between 52 and 59, while the remaining relationship categories have citizenship percentages less than 50 percent.

Table 7.2: Citizenship percentages of other household members by citizenship status of the householder; 2010 CEF cases with business rule citizenship assignments

	Percent citizens when the householder is a citizen	Percent citizens when the householder is a noncitizen
Spouse	97	31
Biological Son/Daughter	99	52
Adopted Son/Daughter	97	53
Stepson/Stepdaughter	96	59
Brother/Sister	94	18
Father/Mother	87	25
Grandchild	99	75
Parent-in-law	82	32
Son/Daughter-in-law	90	46
Other relative	88	24
Roomer, Boarder	93	30
Housemate, Roommate	96	27
Unmarried Partner	97	35
Other Non-Relative	94	33

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

At the end of the three steps above, all 2010 CEF persons received a probability of being a citizen in the following way:

- Householder: Probability of being a citizen is either assigned based on the business rules (as 0 or 1) or is predicted from the logistic regression in Step 1 above.
- Other household members: The probability of their being a citizen is either assigned from the business rules (as 0 or 1) or is calculated as follows:

$$\begin{aligned} \text{Pr}(\text{citizen}) &= \text{Pr}(\text{HH is citizen}) \times (\text{Prob Other is Citizen}|\text{HH is citizen}) \\ &+ [1 - \text{Pr}(\text{HH is citizen})] \times (\text{Prob Other is Citizen}|\text{HH is noncitizen}) \end{aligned}$$

We implemented a separate logistic regression model to predict citizenship probabilities for the group quarters residents not assigned citizenship status by the business rules. The model was estimated using the data from the group quarters resident with assigned citizenship status. Due to the small sample sizes within some group quarters types and race/ethnicity categories, group quarter type was collapsed down to two levels: institutional and non-institutional. For race and ethnicity, reporting categories with less than 100 people were collapsed together. The logistic

regression model used the two-level group quarters type, the collapsed race and ethnicity categories, and age groupings (18-29, 30-40, and 50+) as main effects in the model.

7.3 Modifications for Testing with the 2018 ACS Data

For the testing with the 2018 ACS data, the approach discussed above was not applied separately for each state, but instead applied in one national run, due to the more limited data being used compared to testing done with the 2010 CEF data. The business rules assignment was modified to include additional administrative record sources available for 2018. These included available ADIS, SEVIS, WRAPS, drivers' license records, and SNAP/TANF data. The decision tree methodology was expanded to include these additional sources and their PIK/EPIK assignments.

Business rules similar to those of Table 7.1 above were then implemented to assign citizenship status whenever possible for each record in each new data source.⁴³ An overall citizenship status was then assigned to the 2018 ACS records when possible in the same way that was described for the analysis with the 2010 CEF data. Running this test on the 2018 ACS one-year sample showed that the business rules assignment and the decision tree methodology can be expanded to account for more data sources. For the 2020 processing, the expansion of the business rules and the decision tree methodology to include the other sources could be done when processing each state.

For the cases not assigned citizenship by the business rules, the logistic regression modeling applied with the 2010 CEF data as described in the previous section used fixed effects for tracts. Modeling attempts with the more limited 2018 ACS data found it was unable to support estimation of models with fixed effects for tracts. The householder logistic regression was thus changed to use county fixed effects instead of tract fixed effects in a national-level run. This was also done to provide estimates for comparison with other methods. Changing from tract-level to county-level main effects is a big factor affecting any differences when comparing this approach to the others for the 2018 analysis results. Since tract-level main effects would be used if this methodology were applied to 2020 data, the differences seen between results from this and the other approaches with 2018 one-year ACS data are not as meaningful as those seen from using the 2010 CEF and related data.

⁴³ The approach to constructing the business rules is the same as for the 2010 CEF, but additional sources available only after 2010 are included here.

8. Business Rules and ACS Logistic Regression Approach

As was the case for the approach of Section 7, this approach accepts business rules citizenship assignments for the records that have them, and then uses predicted probabilities from logistic regression models for the remainder. Differing from the approach presented in Section 7, the models here are trained on past ACS data. This section provides motivation for the model design, describes the model estimation in detail, and provides descriptive statistics from the models.

8.1 Motivation

The people without business rules (BR) citizenship information can be divided into three groups.⁴⁴ The first group is persons with Protected Identification Keys (PIKs) but no BR (NBR-PIK). Most of these are foreign-born in the NUMIDENT, but their citizenship variable is missing. This is because they applied for an SSN before May 1981, and they have not updated their information with SSA. These people are generally long-term U.S. residents. They may have obtained a U.S. passport prior to 1978 (the first year of our passport data) and not renewed it since or could have a U.S. Customs and Immigration Services (USCIS) naturalization certificate issued many years ago before the naturalization data coverage was complete.

The second group is persons without PIKs who were sent to the Person Identification Validation System (PVS) search (NBR-SS). A record could fail the PVS search due to discrepancies in how the Personally Identifiable Information (PII) is reported in the population frame survey versus other survey or administrative data, use of a different address, having a common name, or because the person is genuinely absent from the NUMIDENT and other PVS reference files. Persons in the latter category are highly likely to be noncitizens, since all citizens are eligible for a Social Security Number (SSN). Both citizens and noncitizens could have discrepant PII.

The third group is persons without PIKs who were not sent to PVS search due to insufficient PII (NBR-NSS). Insufficient PII could occur if the respondent has confidentiality concerns, e.g., that the data may be used for individually targeted law enforcement. PII-deficient cases also often arise in the census when the respondent is a proxy, and the proxy does not know or will not report the neighbor's PII. There are also census count imputations and substitutions, which are person records that lacked any PII because all characteristics were imputed.

Table 8.1 shows that the as-reported 2018 ACS citizen share is much lower in the NBR-SS group than the others. The NBR-PIK and NBR-NSS groups also have lower shares than the BR group, but the differences are not nearly as pronounced. The citizen share difference between the BR and NBR-SS groups is quite large for Hispanics, in particular, when measured by percentage point difference, and it is also significant for non-Hispanic Asians. Though the percentage point difference is small for non-Hispanic Whites and Black or African Americans, the White NBR-SS noncitizen share is more than three times as high as the BR share, and it is 89 percent higher for Black or African Americans.

⁴⁴ Hereafter we refer to the BR group and these three no-business rules (NBR) groups as BR/PVS groups.

Table 8.1 As-Reported 2018 ACS Percent Citizens

Category	Total	BR	NBR-PIK	NBR-SS	NBR-NSS
Total	92.21	93.68 (91.73)	91.53 (0.05)	71.06 (6.27)	90.90 (1.95)
NH Asian	68.99	70.63 (90.48)	73.76 (0.07)	55.19 (7.50)	45.60 (1.94)
Hispanic	72.61	78.29 (83.93)	66.35 (0.04)	39.56 (14.56)	75.86 (1.47)
NH White	98.43	98.60 (94.41)	98.71 (0.05)	95.33 (3.71)	96.11 (1.84)
NH Black or African American	95.67	95.94 (89.03)	D (0.05)	92.34 (7.90)	96.53 (3.02)

Notes: The sample is persons age 18 and over with as-reported 2018 ACS citizenship. The column group’s percentage of the row sample observations is in parentheses. The results use ACS person weights. “D” indicates that the cell is suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

In the 2018 ACS sample with as-reported citizenship, the NBR-SS group is the largest among those without BR, and the NBR-PIK group size is negligible. Higher shares of Non-Hispanic Whites have BR compared to Hispanics (94.4 versus 83.9 percent, respectively), while Hispanics have much larger NBR-SS shares (14.6 versus 3.7 percent). Non-Hispanic Black or African Americans have the greatest share of NBR-NSS cases.

The table strongly suggests that BR information is not missing at random with respect to citizenship status, and that the nonrandom nature varies by missingness category and race/ethnicity. The ACS logistic approach is designed to address this. That is, these data are inconsistent with an ignorable missing data mechanism (Little and Rubin, 2002, page 119).

One potential concern is that differences in the survey design and protocols for the ACS versus the 2020 Census may result in incompatible BR/PVS groups across the two surveys. For example, the census includes proxy responses and whole person imputations, while the ACS does not. Proxy respondents may not know their neighbor’s PII, and whole household imputation cases don’t have PII. This could lead to a larger share of NBR-NSS cases in the census. Table 8.2 shows the distribution across BR/PVS groups in the 2010 CEF and 2010-2012 ACS. The NBR-NSS group is many times larger in the census than the ACS.⁴⁵ The BR group is 1.8 percentage points smaller, and the NBR-SS group is 1.4 percentage points smaller in the census. The NBR-PIK group size is similar in the two. Assuming that the ACS is a representative sample of the population, it appears

⁴⁵ It is much larger in the 2018 ACS than the 2010-2012 ACS, though, so the incidence of not providing PII may be increasing.

that the 2010 CEF NBR-NSS group is made up of a comparable proportion of persons who would be BR versus NBR-SS in the ACS. The relative declines of the BR and NBR-SS groups (comparing ACS to census) vary by race/ethnicity. The NBR-SS decline is larger than the BR decline for non-Hispanic Asians, but the BR drop is much bigger than the NBR-SS drop for non-Hispanic Black or African Americans. Thus, the composition of the NBR-NSS group appears to differ across race/ethnic groups. We will return to this issue in Section 8.3.

Table 8.2 PVS/BR Group Shares by Race/Ethnicity

Category	BR	NBR-PIK	NBR-SS	NBR-NSS
	2010 CEF			
Total	90.85	0.12	5.76	3.27
NH Asian Alone	89.09	0.22	7.23	3.47
Hispanic	83.20	0.17	12.83	3.79
NH White Alone	93.19	0.12	3.91	2.79
NH Black or African American	87.95	0.04	6.85	5.16
	2010-2012 ACS			
Total	92.68	0.10	7.16	0.06
NH Asian Alone	90.41	0.11	9.45	0.04
Hispanic	85.17	0.14	14.66	0.04
NH White Alone	94.62	0.10	5.21	0.07
NH Black or African American	91.93	0.02	7.99	0.06

Notes: The 2010-2012 ACS estimates use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

8.2 ACS Logistic regressions

A crucial question is which people with observed citizenship status are most similar to the people in the three groups above without BR, after controlling for other observable characteristics such as race/ethnicity and location. One option is to try to select persons with BR who have otherwise similar observable characteristics to persons without BR, using data from the same population frame. We previously explored this using 2009 ACS data to estimate a model predicting whether a person had BR or not. We applied the estimated coefficients to obtain the person's predicted probability of having BR. We then estimated three citizenship models with ACS citizenship as the dependent variable: one with all observations with BR, a second with observations without BR, and a third with BR observations having low BR predicted probabilities. This last group was, by construction, similar to the group with no BR according to other demographic characteristics. We then applied the coefficients from these three models to obtain alternative citizenship probabilities for observations without BR. The citizenship probabilities from the model using all BR were the highest on average, followed by those from the model using the BR group with low BR

probabilities, and then the model using the group without BR.⁴⁶ The model using the group without BR produced estimates much closer to the ACS responses. Thus, we were unable to account for non-ignorable missingness when using BR citizenship cases to train the models.

An alternative way to address non-ignorable missingness is to borrow from another dataset where citizenship information is observed for persons in the non-BR groups. The ACS provides such a data source. We thus estimate a logistic regression model on ACS data for each particular BR/PVS group and use the estimated model to predict citizenship for the persons in the same BR/PVS group in the population frame being used for the citizenship estimates.

2010 CEF Test Specifications

For the 2010 CEF test, we use the “Primary, Then Secondary” business rules described in Section 4.2.

The dependent variable in each logistic regression model is binary (citizen versus noncitizen) as reported ACS citizenship. The explanatory variables include age groups (18-29, 30-49, and 50-plus); a female indicator; indicators for the CVAP race/ethnicity categories other than non-Hispanic Asian and Hispanic, Hispanic origin subgroups (Mexican, Puerto Rican, Cuban, Central American, Latin American, and other Hispanic), and Asian subgroups (Asian Indian alone, Chinese alone, Filipino alone, Japanese (or Okinawan or Iwo Jiman) alone, Korean alone, Vietnamese alone, or other non-Hispanic Asian); an indicator for whether the home is owned with or without a mortgage versus rented or occupied without rent; household size group indicators (1, 2, 3, 4, or 5 or more); six relationship to the householder categories (householder, spouse/partner, child, all other relatives, unrelated individuals, and group quarters); the shares of other household members by citizenship/PVS category (BR citizen, BR noncitizen, NBR-PIK/SS (combining NBR-PIK and NBR-SS)⁴⁷, or NBR-NSS); the citizenship category of the householder interacted with being a relative of the householder and a non-relative of the householder; indicators for whether the language of the survey data is English, Spanish, other language, or missing (the missing category is included in the NBR-PIK and NBR-NSS regressions, while it is combined with English in the NBR-SS regression); the county ACS citizen share among those who are in the regression sample’s BR/PVS group; and state indicators. The NBR-SS and NBR-NSS regressions include age-state, tenure-state, age-tenure, and relationship category-tenure interactions. The NBR-SS regression also includes an indicator for records without a PIK that received multiple links to the reference files.

⁴⁶ The proportion of individuals with a predicted citizenship probability above 99% was 0.65 when using the coefficients from the model using all observations with BR citizenship, 0.56 from the model using those with BR citizenship information and low probabilities of having BR citizenship, and 0.48 from the model using those with no BR citizenship information.

⁴⁷ The NBR-PIK group has a very small number of observations in the 2005-2009 ACS, making it difficult to estimate a separate coefficient for it by itself.

The logistic regression models for the NBR-PIK and NBR-NSS groups use 2005-2009 ACS data. The logistic regressions for the BR and NBR-SS groups use 2010-2012 ACS data.⁴⁸ We do this for the NBR-SS group because the PVS process for the 2010 CUF and 2010-2012 ACS use 2011-2012-vintage reference files containing ITINs, while the reference files used by PVS for the 2005-2009 ACS do not.⁴⁹ Initially no ITIN records in the 2005-2009 ACS received PIKs. When the 2005-2009 ACS records in the NBR-SS group were reprocessed in the PVS using the 2011-2012-vintage reference files, about 10,000 cases received PIKs from the ITIN reference file. This is a quarter of the number of ITIN PIKs in the 2010 ACS alone, suggesting that most of the 2005-2009 ACS ITINs still do not have a PIK. This makes the 2005-2009 ACS sample for the NBR-SS group incompatible with the 2010 CEF sample for the NBR-SS group, because there is almost surely a higher share of ITINs in the former sample. That causes the citizen share of the former sample to be lower than that of the latter sample. If the model were trained on the 2005-2009 ACS sample and applied to the 2010 CEF sample, the citizenship imputations would be biased downward. If this method is used in 2020 Census production, the training data would be the 2015-2019 ACS, which used the same production PVS that is planned for the 2020 CUF, so these two samples for persons without business rules citizenship and sent to PVS search will be compatible.

We considered using a single year of ACS data to train the models (e.g., 2019 for the 2020 Census production). This would have the advantage of being from a period closer to the reference date for the statistics, which could be important if associations between citizenship and other observable characteristics change rapidly. Identical models using 2005-2009 and 2009 ACS samples produced very similar citizen shares for the 2010 CEF at the state and national levels, suggesting that there isn't a meaningful advantage to training the models on the single most recent year. We recommend using the five-year ACS sample (2015-2019 for 2020 production), because having more years of data in the training sample facilitates estimation of richer models.

2018 ACS Test Specifications

Several changes have been made to the specifications for the 2018 ACS test. We have added two variables with administrative record citizenship information for the housing unit. Using a large number of administrative sources,⁵⁰ we selected all person records associated with each housing unit in the ACS.⁵¹ We did so for each interview year and year prior to the interview, and we removed duplicate observations by PIK. For each unduplicated PIK, we then linked in a BR

⁴⁸ The area under the Receiver Operating Characteristic (ROC) curve goodness of fit statistic is 0.9607, 0.7610, 0.9442, and 0.9217 for the models using ACS BR, NBR-PIK, NBR-SS, and NBR-NSS plus initially insufficient partial response observations, respectively (DRB clearance number CBDRB-FY21-CED005-0001).

⁴⁹ We use the 2010-2012 ACS for the BR group, because we want it to use the same sample as the NBR-SS group in the decomposition exercise in Table 10.3.

⁵⁰ These include records from Internal Revenue Service 1040 returns and 1099 information returns; Medicare enrollment; Housing and Urban Development Public and Indian Housing Information Center (PIC) and Tenant Rental Assistance Certification System (TRACS); Federal Housing Authority; Selective Service System registrations; Indian Health Service Patient Registration System; the U.S. Postal Service National Change of Address file; Supplemental Nutrition Assistance Program (SNAP), Temporary Assistance for Needy Families (TANF), and Special Supplemental Nutrition Program for Women and Children (WIC) from several states; Nebraska and South Dakota driver's licenses; passports; USCIS; ADIS; SEVIS; WRAPS; BOP; USMS; and the Veteran Service Group of Illinois (VSGI).

⁵¹ In the 2013-17 ACS sample, 90.20% of all observations were in housing units which linked to administrative records. For the 2018 ACS sample, the coverage rate was 91.46%.

citizenship value. We used these BR citizenship values to construct the citizen share among the PIKs in the housing unit. We also constructed a variable measuring the share of PIKs in each housing unit with any valid BR citizenship assignment. The motivation for these variables is that the administrative data could be for the people without a PIK in the population frame, in which case the BR citizenship information would be theirs as well. These variables could be particularly important in cases where all the person's characteristics are imputed in the population frame. By including these housing unit administrative record variables, at least some variables in the model are based on actual data from the housing unit.

We separated the other Hispanic category into one where the person is in the other Hispanic category (e.g., from Spain) in the detailed Hispanic origin question versus answering that they are Hispanic without providing an answer to the detailed Hispanic origin question. We separated the non-Hispanic other Asian category in an analogous manner.

The 2018 ACS test uses the recommended business rules discussed in Section 4. A handful of records have citizenship information, but are not covered by the recommended business rules. For these cases we use a model trained on observations with both linkable citizenship information (whether in the recommended rules or not) and as-reported 2013-2017 ACS citizenship, where the dependent variable is the ACS citizenship response.

Since the NBR-NSS group in the ACS has a small number of observations, we have expanded the training sample to include ACS responses that were initially partial. The larger sample permits a richer model specification. The ACS contains some internet responses that did not continue after the household roster section, as well as blank paper questionnaire mail returns (where blank is defined as not having a data defined person). Some of these were sampled for Computer-Assisted Personal Interview (CAPI) follow-up. Of the ones sampled for CAPI follow-up, a sufficient partial or complete response was obtained for a subset of them. This could have been done through the field representative prompting the respondent to provide a late internet response or late mail return. Alternatively, the field representative obtained a sufficient partial or completed interview. These people are arguably reluctant responders like those in the NBR-NSS group. Of these cases, some have as-reported ACS citizenship.

The training samples for the 2018 ACS test all come from 2013-2017 ACS responses with as-reported citizenship.

Due to an even smaller number of observations in the 2013-2017 ACS compared to the 2005-2009 ACS for the NBR-PIK group, no interaction terms are included in that specification for the 2018 test.⁵²

8.3 Strengths and Weaknesses

The principle strengths of this approach are that its predictions are trained using data for persons with ACS citizenship information who are most similar to the persons in the population frame

⁵² The area under the ROC curve goodness of fit statistic is 0.9682, 0.7517, 0.9398, and 0.9509 for the models using ACS BR, NBR-PIK, NBR-SS, and NBR-NSS plus initially insufficient partial response observations, respectively (DRB clearance number CBDRB-FY21-CED005-0001).

without linkable citizenship information, and that it captures a number of key associations between citizenship and observable characteristics. The large and race/ethnic group-specific differences in citizen shares for persons with versus without linkable citizenship information suggest that these features are quite important for producing high-quality estimates.

The fact that all model estimation can occur before the 2020 CEF becomes available is an operational advantage. Once the CEF files arrive, the model coefficients just need to be applied to produce predictions for the CEF, saving time.

One weakness is that, unlike in the latent class model approach, citizenship source records used in the business rules are treated as truth, ignoring measurement error. A question to consider is how important this factor is for this use case. As shown in Sections 3 and 4, most of the citizenship sources do not appear to exhibit significant record linkage error, the bulk of the disagreements among the main sources were anticipated and have been addressed in the business rules, and the sources are updated through Census Day.

Another weakness is that the logistic regression imputations implicitly assume that the as-reported ACS citizenship responses are correct. Note, however, that there are no errors due to incorrect record linkage or out of date response information here, unlike when directly linking past citizenship information to the population frame for the business rules. The models use citizenship information and other characteristics collected at the same time for the same person, then apply the coefficients to a different population. No record linkage is involved. The remaining error is response error, which is unobservable.⁵³

Model coefficient sensitivity to compositional differences between the ACS and census BR/PVS groups could introduce error in the estimates. The characteristics of the ACS training samples and the census are likely to differ somewhat. Reluctant responders may be more likely to appear in the census than the ACS due to the census having a shorter questionnaire and a campaign to boost response. And some people are enumerated via neighbors' proxy responses in the census, but proxies are not used in the ACS. Proxy responses often produce NBR-NSS cases, due to lack of knowledge about the PII. Someone in the BR or NBR-SS group in the ACS may be in the NBR-NSS group in the census.

To address this compositional difference between the ACS and the census, we will also test an alternative approach for producing NBR-NSS group estimates. The idea is to train the citizenship model with all as-reported ACS citizenship cases regardless of their BR/PVS status. Variables for controlling for the probability of being in the 2020 Census NBR-NSS group would be included to make the coefficients more appropriate for 2020 NBR-NSS group.

If this approach were used in production, we would estimate a logistic regression model with being in the NBR-NSS group as the dependent variable. The independent variables could be similar to those used for the BR group model. These coefficients would be applied to the 2015-2019 ACS to

⁵³ We can study response error for persons with both BR and as-reported ACS citizenship, but not for persons without linked BR. There are likely to be systematic differences in ACS citizenship reporting error across the BR/PVS groups. For example, NBR-SS persons may have greater confidentiality concerns than BR persons, leading to more misreported citizen responses in the former group.

produce probabilities of being in the NBR-NSS group in 2020. The citizenship model would be trained on records with as-reported 2015-2019 ACS citizenship in all BR/PVS groups. Indicator variables for ranges of the NBR-NSS probability distribution would be included in the model as controls. Then the model estimates would be applied to 2020 CEF NBR-NSS records to produce the citizenship probabilities.

We plan two tests of this approach. One uses the 2010 CEF in place of the 2020 Census and the 2005-2009 ACS in place of the 2015-2019 ACS. We will compare the resulting citizenship estimates to the ones obtained using the model trained on 2005-2009 ACS NBR-NSS records described above. If the ACS and census truly have different NBR-NSS groups, and if this method works properly, the two sets of citizenship estimates should noticeably differ.

The second test uses the 2018 and 2013-2017 ACS to stand in for the 2020 Census and 2015-2019 ACS, respectively. Rather than using the 2018 ACS NBR-NSS group to stand in for the 2020 NBR-NSS group, we will use the 2018 ACS initially insufficient partial group, which is larger and may be more like the 2020 Census NBR-NSS group. These citizenship estimates will be compared to as-reported 2018 ACS citizenship, 2018 BR citizenship, and the estimates from the NBR-NSS model described above.

9. Latent-Class Modeling Method

In the approaches described thus far, citizenship status is assigned for the great majority of the census records based on business rules. For the hot-deck approach of Section 6, citizenship status is then imputed for the remaining records, whereas for the logistic regression approaches of Sections 7 and 8, the remaining records are given probabilities of citizenship predicted from logistic regression models. In this section, we present an alternative that makes no binary assignments of citizenship status, but synthesizes the available information from administrative and survey sources to assign probabilities of citizenship to every census person. The mathematical formula for computing these probabilities comes from multivariate statistical models for the variables measuring citizenship and the true citizenship status, where the true status is regarded as random and unknown. Results from this approach are illustrated here using 2010 data for the state of Delaware, and results for the full U.S. 2010 data are shown in Section 10. Due to the complexity of these models and the algorithms required to implement them, we are not recommending this approach for production of 2020 CVAP statistics. It may, however, play an important role in evaluation of those data products and provide a template for future programs that combine census and survey responses with information gleaned from administrative sources.

9.1 Latent-Class Regression Analysis

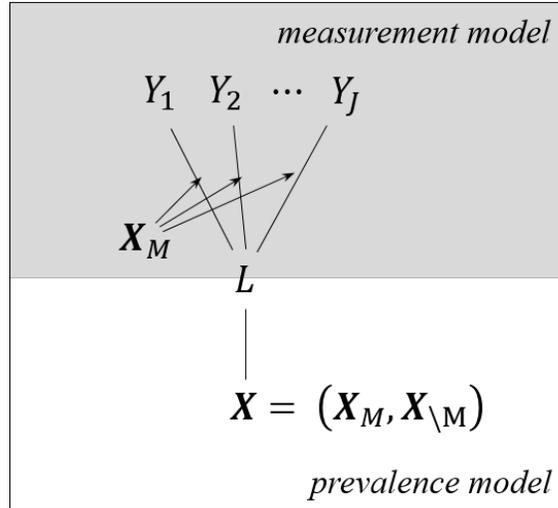
Latent-class (LC) analysis describes relationships among a set of categorical variables by positing associations with an unseen (latent) classifier. The method originated in the social sciences as a tool for synthesizing responses to multiple questions from attitudinal surveys (Lazarsfeld and Henry, 1968; McCutcheon, 1987). LC models have been used to predict disease status from multiple medical or psychiatric diagnostic tests (Formann and Kohlmann, 1996; Kendler et al., 1998), assess the reliability of raters (Agresti, 1992), and estimate rates and patterns of response error in surveys, censuses and administrative data without a gold standard (Biemer et al., 2001; Biemer and Wiesen, 2002; Prosser et al., 2008). For a binary classification, LC analysis typically needs three or more data sources for the model parameters to be estimable. With three items, Kreuter et al. (2008) concluded that:

...in the absence of a gold standard, the LCA model provides better estimates of the true proportion than any of the individual items would have, despite the presence of measurement error in all three indicators.

To predict citizenship status for persons enumerated in the census, we adapted a version of LC analysis known as LC regression. The model, which is depicted in Figure 9.1, has two parts. The upper part, which we call the *measurement model*, describes the associations between citizenship and the items assembled from the various data sources used to measure it. The items, denoted by Y_1, Y_2, \dots, Y_J , are assumed to be conditionally independent given the latent citizenship variable, which is denoted by L . An additional set of covariates, denoted by \mathbf{X}_M (the use of boldface indicates that it may be a vector), moderates the relationships between the items and the latent variable. Candidates for \mathbf{X}_M include person-level variables that influence the reliability of the item: scores from the PVS related to the probabilities of true/false match, the age of the record from which an item was obtained, and so on. The lower part of Figure 9.1, which we call the *prevalence model*, describes how rates of citizenship vary with respect to another set of covariates denoted by $\mathbf{X} =$

$(\mathbf{X}_M, \mathbf{X}_{\setminus M})$, which may include any of the variables from \mathbf{X}_M , plus an additional set $\mathbf{X}_{\setminus M}$ (the backslash is read as “not,” in the sense of a set difference operation). Candidates for $\mathbf{X}_{\setminus M}$ include person and housing-unit characteristics from the 2020 Census (age, sex, race/Hispanic origin, relationship to householder, tenure), indicators of geography, and other available covariates that do not directly measure citizenship status but may be useful for predicting it.

Figure 9.1: Latent-class regression model



9.2 Parameter Estimation

Under the LC regression model, the joint distribution of the items and latent-class variable given the covariates may be factored as

$$P(L, Y_1, \dots, Y_J | \mathbf{X}; \boldsymbol{\theta}) = P(L | \mathbf{X}; \boldsymbol{\beta}) \prod_{j=1}^J P(Y_j | L, \mathbf{X}_M; \boldsymbol{\rho}),$$

where $\boldsymbol{\rho}$ denotes the unknown parameters in the measurement model, $\boldsymbol{\beta}$ denotes the unknown parameters in the prevalence model, and $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\beta})$ is the vector of combined parameters. The measurement parameters $\boldsymbol{\rho}$ are typically expressed as item-response probabilities $\rho_{jk|c} = P(Y_j = k | L = c)$, where $k = 1, \dots, K_j$ indexes the response categories for item Y_j and $c = 1, \dots, C$ indexes the latent classes. The prevalence parameters $\boldsymbol{\beta}$ are commonly expressed as coefficients in a logistic regression for a binary outcome when $C = 2$, or a baseline-category logistic model for a multinomial response when $C > 2$ (Agresti, 2013).

A procedure for computing a joint maximum-likelihood (ML) estimate for $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ using an Expectation-Maximization (EM) algorithm was described by Dayton and Macready (1988) and Bandeen-Roche et al. (1997). More recently, Bakk and Kuha (2018) advocated a two-step procedure that:

- (i) estimates $\boldsymbol{\rho}$ under a reduced model that eliminates $\mathbf{X}_{\setminus M}$, and then
- (ii) estimates $\boldsymbol{\beta}$ with $\boldsymbol{\rho}$ fixed at its estimated value from step (i).

The results from this procedure are consistent if the model is correctly specified and nearly as efficient as joint ML estimates. The primary motivation for this two-step method is robustness to model misspecification; it protects the latent variable from being distorted if residual associations are present between the items Y_1, \dots, Y_J and the covariates $\mathbf{X}_{\setminus M}$ that are not fully explained by L . For our purposes, dividing the estimation into steps (i) and (ii) is also operationally convenient, as it allows us to implement and tune the measurement and prevalence models separately.

To perform the necessary computations, we have created a new R package called *cvam*, which embeds LC analysis within a more general framework of log-linear models with missing and coarsened data (Schafer, 2020).⁵⁴ Full details of the computational methods used in *cvam* are provided in the package documentation, vignettes, and technical appendices.

9.3 Measurement Model

Number of Classes

Because the ultimate purpose of our LC model is to weigh evidence for and against citizenship, we began by fitting models with $C = 2$ classes to represent citizens and noncitizens. However, it soon became apparent that two-class models failed to explain certain features of the administrative data. In particular, they struggled to give adequate weight to reports of citizenship from the U.S. Citizenship and Immigration Services (USCIS) and Passport files because they did not distinguish between citizens who are foreign-born and citizens who were born in the United States. Native-born persons should not be present in the USCIS file; consequently, being absent from this file provides some evidence that a person is native-born and therefore a citizen. The Passport file should include only citizens, but foreign-born citizens are much more likely to hold passports than native-born citizens are. Consequently, being absent from the passport file provides moderate evidence against citizenship among those who are foreign-born. For both the USCIS and Passport sources, country of birth is a lurking variable that is related to the chance of being present in the file, and this variable may introduce selection bias into the modeling if it is not explicitly taken into account.

Upon further investigation, we obtained better results from a three-class model whose categories correspond to:

- (i) noncitizens born outside of the United States,
- (ii) citizens born outside of the United States, and
- (iii) citizens born in the United States.

⁵⁴ Coarsened data, a term coined by Heitjan and Rubin (1991), refers to incomplete data that may convey intermediate amounts of information between fully observed and fully missing. Examples include values that are truncated, grouped or top-coded. Some of the administrative and survey data related to citizenship have this form. For example, the NUMIDENT records some individuals as having been born outside the United States, but does not indicate whether they are U.S. citizens. The *cvam* software allows this partial information contained in coarsened values to be used in model fitting.

As shown in Table 9.1, these groups belong to a 2×2 classification with one empty cell. The empty cell represents US-born noncitizens. As noted in Section 4, this is a tiny group that includes children born on American soil to foreign diplomat parents, and native-born citizens who relinquish their status to become citizens of another country and almost always live outside the U.S. Because they are so rare, our ability to detect them is likely to be swamped by errors from record linkage and other sources of misclassification, and so our model assigns them a probability of zero.

Table 9.1 Latent Classes Represented by the Three-Class Model

	Citizen?	
Born in the United States?	No	Yes
No	(i)	(ii)
Yes		(iii)

Measurement Items

To illustrate the behavior of this three-class model, we present results for a 2010-vintage dataset from the state of Delaware, with approximately 692,000 persons of voting age from the 2010 Census Edited File (CEF). This example has certain features that make LC analysis challenging. Delaware is a very small state with a lower-than-average proportion of noncitizens, which produces many empty and near-empty cells when individuals are cross-classified by key variables. Moreover, by working with 2010-vintage files, we reduce the coverage of some primary sources of citizenship and eliminate some secondary ones, making the model parameters more difficult to estimate. These features suggest that, if the model performs well for this test case, it is also likely to perform well in other contexts.

As described in Section 8, we have indirect evidence that individuals from the 2010 CEF, who were sent to search but could not be assigned a PIK, may have lower rates of citizenship than those who were assigned a PIK. This evidence comes from examining similar individuals in the ACS who were sent to search and not assigned a PIK, but who responded to the ACS questionnaire items on nativity and citizenship. To help capture this evidence, we augmented the 2010 CEF data with a supplementary file of 4,000 persons from the ACS who did not receive a PIK but were sent to search and included PIK status as an extra item in the measurement model. Data from these extra persons helped to identify the item-response parameters for this extra item, but after fitting the measurement model, the cases were removed from the analysis and did not contribute to estimation of the prevalence model, nor to forming predictions of citizenship probabilities for the CEF cases.

The items used in this 2010-vintage measurement model were:

- Nine-digit tax identification number in the ITIN range (ITIN status), with categories (i) Yes and (ii) No;⁵⁵
- NUMIDENT status, with categories (i) Noncitizen, (ii) Citizen, Foreign-Born and (iii) Citizen, US-Born;⁵⁶
- USCIS status, with categories (i) Noncitizen, (ii) Citizen, and (iii) Absent;⁵⁷
- Passport status, with categories (i) Foreign-Born, (ii) US-Born, and (iii) Absent;⁵⁸
- Census Bureau survey status, primarily the ACS, with categories (i) Noncitizen, (ii) Citizen, Foreign-Born, and (iii) Citizen, U.S. Born;⁵⁹ and
- PIK status, with categories (i) Not Sent to Search, (ii) No, and (iii) Yes.⁶⁰

Notice that, for the USCIS, “Absent” is not considered to be a missing value, but a distinct level indicating that the person was not located in the USCIS database, which provides mild evidence that the person was born in the United States and is therefore a citizen. Similarly, for Passport, “Absent” indicates that the CEF person does not have a U.S. Passport, which provides some evidence against citizenship.

For each item of citizenship information, individuals who could not be definitively assigned to one of the response categories contributed coarsened or missing values. For example, in the NUMIDENT, some individuals were recorded as foreign-born but their citizenship was unspecified. They represent a mixture of NUMIDENT categories (i) and (ii), with the mixing proportions unknown. For persons who did not receive a PIK, it was not possible to determine if they were present in the USCIS file. This group represents a mixture of USCIS categories (i), (ii) and (iii), with the mixing proportions unknown. The modeling procedures implemented in *cvam* accept codes corresponding to these mixtures and make use of all the complete and partial information available for each individual when computing parameter estimates.

Prior Information

LC models have unusual features that warrant caution during the fitting process. Maximum-likelihood (ML) estimates from an LC model are not unique. With C latent classes, there are $C! = 1 \times 2 \times \dots \times C$ equivalent solutions corresponding to all possible permutations of the class labels, and the EM algorithm will converge to different solutions depending on the starting values that are used. The loglikelihood function may have additional minor modes, and an unfortunate choice

⁵⁵ ITIN status = Yes means that the assigned PIK was assigned based on a nine-digit taxpayer ID that fell within the range of ITINs, and No means that it fell outside that range. This item was considered to be missing for persons who did not receive a PIK.

⁵⁶ For NUMIDENT status, any person reported to born in the United States was coded as (iii) citizen, US-born, and a very small number of reported US-born noncitizens were coded as missing values. This item was considered to be missing for persons who were not matched to the NUMIDENT.

⁵⁷ This item was considered to be missing for persons who did not receive a PIK or an EPIK.

⁵⁸ This item was considered to be missing for persons who did not receive a PIK or an EPIK.

⁵⁹ This item was considered to be missing for persons who did not receive a PIK, were not sampled for a Census Bureau survey, or were sampled and did not respond to questions on nativity and citizenship.

⁶⁰ In retrospect, it may have been better to define PIK status with two categories, “No” and “Yes,” and regard persons who were not sent to search as having missing values for this item. The latter would then have missing values for all six measurement items, and these persons would in effect be dropped from the fitting of the measurement model. This was done in follow-up analyses, and the results were essentially unchanged.

of starting values may cause EM to get stuck in one of these. Solutions on a boundary of the parameter space are also common, which (depending on how the M-step is implemented) may cause numerical instability.

The issue of class permutation, also known as label switching, can be addressed in various ways. For example, Richardson and Green (1997) imposed prior constraints on the parameter space to eliminate all but one of the equivalent solutions. Chung, Loken and Schafer (2004) made *a priori* assignments of a small number of sample units to latent classes, treating their class memberships as known, to break the symmetry of the loglikelihood function and guide the EM algorithm toward a solution with the desired labels. Prior assignments of in-sample cases can be viewed as a kind of data-dependent Bayesian prior distribution, and the resulting EM solution can be interpreted as a posterior mode. In a similar vein, one could augment the dataset with fictitious out-of-sample cases whose class memberships are known, which introduces a kind of prior distribution known as a data-augmentation prior (Bedrick *et al.*, 1996).

The *cvam* package has special features for adding data-augmentation priors to LC models. One feature is a flattening constant, a small positive value placed in every cell of the multidimensional contingency table that cross-classifies individuals by all model variables. Using a flattening constant is functionally equivalent to adding fictitious observations that are analogous to frequencies but are not necessarily integers, spread uniformly across all cells. In effect, a flattening constant adds a penalty function to the loglikelihood that penalizes the fit for parameter estimates near a boundary. The software also accepts “prior nuggets,” integer counts that are assigned to individual cells or spread across groups of cells. These nuggets can effectively guide the EM algorithm toward the mode that is consistent with the desired ordering of the latent-class labels. In our 2010-vintage model for Delaware, we applied:

- a flattening constant equivalent to 50 persons in total, spread equally across all cells of the table;
- a prior nugget equivalent to 200 persons observed within the latent class “Noncitizen,” with ITIN Status = “Yes,” NUMIDENT Status = “Noncitizen,” USCIS Status = “Noncitizen,” Passport Status = “Absent,” and Survey Status = “Noncitizen”;
- a prior nugget equivalent to 200 persons observed within the latent class “Citizen, Foreign-Born,” by ITIN Status = “No,” NUMIDENT Status = “Citizen, Foreign-Born,” USCIS Status = “Citizen, Foreign-Born,” Passport Status = “Citizen, Foreign-Born,” and Survey Status = “Citizen, Foreign-Born”; and
- a prior nugget equivalent to 200 persons observed within the latent-class “Citizen, US-Born,” with ITIN Status = “No,” NUMIDENT Status = “Citizen, US-Born,” USCIS Status = “Absent,” Passport Status = “Citizen, US-Born,” and Survey Status = “Citizen, US-Born.”

This prior information, which was chosen after some trial and error, was strong enough to reliably guide EM toward the correctly labeled solution, but weak enough to have a negligible effect on the parameter estimates of interest, which we now present.

Item-Response Probabilities

The estimated item-response probabilities for the 2010-vintage model for Delaware are shown in Table 9.2.

Table 9.2 Estimated Item-Response Probabilities for 2020-Vintage Latent-Class Model for Delaware

Item's Response Status	Noncitizen	Citizen, Foreign-Born	Citizen, US-Born
ITIN Status			
Yes	0.1959	0.0004	0.0000
No	0.8041	0.9996	1.0000
NUMIDENT Status			
Noncitizen	0.9798	0.3626	0.0000
Citizen, Foreign-Born	0.0165	0.6299	0.0003
Citizen, US-Born	0.0038	0.0075	0.9996
USCIS Status			
Noncitizen	0.6246	0.1445	0.0001
Citizen, Foreign-Born	0.0174	0.6210	0.0001
Absent	0.3579	0.2346	0.9998
Passport Status			
Passport, Foreign-Born	0.0005	0.8675	0.0002
Passport, US-Born	0.0005	0.0004	0.4406
Absent	0.9990	0.1321	0.5592
Census Bureau Survey Status			
Noncitizen	0.8886	0.1232	0.0015
Citizen, Foreign-Born	0.0614	0.8158	0.0019
Citizen, US-Born	0.0500	0.0609	0.9966
PIK Status			
Not Sent to Search	0.0431	0.0324	0.0470
No	0.2282	0.3495	0.0132
Yes	0.6687	0.6182	0.9397

Note: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

The story told by these estimates is highly plausible.

- **ITIN Status:** The model estimates that about 20% of noncitizens have ITINs but essentially no citizens have them.
- **NUMIDENT Status:** Noncitizens are very likely to be reported as noncitizens in the NUMIDENT, and US-born citizens are very likely to be reported as US-born citizens. For foreign-born citizens, however, the model estimates that more than one-third (36%) are misclassified as noncitizens, presumably because these naturalizations were not reported to the SSA. These item-response probabilities provide a strong justification for using a three-class model rather than a two-class model. The misclassification rates for foreign-born and US-born citizens are very different, and collapsing them into a single category would muddle this important effect.
- **USCIS Status:** Summing the first two rows, about 64% of noncitizens and 76% of foreign-born citizens show up in this data source, but virtually none of the US-born citizens do. A

small minority of noncitizens are misclassified as citizens, but a non-negligible portion of citizens are misclassified as noncitizens. The three-class model correctly limits the coverage of this source to foreign-born persons.

- **Passport Status:** The estimated probability of a noncitizen holding a U.S. Passport is essentially zero, which agrees with our understanding that only U.S. citizens may be granted U.S. passports. The model estimates that about 87% of foreign-born citizens have received passports, compared to only 44% of U.S.-born citizens. Again, these estimates show the advantage of a three-class model; the rates of holding a passport are vastly different for the two groups, and combining them would mask this effect, distorting the evidence from this key data source.
- **Census Bureau Survey Status:** Nearly all US-born citizens are correctly classified in Census Bureau surveys. Approximately 11% of noncitizens are misclassified as citizens, and about 12% of foreign-born citizens are misclassified as noncitizens. As with the NUMIDENT, the misclassification rates among foreign-born citizens and US-born citizens are very different, a distinction that would be hidden by a two-class model.
- **PIK Status:** This model estimates that about 94% of US-born Citizens successfully received a PIK compared to about 62% of foreign-born citizens and 67% of noncitizens. The estimated probabilities of not being sent to search are fairly constant (between 3% and 5%) across the three groups, but the estimated probabilities of being sent to search but not receiving a PIK are vastly different (between 1% and 35%). These results strongly support the notion that being sent to search but not receiving a PIK provides some evidence of noncitizen status that should not be ignored.

Limitations of the Measurement Model

This version of a three-class measurement model is limited in one crucial respect: It does not yet include any moderator variables, covariates \mathbf{X}_M that may affect the relationships between L and Y_1, \dots, Y_J . That is, it assumes that the item-response probabilities shown in Table 9.2 are constant across all subpopulations in the state of Delaware. We have strong reasons to believe this is not the case. For example:

- The accuracy of a NUMIDENT report of “noncitizen” may vary by the year in which the individual’s SSA records were last updated. The probability of a foreign-born citizen being correctly recorded as such could be higher for those whose NUMIDENT reports are more recent.
- The accuracy of a USCIS report of “noncitizen” may vary by age. The probability of foreign-born citizens being correctly recorded as such could be lower for young persons, because that group includes some who have derived citizenship but did not apply for naturalization certificates.
- The accuracy of a Census Bureau survey report of “noncitizen” may vary by the year of the survey. The probability of a foreign-born citizen being correctly reported as such could be higher for those whose survey responses are more recent.
- The accuracy of any data source may vary by source-specific measures of PVS match quality, which are related to false-match rates.

- The accuracy of any data source may vary across subpopulations of race and ethnicity.

In future rounds of model-fitting, we will introduce moderator variables to see how the results change. The item-response probabilities shown in Table 9.2 will then be replaced by regression coefficients from multinomial baseline-category logistic models for each item, within each latent class.

9.4 Bayes Factors

Once the measurement model has been fit, evidence about the latent variable contained in the items can be conveniently summarized with Bayes factors. For any particular pattern of responses to the items Y_1, \dots, Y_j , the Bayes factor for membership in class $L = c$ relative to $L = c'$ is defined as the conditional likelihood of observing that pattern assuming that $L = c$, divided by the conditional likelihood of observing that pattern assuming that $L = c'$. For our purposes, it is helpful to compute a Bayes factor for citizen versus noncitizen status, with the numerator representing a combination of two classes (foreign-born citizens and US-born citizens) mixed in proportions estimated under the model. To interpret the Bayes factor, we compute its base-10 logarithm and compare it to the values shown in Table 9.3, which was adapted from Kass and Raftery (1995).

Table 9.3 Interpretation of Bayes Factor (BF) and its Base-10 Logarithm

<i>BF</i>	$\log_{10} BF$	Interpretation
Greater than 100	Greater than 2.0	Extreme evidence for citizenship
Between 30 and 100	Between 1.5 and 2.0	Very strong evidence for citizenship
Between 10 and 30	Between 1.0 and 1.5	Strong evidence for citizenship
Between 3 and 10	Between 0.5 and 1.0	Moderate evidence for citizenship
Between 1 and 3	Between 0 and 0.5	Weak evidence for citizenship
Equal to 1	Equal to 0	No evidence either way
Between 1/3 and 1	Between -0.5 and 0	Weak evidence for noncitizen status
Between 1/10 and 1/3	Between -1.0 and -0.5	Moderate evidence for noncitizen status
Between 1/30 and 1/10	Between -1.5 and -1.0	Strong evidence for noncitizen status
Between 1/100 and 1/20	Between -2.0 and -1.5	Very strong evidence for noncitizen status
Less than 1/100	Less than -2.0	Extreme evidence for noncitizen status

In results not shown here, we have studied the log-10 Bayes factors for the most common response patterns seen in the Delaware test data. Passports provide the strongest support for citizenship, while ITINs are the clearest indicator of noncitizen status. In a few cases the log-10 Bayes factor is negative for combinations that business rules may classify as citizens, such as persons listed as noncitizens in the NUMIDENT, citizens in the USCIS file, and without a passport. It is reasonable to think that many of these are naturalized citizens whose naturalizations were not reported to the SSA. Because a very high percentage of foreign-born citizens are estimated to hold passports, when a foreign-born person does not have one, that fact provides moderate evidence against citizenship. Nevertheless, many of the persons in this group would still be classified as citizens under this model, because overall rates of citizenship in the population are still high. Bayes' Theorem implies that the posterior probability of citizenship for an individual depends not only on

the Bayes factor, but on the prevalence of citizenship in the reference group to which that individual belongs, as determined by prevalence-model covariates. The impact of the Bayes factor in groups of varying prevalence is shown in Table 9.4. For example, suppose that a person with a log-10 Bayes factor of -0.83 belongs to a group consisting of 80% citizens and 20% noncitizens. Such a person has a posterior probability of citizenship somewhere between 0.286 and 0.558 (about 0.47 with a linear interpolation).

Table 9.4 Posterior Probabilities of Citizenship by Bayes Factor and Prevalence

$\log_{10} BF$	Prevalence of Citizenship							
	0.50	0.60	0.70	0.80	0.90	0.95	0.98	0.99
-3.00	0.001	0.001	0.002	0.004	0.009	0.019	0.047	0.090
-2.50	0.003	0.005	0.007	0.012	0.028	0.057	0.134	0.238
-2.00	0.010	0.015	0.023	0.038	0.083	0.160	0.329	0.497
-1.50	0.031	0.045	0.069	0.112	0.222	0.375	0.608	0.758
-1.00	0.091	0.130	0.189	0.286	0.474	0.655	0.831	0.908
-0.50	0.240	0.322	0.425	0.558	0.740	0.857	0.939	0.969
0.00	0.500	0.600	0.700	0.800	0.900	0.950	0.980	0.990
0.50	0.760	0.826	0.881	0.927	0.966	0.984	0.994	0.997
1.00	0.909	0.938	0.959	0.976	0.989	0.995	0.998	0.999
1.50	0.969	0.979	0.987	0.992	0.996	0.998	0.999	1.000
2.00	0.990	0.993	0.996	0.998	0.999	0.999	1.000	1.000
2.50	0.997	0.998	0.999	0.999	1.000	1.000	1.000	1.000
3.00	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000

Prevalence Model

The prevalence model describes how the distribution of the latent classes varies over the population in relation to characteristics of persons, housing units and areas in which they are tabulated in the Decennial Census. Our proposed model is a baseline-category logistic regression (Agresti, 2013). Choosing one of the C latent classes, c' , as the baseline class, the model is

$$\log \frac{P(L = c | \mathbf{X}; \boldsymbol{\beta})}{P(L = c' | \mathbf{X}; \boldsymbol{\beta})} = \mathbf{X}^T \boldsymbol{\beta}_c$$

for $c = 1, \dots, C$, where $\boldsymbol{\beta}_c$ is a vector of regression coefficients specific to class c , and the superscript ‘ T ’ denotes a transpose. The choice of baseline class affects the meaning and values of the coefficients, but it does not change the fitted values or predictions. The coefficients for the baseline class are not identified, and it is customary to set the elements of $\boldsymbol{\beta}_{c'}$ to zero. At any fixed values for the parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_C)$, the class probabilities may be computed as

$$P(L = c | \mathbf{X}; \boldsymbol{\beta}) = \frac{\exp(\mathbf{X}^T \boldsymbol{\beta}_c)}{\sum_{d=1}^C \exp(\mathbf{X}^T \boldsymbol{\beta}_d)}.$$

Although the prevalence model does not appear to involve the citizenship items Y_1, \dots, Y_j or the measurement parameters $\boldsymbol{\rho}$, those quantities interact with the prevalence model in two important

ways. First, they enable us to update the prevalence-model predictions to incorporate information from the citizenship items. Let $\mathbf{Y}^* = (Y_1^*, \dots, Y_j^*)$ denote a set of citizenship items observed for an individual, where Y_j^* is an expanded version of Y_j that may be a coarsened value or a missing-value code. By Bayes' Theorem, the posterior probability that an individual belongs to class c , given the covariates \mathbf{X} and citizenship items \mathbf{Y}^* , is

$$P(L = c | \mathbf{X}, \mathbf{Y}^*; \boldsymbol{\beta}, \boldsymbol{\rho}) = \frac{P(L = c | \mathbf{X}; \boldsymbol{\beta}) \mathcal{L}(\mathbf{Y}^* | L = c, \mathbf{X}; \boldsymbol{\rho})}{\sum_{d=1}^C P(L = d | \mathbf{X}; \boldsymbol{\beta}) \mathcal{L}(\mathbf{Y}^* | L = d, \mathbf{X}; \boldsymbol{\rho})},$$

where $\mathcal{L}(\mathbf{Y}^* | L = c, \mathbf{X}; \boldsymbol{\rho})$ denotes the conditional likelihood of observing \mathbf{Y}^* for an individual with covariates \mathbf{X} who belongs to class c .⁶¹ These likelihood values depend on the measurement parameters $\boldsymbol{\rho}$, and the R package that we have developed includes a function that quickly computes the likelihoods after the measurement model has been fit. If all citizenship items are missing for an individual, i.e., if \mathbf{Y}^* contains only missing-value codes, then the likelihoods $\mathcal{L}(\mathbf{Y}^* | L = c, \mathbf{X}; \boldsymbol{\rho})$ become equal for $c = 1, \dots, C$, and $P(L = c | \mathbf{X}, \mathbf{Y}^*; \boldsymbol{\beta}, \boldsymbol{\rho})$ becomes identical to $P(L = c | \mathbf{X}; \boldsymbol{\beta})$.

The second way that citizenship items and measurement parameters interact with the prevalence model is that they play an essential role in estimating $\boldsymbol{\beta}$. The prevalence model cannot be fit using ordinary logistic-regression software, because the model's dependent variable is latent. To fit the prevalence model, we have implemented an EM algorithm that is formally equivalent to step (ii) of the two-step procedure of Baak and Kuha (2018) described in Section 9.2. In this iterative procedure, we initialize the elements of $\boldsymbol{\beta}$ to zero, and then update them by an Expectation or E-step followed by a Maximization or M-step. The E-step involves computing the fitted values $P(L = c | \mathbf{X}; \boldsymbol{\beta})$, $c = 1, \dots, C$ for each individual, then using the likelihoods from the measurement model to convert them to $P(L = c | \mathbf{X}, \mathbf{Y}^*; \boldsymbol{\beta}, \boldsymbol{\rho})$. The M-step resembles the fitting of a conventional baseline-category logistic model, except that the dependent variables, the unseen indicators for class membership, are replaced by fractional observations obtained from the E-step. Repeating the E- and M-steps, the coefficients $\boldsymbol{\beta}$ eventually converge to the ML solution with $\boldsymbol{\rho}$ fixed at their estimated values from the measurement model.

To carry out this procedure in the 2020 CVAP project, we need to formulate a prevalence model that preserves the associations between citizenship and key characteristics of persons (notably, the twelve-category classification by race and Hispanic origin) and, at the same time, is sufficiently responsive to local variation to yield plausible and efficient predictions for small geographies. In the final tabulations, however, the choice of prevalence model is unlikely to have a large influence on the results. For a vast majority of individuals, the available citizenship items carry enough information to push the posterior probabilities very close to zero or one regardless of what the prevalence model predicts. Changes to the prevalence model will noticeably influence the results

⁶¹ A likelihood is a non-negative number that is proportional to a probability, and the constant of proportionality is arbitrary. Without loss of information, the set of conditional likelihoods for classes $c = 1, \dots, C$ may be re-expressed as a set of $C - 1$ Bayes factors for each class relative to a baseline class, because the proportionality constants cancel out when a ratio is taken.

only for the roughly 9% of the persons for whom no citizenship measures are available, plus a small minority for whom the evidence is contradictory or inconclusive.

Thus far, we have experimented with three types of prevalence models using the Delaware test dataset:

- (i) Statewide models with fixed effects for race/ethnicity and other person-level characteristics from the Decennial Census, plus fixed main effects for counties and tracts. A statewide fixed-effects model can make use of a rich set of demographic covariates, but these cannot be interacted with geography without introducing zero cells that destabilize the model fit. Fitting these models to Delaware takes several hours or more, and they may be impractical or impossible for larger states.
- (ii) Statewide models with fixed effects for person-level characteristics, plus spatially correlated random effects for blocks. Spatial models are conceptually attractive and strengthen estimates for small areas by borrowing information from similar areas nearby. They would require more specialized computational routines such as Markov chain Monte Carlo (MCMC) in place of the EM algorithm described above, and fitting them may take several days. Preliminary results suggest that the extra effort and time needed for these models is not worthwhile, because the impact of the variation across blocks is small relative to the influence of the important demographic characteristics, namely race/ethnicity.
- (iii) County-level models with fixed effects for a few person-level characteristics. Fitting a separate model to each county is equivalent to fitting a statewide model that interacts person-level variables with county indicators. These models can be fit very quickly and are responsive to local variation, at least at the county level. However, some adjustments are needed to stabilize these models to handle the sparseness and empty cells that arise as more person-level characteristics are introduced. A Bayesian method used by Clogg et al. (1991), which adds a small amount of prior information to each distinct covariate pattern observed in a tract, guarantees convergence in most cases while retaining the simplicity of ML computations.

For the 2010-vintage test, we adopted strategy (iii) and fit a separate three-class prevalence model to each county, with main effects for the twelve categories of race/ethnicity plus a binary dummy indicator for whether a foreign language was spoken at home, using the Bayesian technique of Clogg et al. (1991). The estimation procedure converged for nearly every county in the United States. For a small handful of sparsely populated counties, the procedure did not converge, and for these cases, we computed posterior probabilities by substituting the coefficients obtained by fitting the same prevalence model to the entire state. The results from this procedure are summarized in Section 10.

The fitting procedures just described do not permit missing values in \mathbf{X} . For this exercise, the person-level characteristics appearing in \mathbf{X} came from the 2010 Census Edited File (CEF), in which the missing values had already been filled in using hot-deck imputation. These imputed values were those used in the hot-deck and logistic approaches described in Sections 6-8.

10. Empirical Results Comparing the Approaches

This section shows citizenship estimates produced by the four approaches using the 2010 Census Edited File (CEF) and 2018 American Community Survey (ACS) population frames, each of which has advantages.⁶² The 2010 CEF is the closest population frame to the 2020 CEF. It is of similar size, which allows us to test whether and how quickly each approach can produce citizenship estimates for a decennial population. The 2010 and 2020 Censuses share common data collection methods (e.g., fieldwork for all non-responding housing units rather than just a sample like in the ACS, and the use of proxy responses and whole household imputations),⁶³ so the types of cases for which estimates are generated are more similar than with the ACS frame.

Results based on the 2018 ACS have important advantages as well. As discussed in Section 4, all the citizenship sources contain 2018-vintage data, but many do not include 2010-vintage data. The 2018 ACS contains citizenship responses, unlike the 2010 CEF. Having the ACS citizenship responses in the same population frame as the modeling approaches' estimates greatly facilitates comparisons, as discussed further below.

We compare the results to ACS estimates. The ACS has several advantages as a comparator. It is the largest representative survey with citizenship information. It contains citizenship not only for persons with linkable administrative and prior-survey citizenship information, but also for those who failed to receive a PIK for different reasons. The non-business rule (NBR) records are particularly important for our evaluation, since it is more challenging to estimate citizenship when external information can't be linked. Like any survey, though, the ACS also has sampling error, nonresponse bias, and response error. As shown in Section 4, response error for as-reported citizenship appears to be relatively low for BR cases, and the two types of error nearly cancel each other out. We cannot observe the nature of response error for NBR cases, however, and it may differ from that for BR cases.

10.1 2010 Census Edited File (CEF) Test Results

All four methods produced citizenship estimates for all 2010 CEF person records within 24 hours. The execution time could most likely be shortened using a dedicated computing cluster and parallel processing. A 24-hour processing time, though, is adequate to meet production deadlines. The test results thus confirm that each method is feasible.

Table 10.1 provides the citizen share estimates for each of the approaches, as well as estimates from the 2010-2012 ACS as a comparator.⁶⁴ All approaches produce lower, but quite similar,

⁶² Some of the approaches produce citizenship probabilities between 0 and 1 for part or all of the population. We do not impute a person's citizenship by converting their citizenship probability into a 0 or 1. Rather, the citizenship probabilities are summed across individuals in each reported cell.

⁶³ There remain some differences in data collection between the two, though. Unlike the 2010 Census, the 2020 Census uses internet self-response as well as administrative record enumeration for some non-responding housing units.

⁶⁴ We have chosen to show 2010-2012 ACS estimates rather than 2008-2012, because the 2008-2009 ACS PVS information is not compatible with that in the 2010 CEF, as discussed in Section 7. This confounds comparison of BR/PVS group results between the 2008-2009 ACS and estimates applied to the 2010 CEF.

citizen shares to the ACS overall, as well as by race/ethnicity (Table 9.1.A). The Hot Deck produces the closest overall estimates, and the Latent Class Model estimates are furthest away.

The BR sample estimates shown in Table 10.1.B are extremely similar across approaches.⁶⁵ This is not surprising for the three methods using nearly identical business rules. The fact that the Latent Class estimates are also very close provides important confirmation that the business rules function well. As discussed in Section 9, the Latent Class approach imposes fewer assumptions on the data than business rules and doesn't force values to be only zero or one.

Across the board, the model estimates for BR cases show fewer citizens than ACS self-response would suggest. These differences may be due to the differing frames between the ACS and the CEF. To measure this, the rightmost column in Table 10.1.B applies a set of business rules to the 2010-2012 ACS frame. As can be seen, applying business rules to the same ACS records generally produces slightly higher citizen shares than the ACS citizenship self-responses do. This suggests that the ACS population frame has a higher citizen share than the CEF population frame, at least among persons with linkable citizenship information. It also means that the true citizen share in the CEF is probably lower than the 2010-2012 ACS estimates shown in Table 10.1.A. This CEF-ACS discrepancy is particularly pronounced for Hispanics, whose BR citizen share is more than four percentage points higher in the ACS (~71.6%) than the CEF (~67-68% depending on the method).

The NBR-PIK sample estimates vary widely (Table 10.1.C). Part of this variation is attributable to the business rules used by the BR Logistic method, which classify all foreign-born NUMIDENT persons with missing citizenship as noncitizens and therefore estimates a lower citizenship rate than the others. The ACS Logistic produces estimates closest to the ACS by a considerable margin.

The Hot Deck and BR Logistic estimates are higher than the ACS for the NBR-SS group, while the ACS Logistic and Latent Class estimates are lower (Table 10.1.D). The ACS Logistic citizen shares are particularly low compared to the others for non-Hispanic Asians and Hispanics, and they are also somewhat lower than the ACS as well. The higher citizen shares for the Hot Deck and BR Logistic approaches may be a result of using BR cases rather than NBR-SS cases for their imputations. This is consistent with the finding in Section 8 that BR cases have systematically higher ACS citizen shares for these race/ethnic groups compared to NBR-SS cases. The Latent Class estimates for non-Hispanic Asians, Hispanics, and non-Hispanic Asian and Whites are above the ACS estimates, while the others are below the ACS estimates. This is likely due to the fact that the Latent Class approach does not distinguish the race of the NBR-SS persons in the measurement model. Future versions of the Latent Class approach will attempt to address this.

Table 10.1.E shows results for detailed Asian and Hispanic groups in the NBR-SS subgroup. A high percentage of people of Puerto Rican heritage are citizens, and the ACS estimate reflects that. The ACS Logistic and Hot Deck estimates for Puerto Ricans are very close to the ACS, while the Latent Class and BR Logistic estimates are much lower. In general, the BR Logistic estimates vary less within Asian and Hispanic groups than those for the other approaches; note that the BR

⁶⁵ The BR sample is defined as observations having nonmissing values in the "Primary, then Secondary" rules described in Section 4.

Logistic model doesn't distinguish these detailed groups, while the other approaches do. The ACS Logistic approach produces the greatest variation in estimates across detailed groups, closely tracking the ACS in most categories.

One reason the ACS estimates for the NBR-SS group are higher than the ACS Logistic estimates is that the ACS Logistic approach does not include edited or imputed ACS citizenship responses in the model fitting, while the ACS estimates include those observations. Tables 10.2.A and 10.2.B show that as-reported ACS citizen shares in the NBR-SS group are systematically lower than edited and imputed shares for many race groups, due to the edit and imputation process not taking the PVS category into account. The as-reported versus imputation difference is especially strong for persons of Chinese, Korean, Mexican, Central American, and Latin American origin as shown in Table 10.2.B. In contrast, the citizen share for Puerto Rican imputations is nearly 13 percentage points lower than the as-reported share.⁶⁶

Turning back to model estimates, all the approaches produce lower citizen shares than the ACS for the NBR-NSS group (Table 10.1.F). This probably reflects differences in the composition of the NBR-NSS group in the CEF and ACS, as discussed in Section 8, so the ACS may not be a good comparator for this group. The Hot Deck produces lower estimates for nearly all race groups. This may be because the Hot Deck conditions on whether anyone in the household lacks a PIK. Since both NBR-NSS and NBR-SS persons lack PIKs, this conditioning will be the same for each. It doesn't capture the fact that the ACS citizen share for the NBR-SS group is 21 percentage points lower than for the NBR-NSS group.

To better understand the differences between the ACS Logistic versus Hot Deck and BR Logistic approaches, we decompose them into four parts in Table 10.3. The first column shows the ACS Logistic estimates, trained on ACS NBR-SS cases. The second is estimates trained on ACS BR cases, while still using ACS citizenship as the dependent variable. The third again uses BR cases in the model, but this time also using BR citizenship as the dependent variable rather than ACS citizenship. CEF BR cases are the training sample in the fourth column, while keeping the same model specification as used in the previous columns. The differences between the fourth column and the Hot Deck and BR Logistic estimates are due to model specification differences across approaches. The results show that whether one uses NBR-SS or BR cases for the training sample is the most important reason for the differences in the estimates across approaches. Note that it matters a lot for non-Hispanic Asians, Hispanics, and Asian and Whites, but very little for most other race groups. It is especially pronounced for people of Korean, Mexican, Cuban, Central American, and Latin American origin. Switching the dependent variable from ACS to BR citizenship raises the estimates somewhat for non-Hispanic Asians and Hispanics as well, but has little effect for most other races. The switch from the ACS to CEF population frame matters little. Finally, the ACS Logistic model specification produces higher estimates than the Hot Deck conditioning and BR Logistic specification, but this effect is significantly smaller than the effect from using NBR-SS versus BR cases for the training sample. The lower estimate from the Hot Deck conditioning compared to the ACS Logistic specification is likely due to conditioning on at least one person in the household not having a PIK. The ACS Logistic captures that effect by

⁶⁶ The ACS citizenship edit and imputation process doesn't condition on detailed Asian or Hispanic origin.

training on NBR-SS cases rather than from a model variable. The lower estimates from the BR Logistic specification could be due to the inclusion of tract fixed effects, which may capture small area clustering of persons by citizenship status better than the ACS Logistic variable for county citizen share. Unfortunately, the sample sizes in the ACS Logistic specifications are too small to support inclusion of tract fixed effects.

One could argue that the ACS Logistic approach has an advantage in tracking ACS citizenship estimates in the 2010 CEF test, since it uses the ACS frame and citizenship responses, while the other approaches use primarily or only 2010 CEF data. Switching from ACS to BR citizenship and from the ACS to the CEF population frame are not nearly as important as the type of training sample (NBR-SS versus BR persons) in explaining the differences between the ACS Logistic, Hot Deck, and BR Logistic estimates, however. This suggests that the potential advantages of using the ACS frame or ACS citizenship responses are not driving factors of the differences.

10.2 2018 American Community Survey (ACS) Test Results

We produce Hot Deck, BR Logistic, and ACS Logistic estimates using the 2018 ACS frame. Latent Class results have not been produced. That method is undergoing further development to add moderating variables to the measurement model. This will allow the effect of being in a particular BR/PVS group on one's citizenship status to vary across race groups. Other moderators include citizenship source vintage and record linkage quality.

Unlike in the 2010 CEF test, we impose two important sample restrictions on the 2018 ACS data to facilitate comparison between ACS citizenship and the remaining three approaches' estimates. Only records with as-reported ACS citizenship are included. And records not processed by PVS are excluded, because they cannot be assigned BR/PVS groups.⁶⁷

Unlike the 2010 CEF test, this test could give a potential advantage to the Hot Deck and BR Logistic approaches, since they use 2018 ACS data to produce the estimates, while the ACS Logistic approach uses earlier ACS data. The ACS Logistic approach could experience errors due to differences in associations between explanatory variables and citizenship over time.

In contrast to the 2010 CEF test, all the approaches produce slightly higher overall citizen share estimates than the ACS (Table 10.4.A). The ACS Logistic estimates are closest to the ACS, and the BR Logistic estimates are furthest away. The main reason for the close proximity of the overall estimates is that the BR citizen share is within 0.01 percentage points of the ACS, and the BR cases make up 93.4 percent⁶⁸ of the sample. The ACS Logistic estimates are closest to the ACS for the NBR-PIK and NBR-SS groups, while the BR Logistic is closest for the NBR-NSS group.

All the approaches come very close to the ACS for non-Hispanic White and non-Hispanic Black (Table 10.4.B). The ACS Logistic is closest for non-Hispanic Asians and Hispanics.

⁶⁷ These records were excluded from the PVS process entirely due to being in sensitive group quarters. In contrast, all 2010 CUF records were included in the PVS process. This is another confounding factor when making comparisons between ACS and the 2010 CEF estimates.

⁶⁸ The Census Bureau Disclosure Review Board approval number is CBDRB-FY20-CED006-0031.

The estimated citizen shares for the BR cases are quite similar to the ACS across race groups for each approach (Table 10.4.C). Noticeable differences emerge in the NBR-SS sample, though (Table 10.4.D), particularly for Hispanics. The BR Logistic Hispanic estimate is over 25 percentage points higher than the ACS, while the ACS Logistic is just three percentage points higher. The differences for Central Americans are most stark: the BR Logistic estimate is nearly 40 percentage points higher than the ACS, but the ACS Logistic is less than five percentage points higher (Table 10.4.E). As is the case in the 2010 CEF test, the BR Logistic Puerto Rican estimate is well below the others, due to not conditioning on detailed Hispanic groups in the models.

We are interested not only in how the approaches compare to the ACS at a national level, but also at lower levels of geography. Figures 9.1.A-E provide state-level estimates. When all races and all PIK status groups are combined, the estimates in states with higher citizen shares are virtually the same as the ACS, while they diverge somewhat in states with lower citizen shares like California and Texas. The ACS Logistic estimates are closer to the ACS in the states with lower citizen shares. No approach is closest most of the time for Non-Hispanic Asians (Figure 10.1.B). The ACS Logistic tracks the ACS best for Hispanics, and the differences across approaches are quite large in states with lower citizen shares (Figure 10.1.C). The estimates are all within one percentage point of the ACS for non-Hispanic Whites (Figure 10.1.D), and that's true for non-Hispanic Black estimates in most states as well (Figure 10.1.E).

Drilling down further, we calculate the record-by-record disagreement rates between the approaches and ACS citizenship in Table 10.5. These disagreement rates are based on a confusion matrix, which is generally used to measure agreements and disagreements between two classification methods. To compare among the models, we have modified the traditional form to accommodate probabilistic predicted values. The disagreement rate provides the expected proportion of cases where the models provide different predicted values. Here, in the case of comparing model results to binary ACS citizenship survey response, the overall disagreement rate simplifies to the following process: first, calculate the absolute value of the difference between a model's predicted probability of citizenship and the ACS survey response (coded as 0 or 1); second, multiply each individual's absolute difference by their ACS person weight; third, sum up all of these weighted differences; fourth, divide the sum by the total weighted population to get a rate of disagreement. As can be seen, the ACS Logistic has the smallest disagreement with the ACS overall with approximately 2.5% of cases in expectation.

We also display the disagreement rates by PIK status in Table 10.5. The disagreement rate for BR cases is quite low—only about 1.3-1.4% of BR individuals disagree with their ACS survey responses in expectation. The disagreement rate is much higher for the NBR-SS group—where 18.9% of ACS Logistic NBR-SS cases disagree with their ACS survey responses in expectation, and 26.5% of BR Logistic NBR-SS cases do. The ACS Logistic has the smallest disagreement with the ACS overall and in the NBR-SS and NBR-NSS groups. The Hot Deck's disagreement rate is lowest for the NBR-PIK group.

10.3 Discussion

The analysis shows that citizenship for the vast majority of the voting-age population can be reliably estimated using citizenship information linked from external sources. Both categorizing citizenship via business rules and a Latent Class model that imposes few assumptions produce estimates remarkably consistent with each other and with as-reported ACS citizenship.

The ACS shows that citizenship rates are strongly associated with the reason for not having linked citizenship information (has a PIK versus no PIK and sent to search versus no PIK and not sent to search). Since the ACS contains as-reported citizenship for people in all these groups, it can serve as a valuable training dataset for citizenship models for these groups. For the NBR-SS group, by far the largest non-BR group in the ACS, the ACS Logistic approach produces quite similar estimates to the ACS by training the model on NBR-SS cases with as-reported ACS citizenship from past ACS surveys. These estimates are much closer to the ACS than ones generated from BR cases.⁶⁹

NBR-NSS group citizenship estimation is the most challenging. Imputation rates for many of the variables that can help predict citizenship are high for these cases, reducing the accuracy of the models. The ACS contains very few NBR-NSS people, and the group may not be representative of the much larger census NBR-NSS group. ACS citizenship for the ACS NBR-NSS group may thus not be a good comparator for the census NBR-NSS group. As a result, we don't know which of the approaches produces estimates closest to what one would get from as-reported ACS responses, and it isn't fruitful to use the ACS NBR-NSS group as a guide for fine-tuning the models. The ACS NBR-NSS group may not be a suitable training sample. Not only may it not be representative of the 2020 Census NBR-NSS group, but it contributes few observations for a training sample, which means that only very simple model specifications can be estimated. Using BR alone to train the model may also not be appropriate, because a comparison of ACS and census shares of people in each of the BR/PVS groups suggests that the census NBR-NSS group draws significantly from both the ACS BR and NBR-SS groups. The make-up of the 2020 Census NBR-NSS group could also be different from any past survey or census due to its occurring during the COVID-19 pandemic. People who may have responded in the past may not be willing to do so, resulting in an increase in proxy responses or whole household imputations. It may thus be optimal to use a model based on all ACS cases with as-reported citizenship that controls for the probability the person is like someone in the 2020 Census NBR-NSS group. The training sample would be large enough to permit use of a rich model specification, and it can be tailored to the characteristics of the actual 2020 Census NBR-NSS group. We will continue developing the modeling for this group in particular.

⁶⁹ One could argue that the ACS Logistic approach has an advantage in tracking ACS citizenship estimates in the 2010 CEF test, since it uses ACS data as the training sample, while the other approaches use primarily or only 2010 CEF data. Also note that the results in Table 10.3 suggest that switching from ACS to BR citizenship and from the ACS to the CEF population frame are not nearly as important as the type of training sample (NBR-SS vs. BR persons) in explaining the differences. In contrast, the Hot Deck and BR Logistic approaches may have an advantage in the 2018 ACS test, since they use 2018 ACS data to produce their estimates. The ACS Logistic approach uses earlier ACS data, which could cause errors due to the different vintage. Since the differences in the results across approaches for the NBR-SS group are similar across the two tests, these potential advantages appear not to be important.

Table 10.1.A Percent Citizens, 2010 CEF Full Sample

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS	Share of CEF Population
Total	91.40	91.37	91.14	90.80	91.51	100.00
NH Asian Alone	67.68	67.53	67.23	68.30	67.38	4.81
Hispanic	63.95	64.76	62.52	63.29	65.69	14.22
NH White Alone	98.27	98.09	98.20	97.75	98.25	66.98
NH Black or African American	95.04	94.88	95.01	93.65	95.29	11.65
NH AIAN Alone	99.08	98.39	99.18	97.36	99.41	0.68
NH NHOPI Alone	80.59	80.17	81.25	78.70	82.28	0.15
NH Some Other Race Alone	69.24	69.92	71.13	67.58	72.69	0.16
NH AIAN and White	99.72	99.48	99.73	99.41	99.82	0.37
NH Asian and White	90.45	90.06	89.85	90.16	90.32	0.32
NH Black or African American and White	97.01	96.55	96.91	95.88	97.04	0.21
NH AIAN and Black or African American	98.87	98.54	98.96	98.05	99.36	0.07
NH Remainder of Two or More Races	86.12	85.80	86.35	85.22	87.94	0.39

Notes: The 2010-2012 ACS results use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.1.B Percent Citizens, 2010 CEF BR Sample

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS	BR in 2010-2012 ACS
Total	92.62	92.53	92.62	92.54	92.99	93.02
NH Asian Alone	69.37	68.89	69.36	69.44	69.69	71.14
Hispanic	67.75	67.46	67.74	67.20	71.34	71.63
NH White Alone	98.44	98.39	98.43	98.43	98.46	98.40
NH Black or African American	95.52	95.50	95.52	95.42	95.63	95.45
NH AIAN Alone	99.20	99.19	99.19	99.15	99.48	97.42
NH NHOPI Alone	81.84	81.71	81.84	81.34	83.07	81.82
NH Some Other Race Alone	73.16	73.01	73.15	72.49	75.11	75.55
NH AIAN and White	99.74	99.73	99.74	99.73	99.85	98.67
NH Asian and White	91.05	90.81	91.05	91.13	91.66	91.43
NH Black or African American and White	97.18	97.15	97.18	97.14	97.38	96.55
NH AIAN and Black or African American	98.96	98.95	98.96	98.92	99.31	97.80
NH Remainder of Two or More Race	86.83	86.67	86.83	86.66	88.34	87.88

Notes: The 2010-2012 ACS column uses the ACS citizenship values. BR in 2010-2012 ACS is the business rules used in the Hot Deck (using primary sources only) applied to the same 2010-2012 ACS records as in the 2010-2012 ACS column. The results in the last two columns use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.1.C Percent Citizens, 2010 CEF NBR-PIK Sample

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS
Total	87.25	8.44	69.47	24.93	68.91
NH Asian Alone	63.22	9.57	50.43	18.85	59.81
Hispanic	66.68	2.53	57.12	8.76	54.85
NH White Alone	96.86	9.99	75.25	30.80	73.33
NH Black or African American	86.60	8.01	70.86	20.68	70.08
NH AIAN Alone	100.00	30.91	85.67	46.54	82.74
NH NHOPI Alone	85.71	11.29	87.55	19.66	D
NH Some Other Race Alone	70.00	3.35	67.50	9.77	D
NH AIAN and White	D	19.04	94.15	51.50	D
NH Asian and White	90.91	7.79	91.47	31.38	D
NH Black or African American and White	85.71	8.85	67.63	27.19	D
NH AIAN and Black or African American	D	24.23	74.39	40.97	D
NH Remainder of Two or More Races	91.67	10.06	81.75	23.54	70.13

Notes: The 2010-2012 ACS results use ACS person weights. "D" signifies that the value is suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.1.D Percent Citizens, 2010 CEF NBR-SS Sample

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS
Total	75.30	76.70	66.98	64.83	73.08
NH Asian Alone	53.22	54.83	41.19	55.53	47.31
Hispanic	42.01	48.61	28.97	37.16	33.44
NH White Alone	95.87	94.89	93.63	83.26	95.13
NH Black or African American	91.37	89.50	88.52	70.35	90.93
NH AIAN Alone	98.36	92.90	99.20	80.48	98.79
NH NHOPI Alone	72.45	70.32	72.00	56.01	75.85
NH Some Other Race Alone	51.57	55.04	55.06	38.77	60.78
NH AIAN and White	99.59	96.54	99.48	92.80	99.48
NH Asian and White	82.36	82.59	64.86	75.33	69.16
NH Black or African American and White	95.31	91.09	93.68	78.07	91.70
NH AIAN and Black or African American	97.85	94.17	98.94	83.78	D
NH Remainder of Two or More Races	77.90	76.41	74.10	65.10	82.84

Notes: The 2010-2012 ACS results use ACS person weights. "D" signifies that the value is suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.1.E Percent Citizens, 2010 CEF NBR-SS Sample, Detailed Asian and Hispanic

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS
Asian Indian	46.98	49.72	36.85	53.04	40.66
Chinese	51.74	53.48	39.04	52.73	44.08
Filipino	62.94	61.56	54.53	61.02	63.29
Japanese	57.32	61.63	44.25	58.55	54.84
Korean	48.28	53.71	29.05	54.84	35.12
Vietnamese	64.45	59.68	62.69	61.75	65.32
Other Asian	53.58	54.59	39.71	55.52	47.35
Mexican	39.87	48.23	25.63	33.98	30.28
Puerto Rican	96.25	70.04	97.44	87.71	97.20
Cuban	59.94	60.41	58.33	49.72	62.76
Central American	28.17	39.33	15.28	31.59	19.16
Latin American	37.12	47.58	26.51	36.45	33.69
Other Hispanic	62.65	59.95	47.21	47.87	75.08

Notes: The 2010-2012 ACS results use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.1.F Percent Citizens, 2010 CEF NBR-NSS Sample

Category	Hot Deck	BR Logistic	ACS Logistic	Latent Class Model	2010-2012 ACS
Total	86.06	88.08	89.80	90.96	94.40
NH Asian Alone	54.74	62.79	67.82	68.85	76.15
Hispanic	54.61	63.14	61.70	68.55	69.19
NH White Alone	95.96	96.18	97.85	98.07	98.17
NH Black or African American	91.68	92.21	95.11	94.88	96.61
NH AIAN Alone	98.39	94.81	98.83	97.68	100.00
NH NHOPI Alone	76.74	77.89	90.61	84.63	D
NH Some Other Race Alone	59.88	66.33	83.59	73.69	D
NH AIAN and White	99.50	97.71	99.87	99.51	D
NH Asian and White	85.87	87.99	95.47	91.63	D
NH Black or African American and White	95.98	93.58	96.46	96.73	D
NH AIAN and Black or African American	98.73	96.31	99.19	98.42	D
NH Remainder of Two or More Races	83.78	84.27	96.69	88.47	D

Notes: The 2010-2012 ACS results use ACS person weights. "D" signifies that the value is suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.2.A 2010-2012 ACS Percent Citizens for NBR-SS Sample

Category	As-Reported	Edited	Imputed
Total	72.26	71.48	82.28
NH Asian Alone	44.79	55.04	70.10
Hispanic	31.47	56.18	55.76
NH White Alone	94.98	89.17	96.85
NH Black or African American	90.92	75.48	91.39
NH AIAN Alone	99.03	D	95.01
NH NHOPI Alone	75.73	100.00	74.97
NH Some Other Race Alone	57.74	D	84.77
NH AIAN and White	99.50	D	D
NH Asian and White	67.88	D	82.05
NH Black or African American and White	91.66	D	D
NH AIAN and Black or African American	D	NA	D
NH Remainder of Two or More Races	82.37	D	87.40

Notes: The results use ACS person weights. "D" signifies that the value is suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.2.B 2010-2012 ACS Percent Citizens for NBR-SS Sample, Detailed Asian and Hispanic Groups

Category	As-Reported	Edited	Imputed
Asian Indian	37.72	D	71.51
Chinese	41.65	55.98	66.74
Filipino	61.14	D	77.90
Japanese	53.39	D	72.06
Korean	32.99	D	60.17
Vietnamese	64.55	D	70.70
Other Asian	44.05	D	73.79
Mexican	28.49	45.94	52.67
Puerto Rican	98.30	D	85.44
Cuban	61.60	D	72.84
Central American	16.48	50.16	49.96
Latin American	31.12	D	57.51
Other Hispanic	75.98	D	70.93

Notes: The results use ACS person weights. “D” signifies that the value has been suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.3 Percent Citizens for 2010 CEF NBR-SS Group by Training Sample, Citizenship Source, and Specification

Category	ACS NBR-SS Sample, ACS Citizenship	ACS BR Sample, ACS Citizenship	ACS BR Sample, BR Citizenship	CEF BR Sample, BR Citizenship	Hot Deck	BR Logistic
Total	68.98	82.67	83.83	83.64	75.30	76.70
NH Asian Alone	41.19	57.51	62.02	61.05	53.22	54.83
Hispanic	28.97	46.37	50.31	49.68	42.01	48.61
NH White Alone	93.63	95.70	95.83	95.86	95.87	94.89
NH Black or African American Alone	88.52	91.75	92.15	92.03	91.37	89.50
NH AIAN Alone	99.20	99.26	96.21	98.17	98.36	92.90
NH NHOPI Alone	72.00	69.36	69.75	71.20	72.45	70.32
NH Some Other Race Alone	55.06	67.57	70.27	65.76	51.57	55.04
NH AIAN and White	99.48	99.64	96.99	99.17	99.59	96.54
NH Asian and White	64.86	85.31	85.62	85.00	82.36	82.59
NH Black or African American and White	93.68	95.76	93.78	94.10	95.31	91.09
NH AIAN and Black or African American	98.94	98.95	95.56	98.11	97.85	94.17
NH Remainder of Two or More Race	74.10	80.24	81.50	80.85	77.90	76.41

Notes: The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.3 Percent Citizens for 2010 CEF NBR-SS Group by Training Sample, Citizenship Source, and Specification Continued

Category	ACS NBR-SS Sample, ACS Citizenship	ACS BR Sample, ACS Citizenship	ACS BR Sample, BR Citizenship	CEF BR Sample, BR Citizenship	Hot Deck	BR Logistic
Asian Indian	36.85	53.45	60.21	58.95	46.98	49.70
Chinese	39.04	54.91	60.05	59.05	51.71	53.44
Filipino	54.53	52.02	55.30	53.72	62.93	61.56
Japanese	44.25	47.54	52.45	47.92	57.32	61.63
Korean	29.05	62.80	67.59	67.69	48.26	53.70
Vietnamese	62.69	71.03	72.72	73.33	64.44	59.67
Other Asian	39.71	54.58	58.77	57.32	53.57	54.59
Mexican	25.63	51.66	55.68	55.04	39.84	48.22
Puerto Rican	97.44	99.33	96.50	98.56	96.26	70.03
Cuban	58.33	77.31	78.83	80.24	59.97	60.42
Central American	15.28	38.02	42.55	40.98	28.15	39.31
Latin American	26.51	48.22	52.01	50.60	37.13	47.58
Other Hispanic	47.21	60.58	62.82	67.54	62.63	59.93

Notes: All estimates are applied to the 2010 CEF population frame. The first column estimates the ACS Logistic NBR-SS model on the 2010-2012 ACS NBR-SS sample with as-reported ACS citizenship as the dependent variable. The second estimates the model on 2010-2012 BR cases with as-reported ACS citizenship as the dependent variable. The third estimates the model on 2010-2012 BR cases with BR citizenship as the dependent variable. The fourth estimates the model on 2010 CEF BR cases. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.4.A Percent Citizens, As-Reported 2018 ACS Citizenship, By BR/PVS Status

Category	Hot Deck	BR Logistic	ACS Logistic	2018 ACS	Percent of ACS Sample
Total	92.67	92.87	92.33	92.15	100.00
BR	93.74	93.75	93.69	93.68	93.35
NBR-PIK	81.66	81.17	74.14	54.43	0.22
NBR-SS	77.46	80.50	73.00	71.06	6.38
NBR-NSS	88.88	89.62	94.89	91.53	0.05

Notes: The results use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.4.B Percent Citizens, As-Reported 2018 ACS Citizenship Full Sample

Category	Hot Deck	BR Logistic	ACS Logistic	2018 ACS
Total	92.67	92.87	92.33	92.15
NH Asian Alone	70.84	71.16	70.71	69.04
Hispanic	74.99	76.44	73.11	72.47
NH White Alone	98.46	98.43	98.42	98.43
NH Black or African American	95.65	95.40	95.69	95.63
NH AIAN Alone	97.43	96.82	97.50	99.65
NH NHOPI Alone	81.97	81.45	82.46	83.47
NH Some Other Race Alone	82.83	83.18	82.99	82.61
NH AIAN and White	99.29	99.21	99.27	99.89
NH Asian and White	92.06	92.08	91.29	91.49
NH Black or African American and White	97.81	97.61	97.75	98.57
NH AIAN and Black or African American	98.35	98.40	98.75	D
NH Remainder of Two or More Race	91.62	91.82	91.77	91.94

Notes: The results use ACS person weights. “D” signifies that the value has been suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.4.C Percent Citizens, As-Reported 2018 ACS Citizenship, BR Sample

Category	Hot Deck	BR Logistic	ACS Logistic	2018 ACS
Total	93.74	93.75	93.69	93.68
NH Asian Alone	72.06	72.14	71.70	70.63
Hispanic	78.49	78.49	78.45	78.29
NH White Alone	98.57	98.57	98.55	98.60
NH Black or African American	95.88	95.88	95.85	95.94
NH AIAN Alone	97.32	97.32	97.27	99.70
NH NHOPI Alone	82.12	82.32	82.01	84.04
NH Some Other Race Alone	84.77	84.80	84.68	85.20
NH AIAN and White	99.28	99.28	99.27	99.91
NH Asian and White	92.36	92.34	92.05	92.36
NH Black or African American and White	97.88	97.87	97.82	98.66
NH AIAN and Black or African American	98.75	98.72	98.72	D
NH Remainder of Two or More Race	92.23	92.26	92.18	92.60

Notes: The results use ACS person weights. “D” signifies that the value has been suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.4.D Percent Citizens, As-Reported 2018 ACS Citizenship, NBR-SS Sample

Category	Hot Deck	BR Logistic	ACS Logistic	2018 ACS
Total	77.46	80.50	73.00	71.06
NH Asian Alone	57.92	61.86	58.14	55.19
Hispanic	55.11	64.75	42.71	39.56
NH White Alone	95.96	95.08	95.99	95.33
NH Black or African American	93.03	90.02	94.08	92.34
NH AIAN Alone	98.38	92.90	99.46	99.29
NH NHOPI Alone	80.34	74.88	85.77	78.87
NH Some Other Race Alone	72.18	74.29	73.04	68.96
NH AIAN and White	D	97.74	99.70	99.42
NH Asian and White	85.13	86.55	75.95	74.17
NH Black or African American and White	96.56	93.21	96.93	98.00
NH AIAN and Black or African American	D	94.20	99.22	100.00
NH Remainder of Two or More Race	83.62	86.28	86.42	83.81

Notes: The results use ACS person weights. “D” signifies that the value has been suppressed due to disclosure restrictions. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.4.E Percent Citizens, As-Reported 2018 ACS Citizenship, NBR-SS Sample, Detailed Non-Hispanic Asian and Hispanic

Category	Hot Deck	BR Logistic	ACS Logistic	2018 ACS
Asian Indian	46.67	56.00	49.50	41.94
Chinese	58.06	60.50	49.92	48.91
Filipino	65.18	65.05	75.06	75.55
Japanese	65.63	69.36	66.34	63.60
Korean	54.94	64.62	55.61	58.21
Vietnamese	69.04	67.44	77.67	70.47
Other Asian	61.33	63.69	59.12	56.73
Unspecific Asian	62.65	69.48	79.33	68.86
Mexican	54.44	65.02	41.63	39.14
Puerto Rican	96.98	73.82	98.09	98.94
Cuban	68.04	66.84	61.81	56.76
Central American	42.50	60.55	25.42	20.93
Latin American	53.78	62.94	39.79	33.52
Other Hispanic	64.08	68.79	57.02	49.81
Unspecific Hispanic	74.07	70.69	77.37	70.40

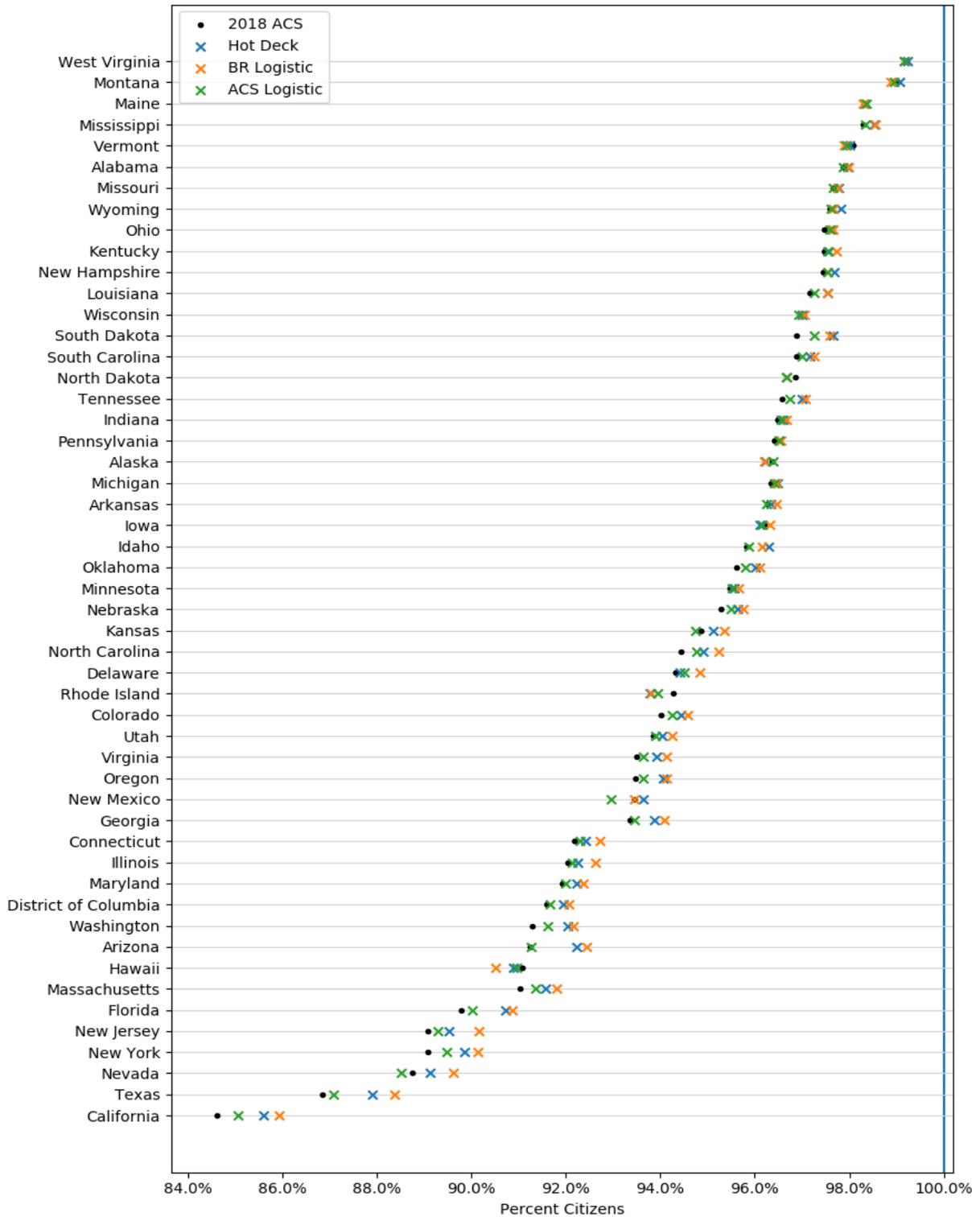
Notes: The results use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Table 10.5 Disagreement Rate with As-Reported 2018 ACS Citizenship

Category	Hot Deck	BR Logistic	ACS Logistic
Total	2.80	3.08	2.53
BR	1.38	1.40	1.34
NBR-PIK	30.92	32.54	35.86
NBR-SS	22.47	26.51	18.86
NBR-NSS	12.42	13.33	8.23

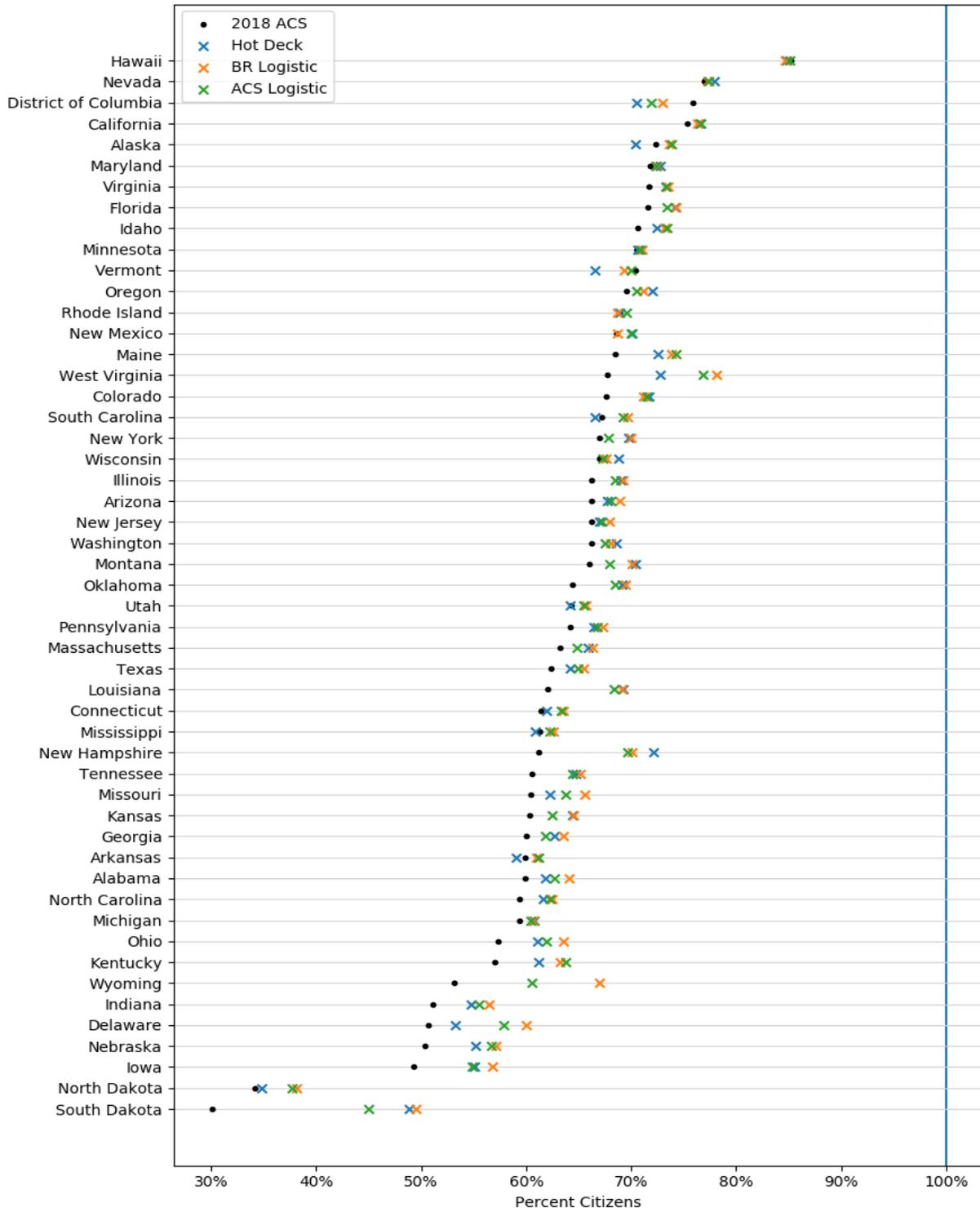
Notes: The results use ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 10.1.A 2018 ACS Test State Citizen Shares, All Races



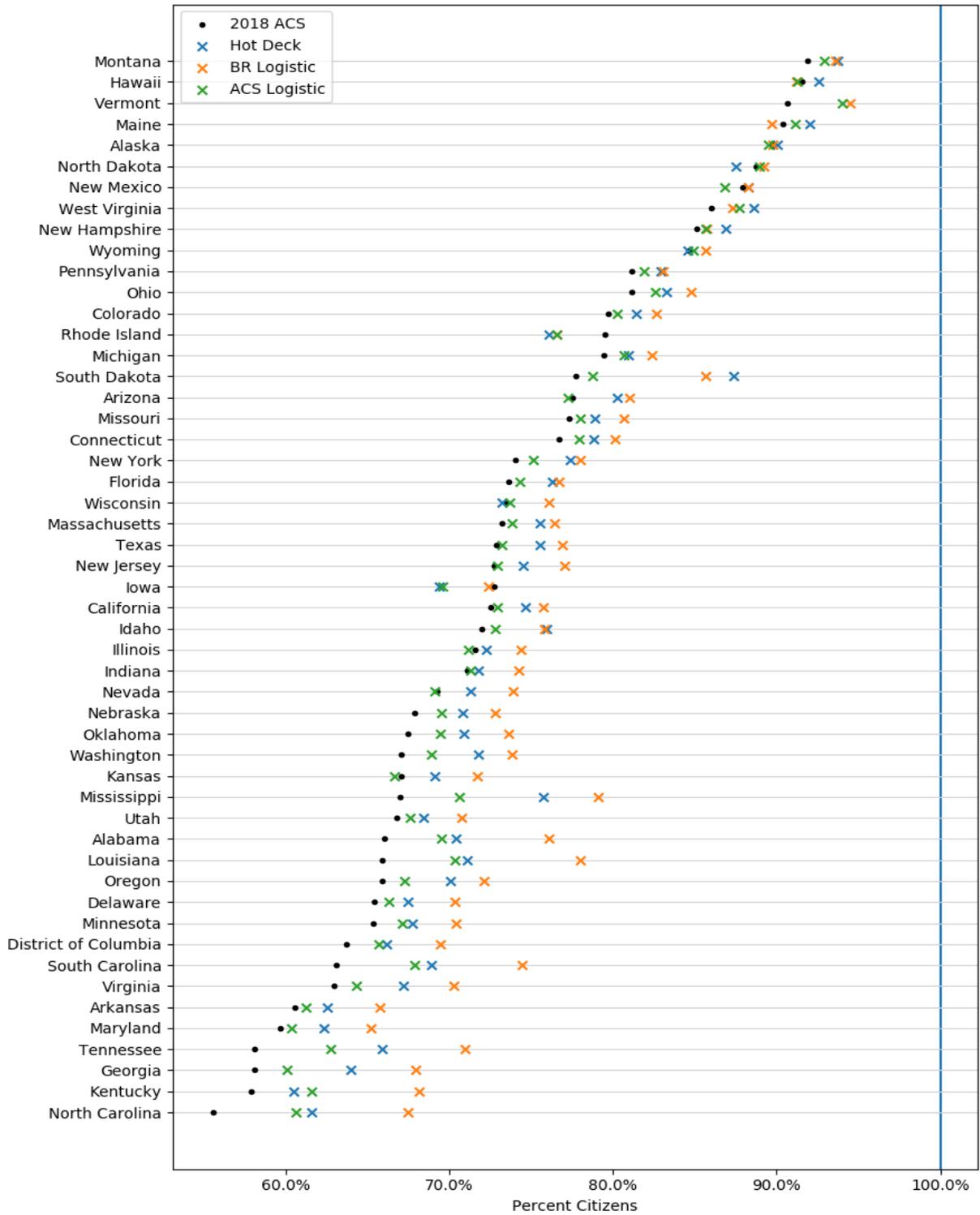
Notes: These shares are weighted using ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 10.1.B 2018 ACS Test State Citizen Shares, Non-Hispanic Asian



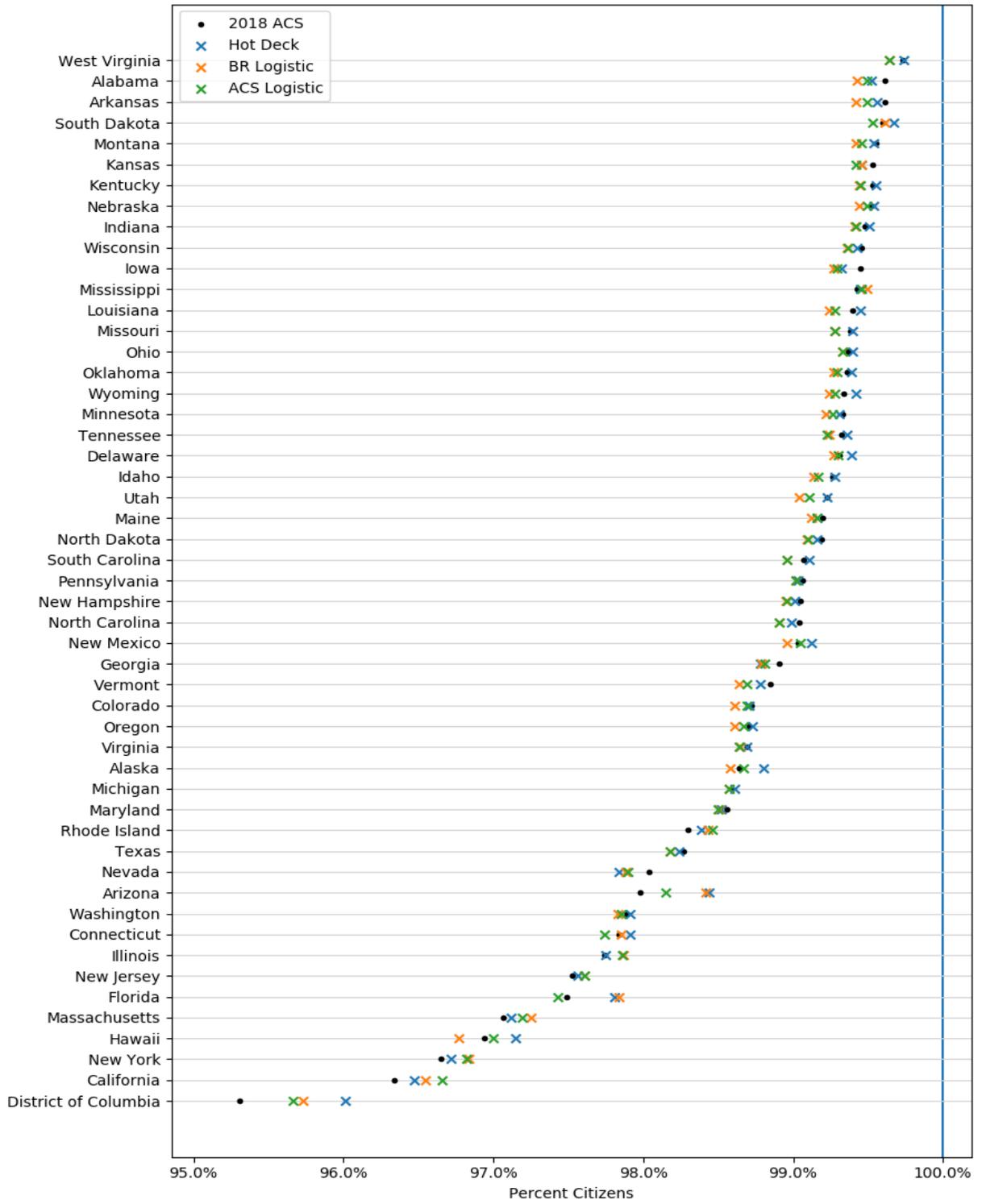
Notes: These shares are weighted using ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 10.1.C 2018 ACS Test State Citizen Shares, Hispanic



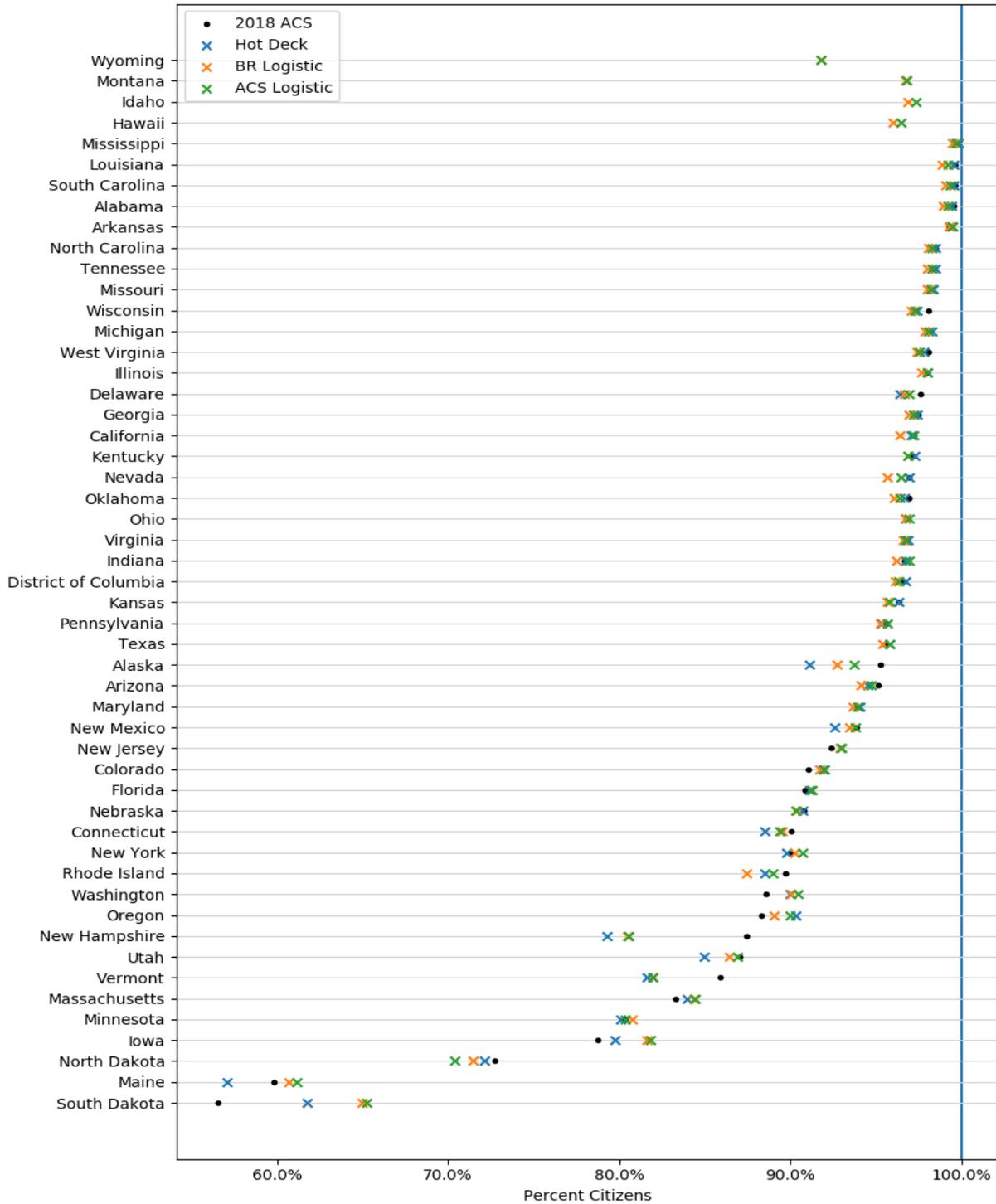
Notes: These shares are weighted using ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 10.1.D 2018 ACS Test State Citizen Shares, Non-Hispanic White



Notes: These shares are weighted using ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

Figure 10.1.E 2018 ACS Test State Citizen Shares, Non-Hispanic Black



Notes: These shares are weighted using ACS person weights. The data presented in this table are approved for dissemination by the DRB (CBDRB-FY21-CED002-B0001).

11. Disclosure Avoidance

11.1 Disclosure Avoidance and Census Data

Title 13, Section 9 of the U.S. Code prohibits the Census Bureau from publishing any information in a manner that may be used to identify the information provided by any census or survey respondent. In order to produce official data products derived from its censuses and surveys, while meeting this legal requirement to protect confidentiality, the Census Bureau has historically relied upon the application of statistical disclosure avoidance methods to alter the published data sufficiently to mitigate the risk that individual respondent data could be reliably re-identified. From the 1990 Census through the 2010 Census, this process involved the introduction of “noise,” or statistical uncertainty, into the data via the swapping of entire households’ records across geographies. Other procedures, including aggregation combined with suppression based on table population thresholds, and synthetic data may have been used in those censuses.⁷⁰

Recognizing the growing privacy threats posed by the proliferation of external data sources that may be leveraged to attempt the re-identification of respondents, and improvements in the computing algorithms that can reconstruct individuals’ records from aggregate, tabular data making linkage of these records to external data easier, the Census Bureau’s Data Stewardship Executive Policy Committee (DSEP) determined that the data swapping methods used in prior censuses are no longer sufficient to protect the confidentiality of census records. On April 5, 2019, DSEP decided that the Census Bureau will use new, mathematically provable, disclosure avoidance techniques for noise injection based on differential privacy for all 2020 Census public data releases.

11.2 Differential Privacy

The Database Reconstruction Theorem, also known as the Fundamental Law of Information Reconstruction, first introduced by Irit Dinur and Kobbi Nissim in 2003, demonstrates that the calculation of any statistic from a confidential data source reveals a tiny amount of private information about the confidential data. If you release too many statistics, at too high a degree of accuracy, then after a finite number of tabulations you will reveal all of the underlying confidential information used to create the tabular summaries. Differential privacy, first conceptualized by Cynthia Dwork et al. in 2006, provides a framework for quantifying this leakage of private information, and in doing so, enables its mitigation through the injection of precisely calibrated amounts of noise. Consequently, differential privacy as an approach to disclosure avoidance allows for quantifiable, future-proof privacy guarantees. These guarantees are set through the establishment of a privacy-loss budget (PLB) and its allocation to each tabular summary. Under this approach, any statistic, tabulation, or calculation to be performed against the confidential data will have a certain amount of noise added to it. The precise amount of injected noise is a function of that query’s allocation of the PLB and the potential impact of any individual’s contribution to the result of that query (the query’s sensitivity). For the 2020 Census, DSEP will establish a global PLB for all 2020 Census Data Products and will allocate that PLB across the various tabulations

⁷⁰ See McKenna (2018) for a history of the publicly-released details of these procedures

necessary to produce the entire array of 2020 Census Data Products, including the proposed 2020 CVAP tables at the block level.

11.2 The 2020 Disclosure Avoidance System

To apply differential privacy to the 2020 Census, the decennial data products were grouped publications into collections of tabular summaries according to their planned publication schedule. The Group I data products include the PL94-171 redistricting data, the Demographic Profiles, and the Demographic and Housing Characteristics data. For these products, the Census Bureau will use the Disclosure Avoidance System's TopDown Algorithm (TDA). The CVAP product is a special tabulation scheduled for production after the redistricting data have been finalized and released, using the TDA technology.

One of the primary requirements for DAS integration into the 2020 Census production process is that the DAS must use the Census Edited File (CEF) microdata as its input, and must output a microdata file with a pre-specified schema (the Microdata Detail File or MDF) for use by Decennial's tabulation system.

As the DAS does not operate on microdata records directly, the system represents the microdata during processing as a functionally equivalent "histogram," which is a matrix of record frequencies, in which every cell reflects a unique combination of variable attributes that a microdata record might contain. Statisticians and social scientists often call such a data structure a "fully saturated contingency table with structural zeros removed."

Throughout this process, the DAS holds a few tabulations invariant; that is, they are preserved as enumerated on instruction from DSEP. While the final list of invariants for the 2020 Census Data Products has not yet been established by DSEP, the current list of invariants includes total population at the state level, number/type of group quarters facilities (not populations) at the block level, and number of housing units (not populations) at the block level.

Once the CEF has been input into the DAS, the system takes an extensive series of measurements (cross-tabulations) the outputs of which are injected with noise from a probability distribution determined by the PLB and the query's sensitivity.

The TDA computes all of these noisy measurements at once, regardless of the level of geography represented by the statistic. It also computes the invariants once, regardless of the level of geography. Armed with these noisy and invariant measurements, the DAS has no further access to the CEF. The DAS then solves for a national-level histogram of record frequencies that reflect those measurements (noisy and invariant) at the national level.

At present, the CVAP will be processed after the PL94-171 redistricting data are finalized and before the balance of the Group I products are ready. Therefore, the histogram available to the DAS is the one embodied in tables P1-P5 and H1 of the PL94-171 redistricting data.

The noisy measurements, taken alone, do not satisfy the usual properties of a histogram: all entries must be non-negative integers. TDA solves a multi-level, multi-pass set of optimization problems

to combine the noisy measurements and invariants into the best non-negative integer representation.

The TDA proceeds from the U.S.-level histogram to the state-level histogram using a similar process. At this stage, the TDA incorporates the national-level histogram into the set of invariants and constraints, then uses the noisy measurements taken at the state-level to solve for the set of histograms for each state (plus DC)⁷¹ that reflect the state-level characteristics, while satisfying the invariants, constraints, and national totals.

In this manner, TDA divides each cell in the U.S-level histogram into 51 cells representing the frequency of that unique combination of attributes for each state and the District of Columbia. Similar optimizations are used to ensure that each state-level histogram is the best non-negative, integer representation of the noisy measurements and invariants.

This process continues down the TDA's geographic hierarchy, using noisy measurements for that level along with the invariants, constraints, and histograms from the higher geographic levels to generate expanded histograms for each successive geographic level. At each level, optimizations are used to ensure that each geographic-level histogram is the best non-negative, integer representation of the noisy measurements and invariants.

Once the block-level histograms have been fully specified, the TDA then converts these histograms into microdata by generating individual records for each cell of the histogram according to the frequency count contained in that histogram cell, with each of those records reflecting the combination of attributes for that cell. The resulting microdata is the Microdata Detail File (MDF), which then enters the 2020 Census tabulation system used to generate the official PL94-171 redistricting data.

11.3 The CVAP Special Tabulation

On October 3, 2019, DSEP decided that production of the CVAP data product would not be integrated into the production of the Group I data products. Instead, the Census Bureau will apply disclosure avoidance to the CVAP data product as a special tabulation of the 2020 Census data after the PL94-171 data have been processed through the DAS. Privacy protections for the CVAP product, as with all 2020 Census data products, will be governed by a dedicated share of the global PLB to be set by DSEP.

Once the DAS has produced the MDF for PL94-171, the CVAP Implementation Team will provide the DAS Team with a person-level file containing two data elements: (1) correct CEF person id; and (2) best citizenship, expressed as a probability, for the universe of persons included in the CEF. Best citizenship is coded "not available" if the enumerated individual is under 18 years of age on the CEF.

The DAS Team will then run the CVAP data through the DAS TDA process, outlined above, using Table P4 from the PL94-171 redistricting data as an additional set of constraints on the TDA's

⁷¹ TDA processes Puerto Rico in a separate hierarchy. Its resident population is not included in the resident population of the United States. The resident population of Puerto Rico is also invariant.

optimization. The output of this special tabulation will be a differentially private person-level file containing just the information required to tabulate CVAP at the designated geographies, including block-level.

Since the final production settings for the PLB will not be available before December 2020, this technical paper contains no analysis of the effects of the DAS on the final CVAP tabulations. No demonstration data product is planned. The effects of disclosure avoidance will be included with any published measures of uncertainty for CVAP data.

12. Recommendations for Experimental Citizen Voting-Age Population Data Products

12.1 Overview and Authority

The 2020 Census Methods Internal Expert Panel (IEP) was charged with recommending a method to produce the highest quality Citizen Voting Age Population (CVAP) data product possible by combining population data from the 2020 Census with citizenship data from various available sources, including administrative and survey sources. This is in line with Department of Commerce (DOC) Secretary Wilbur Ross’s direction of March 26, 2018, the 2020 Census Office of Management and Budget (OMB) Paperwork Reduction Act Clearance Package of December 28, 2018,⁷² and the Presidential Executive Order of July 11, 2019 titled *Executive Order on Collecting Information about Citizenship Status in Connection with the Decennial Census*.⁷³ In collaboration with the Census Bureau’s Redistricting and Voting Rights Data Office, the 2020 Methods IEP determined the content and format for the updated experimental CVAP data products. This defined the statistical estimand: the quantity that the Census Bureau’s methods are trying to estimate.

The IEP met on a regular basis from July 2018 to the present, reviewing the efforts of the 2020 CVAP Technical Working Group. The working group explored four alternative approaches for using multisource data in the production of CVAP statistics. Three of these approaches started with “business rules” for using the citizenship data sources to assign citizenship data to census records. Two experiments, one using 2010 Census data and the other using 2018 ACS data, combining these data sources with corresponding administrative and survey sources appropriate for the two years, found the business rules (BR) could assign citizenship to just over 90 percent of the population. These assignments are believed to be very accurate as they used what are believed to be accurate recordings of citizenship status from the data sources, and avoided using inaccurate citizenship data (such as outdated records). The assignments also were based on linkages across data sources that were assessed to be reliable. The BR assignments left just under 10 percent of cases for whom citizenship status required statistical estimation.

The three approaches pursued to augment BRs with statistical estimation were (i) impute citizenship status of the non-BR (NBR) cases using donors from the BR cases, (ii) predict probabilities of citizenship status for the NBR cases using logistic regression models fitted to the BR cases, and (iii) predict probabilities of citizenship for the NBR cases using logistic regression models fitted to ACS records that could not be given BR citizenship assignments, but that did have reported citizenship. Strengths and weaknesses of the three approaches were reviewed, and their empirical results for the experiments were compared. This formed the basis for the recommendations given here.

⁷² “Accordingly, the Secretary has directed the Census Bureau to proceed with the 2020 Census without a citizenship question on the questionnaire, and rather to produce Citizenship Voting Age Population (CVAP) information prior to April 1, 2021 that states may use in redistricting.” For more information, see OMB PRA 2020 Census Supporting Statement A (full revised final), submitted July 3, 2019, approved July 12, 2019 (<https://www.reginfo.gov/public/do/DownloadDocument?objectID=88197702>).

⁷³ For more information, see: <https://www.whitehouse.gov/presidential-actions/executive-order-collecting-information-citizenship-status-connection-decennial-census/>.

The working group also explored a fourth approach, latent class (LC) modeling, that uses a multivariate model to combine information from multiple citizenship data sources to produce predicted probabilities of citizenship for all person records. Despite not using explicit business rules, the LC modeling was found to produce citizenship estimates for the BR cases that were very close to those from the BR assignments, providing strong confirmation for the BRs. LC estimates for the NBR cases were also broadly reasonable compared to those from the other approaches. However, while the LC modeling has some advantages compared to the other three approaches, certain detailed effects found in the logistic regression modeling for detailed population subgroups could not be fully replicated in the LC model without enhancements to the computer software for fitting the model. While intensive work has been done on these enhancements, they are not complete as of this writing, and this work is ongoing.

12.2. Recommendations

Based on the CVAP Technical Working Group's evaluations of the data sources, empirical results from the four estimation approaches, and CVAP production considerations, the IEP makes the following recommendations.

1. The BRs used for citizenship assignment, which differed slightly across the approaches, can provide accurate citizenship estimates for the census cases that can be reliably linked to the administrative and survey data sources. In the experiments done, differences in these results across the three approaches were minor. The IEP thus recommends proceeding by developing a single harmonized set of BRs as follows:
 - a. Persons are classified as citizens if they are citizens in the SSA NUMIDENT, have a U.S. passport or USCIS naturalization certificate, or don't have SSA NUMIDENT citizenship but are U.S.-born in those data.
 - b. Persons lacking that information are deemed noncitizens if noncitizens in the SSA NUMIDENT, SEVIS, WRAPS, IMARS, LEMIS, BOP, USMS, Nebraska or South Dakota driver's license data, NCRP, ACS, AHS, CPS, or SIPP; have a nine-digit taxpayer ID number in the ITIN range; noncitizens in USCIS data with better record linkage quality; or noncitizens in ADIS with better linkage quality and more recent vintage.
 - c. If none of the above apply, then persons are treated as citizens if they are citizens in ADIS, BOP, USMS, Nebraska or South Dakota driver's licenses, SNAP/TANF, NCRP, ACS, AHS, CPS, or SIPP.

Statistical estimation will be required to estimate citizenship for the cases not covered by the BRs.

2. The ACS logistic method is the preferred method for the production of the 2020 CVAP experimental data products, subject to the caveat listed in 2.b. below.
 - a. The IEP believes that this method best addresses the non-ignorable missing data issue that arises when BR cases of linkable citizenship information are used to develop predictors of citizenship probabilities for the NBR cases. By training models on ACS records that also lack linkable citizenship information, but have as-

reported ACS citizenship responses, the ACS logistic method helps address this issue, especially for those cases with sufficient personally identifiable information (PII) to be sent to search for a PIK.

- b. The evidence about non-ignorable missing data is less strong for the NBR cases with insufficient PII to be sent to search for a PIK, and the IEP recommends that the working group perform further study of these cases. Another reason for further study is the possibility that the size of this group in the 2020 Census could be larger than was the case in the 2010 Census (when it was about 3.3 percent of the population). The IEP recommends that the working group investigate enhancements to the use of logistic regression with either the BR cases or ACS cases, and perform further evaluations of the results. A final recommendation on treatment of these cases will be made following this additional investigation.
 - c. For the cases that received a PIK, but for which no citizenship status could be assigned, the estimates differed across the alternative approaches. However, the IEP recognizes that this is a very small group of records, with little impact on the overall estimates, and with no clear reason to expect significant growth of this group in 2020. Therefore, the IEP recommends that this group be combined with one of the other two NBR groups, based on an assessment of evidence of non-ignorability in this small population.
3. The IEP recognizes that LC modeling is a promising approach for producing CVAP estimates. However, given the need to enhance the software to accommodate LC models with the desired detail, and the fact that this enhanced software is still under development, the IEP recommends that the LC approach be used for evaluation, via comparisons made to the results from the recommended approaches, and not for the CVAP production at this time. The LC approach should also be examined for its ability to produce uncertainty measures (standard errors) for citizenship estimates.
 4. Any newly received citizenship data sources not covered by the tests in this report should be evaluated for use based on the same methods applied to the sources that are included in this report.
 5. The Census Bureau should continue to enhance and develop improved record linkage for the production of official statistics using multisource data, including the production of enhanced CVAP statistics.
 - a. The PVS reference files should be expanded to include records in government sources that have sufficient PII, but have not received a PIK when attempting to link to the current production PVS reference files. This facilitates linkage for individuals without SSNs or nine-digit taxpayer IDs in the ITIN range.
 - b. Record linkage quality measures derived from PVS module, pass, and score information should be used when evaluating records' fitness for use.

References

- Agresti, A. (1992). "Modelling patterns of agreement and disagreement," *Statistical Methods in Medical Research*, 1(2):201–218.
- Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. Hoboken, NJ: John Wiley & Sons.
- Bakk, Z. and Kuha, J. (2018). "Two-step estimation of models between latent classes and external variables," *Psychometrika*, 83(4): 871-892.
- Bandeem-Roche, K., Miglioretti, D.L., Zeger, S.L. and Rathouz, P.J. (1997). "Latent variable regression for multiple discrete outcomes," *Journal of the American Statistical Association*, 92(440): 1375-1386.
- Bedrick, E. J., Christensen, R., and Johnson, W. (1996). "A new perspective on priors for generalized linear models." *Journal of the American Statistical Association*, 91(436):1450-1460.
- Biemer, P. P. and Wiesen, C. (2002). "Measurement error evaluation of self-reported drug use: a latent class analysis of the us national household survey on drug abuse," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165(1):97–119.
- Biemer, P. P., Woltmann, H., Raglin, D., and Hill, J. (2001). "Enumeration accuracy in a population census: An evaluation using latent class analysis," *Journal of Official Statistics*, 17(1):129.
- Brown, J. David, Misty L. Heggeness, Suzanne M. Dorinski, Lawrence Warren, and Moises Yi (2019a). "Understanding the Quality of Alternative Citizenship Data Sources for the 2020 Census," *Center for Economic Studies Working Paper Series, No. 18-38R (June)*. Washington, DC: US Census Bureau.
- Brown, J. David, Misty L. Heggeness, Suzanne M. Dorinski, Lawrence Warren, and Moises Yi (2019b). "Predicting the Effect of Adding a Citizenship Question to the 2020 Census," *Demography*, **56**, 1173-1194.
- Chung, H., Loken, E., and Schafer, J. L. (2004). "Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution." *The American Statistician*, 58(2):152-158.
- Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86(413): 68-78.
- CVAP Working Group (2019), "Producing Federal Statistics on Citizenship and the Citizen Voting-Age Population (CVAP) Tables Using Multi-Source Data," (Final Report), June 20, 2019.
- Dayton, C.M. and Macready, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401): 173-178.
- Fellegi, Ivan P., and Sunter, Alan B. (1969). "A Theory for Record Linkage," *Journal of the American Statistical Association*, Vol. 64 (328): pp. 1183-1210.
- Formann, A. K. and Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5(2):179–211.
- Goodin, Brett. (2020). "Americans Are Renouncing U.S. Citizenship in Record Numbers, But Not For The Reasons You Think." *The Conversation*, available at <https://theconversation.com/americans-are-renouncing-us-citizenship-in-record-numbers-but-maybe-not-for-the-reasons-you-think-145365>, last accessed October 2, 2020.
- Heitjan, D. F. and Rubin, D. B. (1991). "Ignorability and Coarse Data." *The Annals of Statistics*, 19(4): 2244-2253.

- Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430): 773-795.
- Kendler, K. S., Karkowski, L. M., and Walsh, D. (1998). The structure of psychosis: latent class analysis of probands from the roscommon family study. *Archives of General Psychiatry*, 55(6):492-499.
- Kreuter, F., Yan, T., and Tourangeau, R. (2008). Good item or bad—can latent class analysis tell? the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3):723-738.
- Layne, M., Wagner, D., and Rothhaas, C. (2014). Estimating record linkage false match rate for the person identification validation system. *Center for Administrative Records Research and Applications (CARRA) Working Paper Series*, #2014-02. Washington, DC: U.S. Census Bureau.
- Lazarsfeld, P. F. and Henry, N.W. (1968). *Latent Structure Analysis*. Houghton Mifflin Co.
- Little, Roderick J. A. and Rubin, Donald B. (2002). *Statistical Analysis with Missing Data*. Wiley-Interscience.
- McCutcheon, A. L. (1987). *Latent Class Analysis*. Sage Publications, Thousand Oaks, CA.
- McKenna, L. (2018), “Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing,” Working Paper, U.S. Census Bureau, Research and Methodology Directorate, available at <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/DisclosureAvoidanceforthe1970-2010Censuses.pdf>, last accessed 10/1/2020.
- Prosser, R. J., Carleton, B. C., and Smith, M. A. (2008). Identifying persons with treated asthma using administrative data via latent class modelling. *Health Services Research*, 43(2):733-754.
- Rastogi, Sonya, and Amy O’Hara (2012). “2010 Census Match Study,” 2010 Census Planning Memoranda Series No. 247, U.S. Census Bureau.
- Richardson, S. and Green, P.J. (1997). “On Bayesian analysis of mixtures with an unknown number of components (with discussion).” *Journal of the Royal Statistical Society, Series B*, 59(4):731-792.
- Schafer, J.L. (2020). *cvam: Coarsened-variable modeling*. R package, version 0.9.1. To become available at the Comprehensive R Archive Network, <https://cran.r-project.org>.
- Wagner, Deborah, and Mary Layne (2014). “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research & Applications’ (CARRA) Record Linkage Software,” *Center for Administrative Records Research and Applications (CARRA) Working Paper Series*, #2014-01, U.S. Census Bureau.