Distr.: General
19 January 2023

English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Migration Statistics

Geneva, Switzerland, 26−28 October 2022
Item A of the provisional agenda
Improvements in use of administrative data for migration statistics

Evaluating Coverage of the US Census Bureau's Integrated Database for International Migration (IDIM)

Note by U.S. Census Bureau*

Abstract

In recent years the US Census Bureau's International Migration Branch has explored using administrative data to improve foreign-born international migration flow estimates, resulting in a linked database called the Integrated Database for International Migration (IDIM). A limitation of the IDIM is that it is restricted to persons covered by Federal administrative record data sets acquired by the Census Bureau resulting in coverage gaps for the foreign born. We assume the IDIM underestimates, or misses, specific groups of foreign born not covered by the linked data sources including working migrants who did not file tax returns, non-working dependents not claimed as exemptions on tax returns, international students, exchange visitors, and unauthorized migrants. The American Community Survey (ACS) is a large annual household survey conducted by the US Census Bureau that, in theory, should cover these populations missing from the IDIM. To examine IDIM coverage limitations, we match individual records from the ACS to the IDIM.

This paper presents results of exploratory research to assess IDIM coverage and evaluates the magnitude and characteristics of ACS survey respondents who are not included in the IDIM, as opposed to those who are present in the IDIM. Further, this research serves to assess data quality of both the ACS and IDIM, as we can compare results between linked individuals and households within each data source. These findings provide us with information on the potential for using the ACS to help adjust for IDIM coverage limitations.

*Prepared by Jason Schachter, Esther Miller, and Angelica Menchaca. The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. CBDRB-FY23-POP001-0007

I. Introduction

1. The International Migration Branch at the US Census Bureau produces annual estimates of foreign-born immigration flows to the United States with demographic detail (age, sex, race, and Hispanic origin) at national, state, and county-level geographies. These estimates are created using the American Community Survey (ACS) as a primary data source, though survey estimates come with limitations such as increased variance, particularly at subnational geographies, and lagged measurement of migration events. To help overcome these limitations, the Census Bureau has been developing an alternative data source called the Integrated Database on International Migration (IDIM), which incorporates available administrative data from social security and tax records to estimate international migration.

2. One limitation of the IDIM is that it is restricted to persons covered by Federal administrative record datasets, resulting in coverage gaps for the foreign born who are not included in these records. This means the IDIM likely underestimates, or misses, specific groups of foreign born, such as working migrants who did not file tax returns, non-working dependents not claimed as exemptions on tax returns, international students, exchange visitors, and most unauthorized migrants. The ACS is a large annual household survey conducted by the US Census Bureau that is representative of the entire resident population, and thus theoretically should cover these missing IDIM populations. While the ACS likely underrepresents some of these missing foreign-born groups as well, we can still examine IDIM coverage limitations by matching individual records from the ACS to the IDIM. This allows us to compare individuals who are included in both the ACS and the IDIM ("ACS match" or "ACS-IDIM match") to those in the ACS but not in the IDIM ("ACS only," or "ACS-IDIM non-match"). In addition to this comparison, for individuals both in the ACS and IDIM, we can evaluate data quality for a number of variables shared across both data sets, including age, sex, citizenship status, year of entry, and current geography. Finally, based on these findings, this paper discusses the potential of using the ACS to adjust IDIM results to improve its estimates.

II. IDIM overview

3. The IDIM is created by linking administrative data sources which can be used to generate foreign-born immigration estimates. While there are many administrative data sources which could potentially be used in the IDIM, such as those maintained by the Department of Homeland Security (DHS), we are currently limited to data sources readily available at the Census Bureau, namely the Numident from the Social Security Administration (SSA) and tax filing information from the Internal Revenue Service (IRS). Data linking is done via matching of unique Personal Identification Keys (PIKs) which are assigned to individuals across data sets. PIKs are most easily created using directly matched encrypted Social Security Numbers (SSNs), but are also created by probabilistically matching name, sex, age, and address information.

A. Numident

4. The Numident is a micro-record dataset that combines SSA SSN records with Census Bureau death records. It includes data on demographic characteristics, place of birth, and citizenship status. It does not, however, include address data. SSNs can be easily anonymized using PIKs which allow for linking across datasets. Given that most documented immigrants to the United States apply for SSNs, the Numident was chosen to serve as IDIM's spine for initial data integration and research.

5. Numident data are delivered on a quarterly basis and contain records for all persons who have ever received SSNs. In addition to native births, this includes applications for SSNs by the foreign born. Foreign-born individuals who are either authorized to work or have become naturalized citizens are eligible to receive SSNs. Using a combination of citizenship status, place of birth, and date of record creation, we can identify foreign-born migrants at the national level by demographic characteristics. Linking Numident data to other sources can give us additional information, such as place of residence, which would allow us to create sub-national

estimates of the foreign born. It also provides information about "signs of life," which give us additional confidence as to whether the social security holder has moved to the United States for the requisite period of time to establish residency.

B. Internal Revenue Service Tax Filings

6. The IRS provides tax form 1040 filing data to the Census Bureau every four weeks. While these data do not include demographic characteristics, they do include address information and PIKs for the primary filer, spouses and dependents. Where possible, address data are linked to a Master Address File ID (MAFID). IRS data do not include information on foreign-born status, which must come from linked Numident files. They also do not include information for individuals who do not file taxes (either through not having sufficient income or for failing to claim income). SSNs included on tax filing data make for the possibility of directly matching individuals to the Numident.

7. It is also possible to identify tax filings that use Individual Tax Identification Numbers (ITINs), unique identifiers used by individuals without SSNs to file taxes. ITINs are only issued to non-US citizens; thus it is not necessary to link these individuals to the Numident to assign foreign-born status. While this universe is assumed to include migrants unauthorized to work, it also includes non-working dependents of authorized migrant workers. Changes to tax laws in 2017 caused drastic decreases in reported ITINs, as spouses and dependents are no longer eligible for ITINs unless they qualify for specific deductions or file their own separate return.[1] This significantly reduces the usefulness of ITINs for identifying migrants after 2017. Further complicating use of ITINs is that individuals are periodically required to reapply for ITINs, and thus someone can have multiple ITINs over the course of a lifetime. In addition, ITINs holders can apply for SSNs later in life, and thus could have both an ITIN and SSN on file. As ITINs are not registered with the Social Security System, they cannot be directly matched to the Numident, though probabilistic methods could be used to match ITIN records to other data sets. The current version of IDIM does not include ITIN holders, but additional research is being conducted to see how they can possibly be incorporated into future analyses.

C. IDIM Creation

8. As noted earlier, the Numident acts as our spine for identifying foreign-born immigrants. It is used in the first phase of processing whereby the foreign born are identified using citizenship variables from the Numident (this includes non-citizens authorized to work and naturalized citizens). We then use record creation year as a proxy for year of entry into the US. Lastly, we remove individuals who died the same year they migrated. This step results in an estimate of foreign-born immigrants by year with demographic characteristics, albeit a clear overestimation. We expect an overestimation at this point, as this universe includes SSN applicants who received SSNs, but who either only came to the United States for a short period of time or never actually migrated to the United States. The native-born population is retained in the file to have a comparison group to the foreign born. Race and Hispanic origin data are incomplete or missing from the Numident,[2] so it is necessary to use alternative methods to assign race and Hispanic origin by modeling decennial Census 2010 and ACS files. These methods to assign race to the Numident have not been incorporated into this paper, which limits analysis for these variables.

9. In the second phase of processing, we match Numident records to IRS tax form 1040 filings to confirm entry into the United States. The Numident contains all applications for SSNs, including individuals who received SSNs, but never actually migrated to the United States or only stayed for a short period of time. To remove this group from our estimates, we match IRS data to restrict the universe to authorized migrants who worked and filed taxes in the United States, as well as both working and non-working naturalized citizens. This step also

[1] See https://www.irs.gov/individuals/individual-taxpayer-identification-number.

[2] Why Researchers Now Rely on Surveys for Race Data on OASDI and SSI Programs: A Comparison of Four Major Surveys (ssa.gov)

assigns geocodes, giving us foreign-born immigrants with demographic characteristics at national, state, and county geographies. We expect an underestimation of the foreign-born non-citizen population at this point, as we are missing migrants who fail to file taxes, as well as authorized migrants who did not work.

10. At this stage the IDIM includes the following immigrant populations: naturalized citizens, non-citizens authorized to work and who filed taxes, and their non-working dependents and spouses. Populations not included are: US citizens born abroad of American parents, unauthorized migrants, working migrants who did not file a tax return, and non-working dependents not claimed as an exemption. Given the ACS is designed to be representative of the entire US resident population, it should include information on many of the foreign-born groups currently missing on the IDIM.

D. ACS

11. This paper links 2019 ACS micro data to the IDIM to help evaluate IDIM's coverage, as well as data quality of both IDIM and the ACS. The ACS is a large annual continuous household survey of the US population that asks detailed information previously collected on the decennial census long form. Fully implemented in 2005, it currently surveys about 3.5 million addresses per year. Inclusion in the sample is based on having lived, or planning to live, for at least two months in the sampled address. The ACS asks detailed sociodemographic and economic questions, including immigration-pertinent variables such as country of birth, citizenship status, year of entry to the United States, and country of residence one year ago. While SSN information is not collected on the ACS, individuals on the ACS can be assigned PIKs using the Person Verification System (PVS), which assigns probability by matching name, sex, and address information.

12. Since the ACS includes all US residents in its sample universe, and does not distinguish by legal status, we feel the ACS is a potentially good source of information on migrants missing from the IDIM. However, given the hard-to-count nature of recent and unauthorized migrants, it is likely the ACS underrepresents these populations to some degree (Jensen et al, 2015). While it is not our intent to evaluate ACS coverage in this paper, this issue should be kept in mind when interpreting some of our findings.

E. Person Identification Validation System (PVS)

13. The PVS is the Census Bureau's process to identify and verify SSNs and PIKs for person records in surveys, censuses, and administrative records. The Census Bureau attempts to assign PIKs to every administrative record via a probabilistic model known as the Person Verification Model (Wagner and Layne, 2014) that is composed of four modules. First, if the administrative data contain SSNs, the verification module checks for an exact SSN match to the Numident file and verifies that name and date of birth elements sufficiently agree. If they do agree, the SSN is considered verified and PVS assigns the corresponding PIK to the person record. If there is no SSN, such as in the case of the ACS, the PVS continues through three more probabilistic modules to attempt to assign an SSN to the administrative record using geography, name, and date of birth. Approximately 94% of all 2010 ACS records received a PIK, implying that only 6% of all records could not be linked to any administrative data.

14. Those not assigned PIKs can potentially introduce bias to linked data. A study of the 2009 and 2010 ACS concluded that PVS is less likely to validate young children, minorities, residents of group quarters, immigrants, recent movers, low-income individuals, and non-employed individuals (Bond, 2014). In addition to unassigned PIKs, there is the possibility of PIKs being erroneously assigned to individuals (also referred to as record linkage error), though the rate of these misassigned PIKs for the foreign born is not known (Abowd et al., 2020). This is one possible reason for mismatch between variables on the IDIM and ACS, when individuals are matched incorrectly.

F. Linking the ACS to IDIM

15. For this paper, we linked 2019 ACS data to the 2019 IDIM universe. Since the IDIM universe is defined by linked Numident and IRS records, which have undercoverage of older populations, we restricted our universe to those under 65 years of age to make the ACS universe more comparable. For our ACS universe, 14% are identified as foreign born. To define foreign born on the ACS, we use responses to the ACS citizenship question. Those born in the US and born abroad of American parents are defined as "native born," while US citizens by naturalization and non-US citizens are defined as "foreign born." The proportion of foreign born in the IDIM is 12.6%. The foreign born are defined similarly to the ACS using a citizenship variable, due to data quality concerns with the country of birth and foreign-born indicator variables on the Numident file. In the IDIM's case, we use a variable that identifies "US citizens" and "legal aliens," in combination with a variable that denotes if a person was ever naturalized. Neither the ACS nor IDIM definitions of the foreign born disaggregate this group by citizenship status, which is important to note due to data quality concerns for the naturalization variable on both the ACS (Van Hook and Bachmeirb, 2013) and Numident.

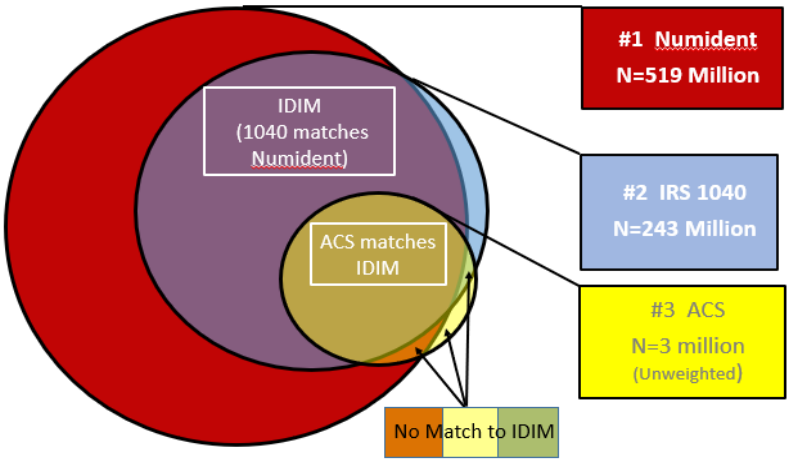
16. For the total ACS universe, 89% of the sample can be assigned an individual PIK, and thus matched to the Numident or assigned an ITIN by the IRS. Among the foreign-born identified on the ACS, 79% of the sample can be assigned a PIK. When further linked to IRS/SSN data, these match rates drop, with 80% of the total ACS universe matched via PIK and 71% of the foreign-born sample. These drops in match rates are expected, since the IDIM is limited to tax filers with SSNs. The lower match rate for the foreign-born universe was expected based on previous research discussed earlier, thus contributing to foreign-born undercoverage in the IDIM.

17. Conversely, the non-match rate for the ACS foreign-born universe is 29%, which provides us with a key comparison group. This universe consists of three distinct groups: those in the ACS for which a PIK is not able to be assigned, those with ITINs who file taxes, and those linkable to the Numident, but who did not file taxes or appear as exemptions on IRS tax returns. Future research will attempt to disaggregate the ACS not matched to IDIM universe, but it is assumed to include a large proportion of unauthorized migrants, as well as groups like international students and dependents not claimed as exemptions on tax returns, hence those assumed to be missing from the IDIM universe.

III. Comparison of IDIM and ACS Foreign-Born Universes

18. There are several different universes that can be compared to evaluate IDIM coverage. Figure 1 is a conceptual diagram of how the different universes are created, linking the Numident, IRS and ACS. The IDIM consists of Numident and IRS matches, and currently excludes those with ITIN records (light blue, outside the Numident and IDIM). The figure also denotes the important comparison groups used in this analysis: (1) total IDIM (purple), (2) the ACS-match or ACS-IDIM match group (shaded yellow), and (3) the ACS-only or ACS-IDIM non-match groups. The ACS-only group consists of three sub-groups: non-PIKable ACS files (light yellow), ACS respondents on the Numident but not on the IDIM (orange, e.g., non-tax filers), and ACS respondents on the IRS but not on the Numident (green, e.g., ITIN holders).

Figure 1. Numident, IRS, IDIM, and ACS Evaluation Universes



Sources: 2020 Census Numident, IRS 1040 TY19, 2019 American Community Survey

19. For comparison purposes, Table 1 shows detailed demographic and socioeconomic characteristics for each of these foreign-born universes. All these characteristics can be derived from the ACS, while a limited number of variables are also available on the IDIM, including sex, age, citizenship status, year of entry, and state of current residence. Thus, for the ACS-IDIM match group, some variables can be derived independently from both the ACS and IDIM, which will allow us to evaluate data quality in the following section.

(See Table 1 in appendix)

20. As expected, we found clear differences between the IDIM/ACS-IDIM matched and the ACS-only universes. The ACS-only universe was more male, younger (under 25), less Asian, more Hispanic (Mexican and Central American), less educated, and more not in labor force and in poverty. There were few differences in terms of year of entry between the IDIM and the ACS-only group. Comparisons between ACS-only and ACS-matched universes showed more extreme differences between groups than total IDIM comparisons. Of particular interest were differences in terms of citizenship status and year of entry.

21. Sex and age distributions can be gleaned from both the IDIM and ACS for the various groups. The ACS-only universe was more male than both the IDIM and the ACS-IDIM-matched universes. The IDIM had a generally older age distribution (50 and older), while the ACS-only group skewed younger, including more college aged and children in its universe.

22. As discussed earlier, race and Hispanic origin are currently not measurable on the IDIM, so this comparison was limited to the ACS-only and ACS-matched groups. Clear differences were found, with far more Asians in the ACS-matched universe, and far more Hispanics in the ACS-only universe. Among Hispanics, far more were of Mexican or Central American descent in the ACS-only universe compared to the ACS-matched universe.

23. Socioeconomic variables are only available on the ACS and clear differences were seen between the ACS-only and ACS-matched universes. In terms of education, the ACS-only universe was far more likely to have less than a high school degree, while the ACS-matched universe was more likely to have at least a college degree. Relatedly, the ACS-matched group was more likely to be employed and not in poverty than the ACS-only group.

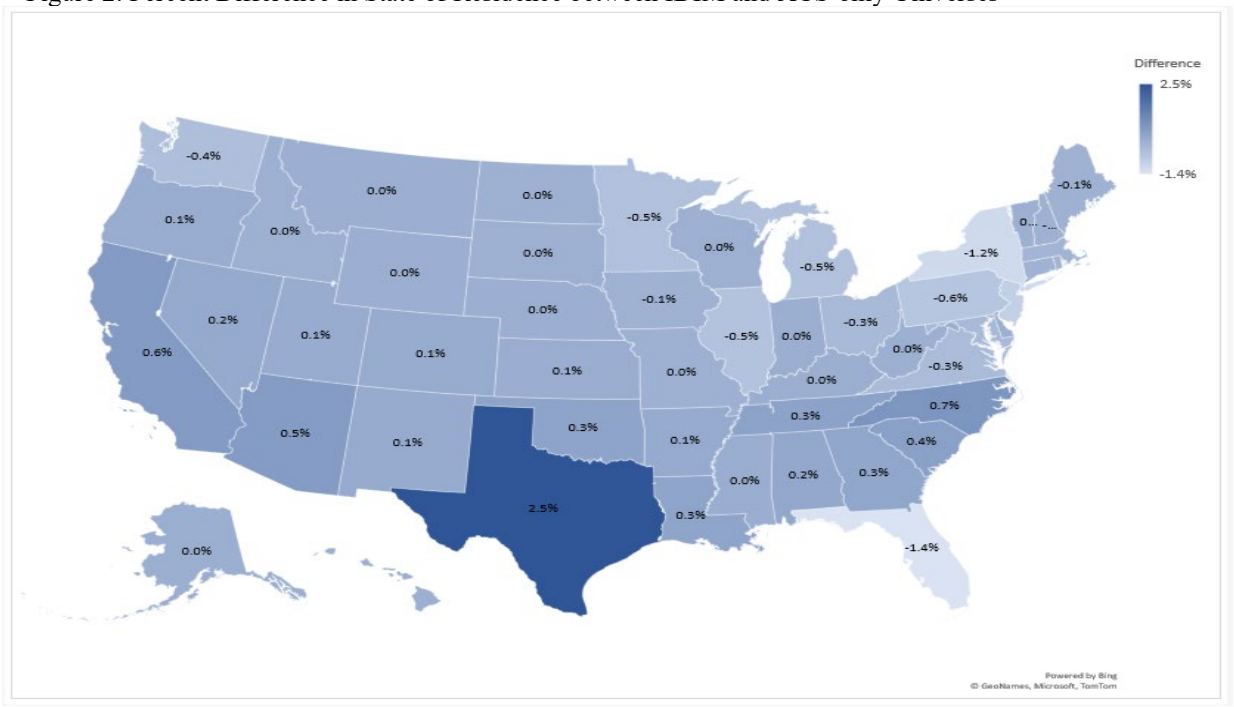
24. Some interesting findings were discovered looking at the year of entry and citizenship variables present on both the ACS and IDIM. Year of entry is defined on the IDIM as the year when an SSN was entered into the Numident, while the ACS asks respondents which year they came to live in the United States, so we would expect to see differences between datasets. This was not the case between the ACS-only and IDIM universes, as their year of entry distributions are quite similar. However, when comparing the ACS-only to ACS-matched universes, which use the same variable of measurement for year of entry, differences do appear. The ACS-only universe was more likely to have been recent migrants (since 2015) than the ACS-matched universe, who were more likely to have arrived before 1999. Given the similar year of entry distribution between the IDIM and ACS-only universes, there were some surprising differences between the ACS-IDIM matched IDIM-based year of entry distributions and the ACS-IDIM matched ACS-based year of entry distributions, which will be examined in greater detail in the following section.

25. Citizenship status also produces interesting results, which bring to question the quality of this variable on the IDIM. The foreign born on the IDIM are far more likely to be non-citizens than the ACS-matched universe, which was an unexpected result. This suggests that IDIM records are not updated on a regular basis after an individual naturalizes or suggests that ACS reporting are of poor quality for this variable (or a combination of both reasons). More telling is the comparison between the ACS-only to ACS-matched universes, where large differences are found. The proportion of non-citizens is far higher for the ACS-only universe, while conversely the proportion of naturalized foreign born is far greater for the ACS-matched universe. This suggests that the ACS-only universe is more representative of persons ineligible for SSNs, including unauthorized migrants and dependents of legal migrants.

26. Specific citizenship status is not relevant from the perspective of how we use the IDIM to produce estimates of the foreign born. It is not important if someone on the IDIM has accurate up-to-date naturalization status information, since what is important is whether they are foreign born or not. However, these findings could have important implications for other types of analysis using administrative data, and would likely require additional linkages to other data sources (e.g. from US Citizenship and Immigration Services (USCIS)) to accurately measure citizenship status.

27. Finally, Figure 2 shows the difference in state of residence for those on the ACS-only and IDIM files. Differences were relatively small, with the IDIM having a bit more representation in states like Florida and New York and the ACS-only universe having a bit more representation in Texas.

Figure 2. Percent Difference in State of Residence between IDIM and ACS-only Universes



Sources: US Census Bureau, Integrated Database for International Migration and 2019 American Community Survey; Social Security Administration; and Internal Revenue Service

28. These findings suggest there is clear underrepresentation of specific foreign-born groups in the IDIM. Despite the ACS's own potential coverage biases for the foreign born, it appears to better measure the hard-to-count foreign-born populations missing from the IDIM, and thus should be useful to adjust undercoverage of specific migrant groups in the IDIM. Some of the underrepresentation seen in the IDIM could be addressed by incorporating ITINs, but this has its own set of challenges that still need to be worked out. In lieu of linking additional administrative data sources to IDIM, integrating the ACS into IDIM to account for some of this undercoverage seems feasible. This is discussed in more detail later in the paper.

IV. Data Quality Analysis

29. The initial analysis brought up some questions about both IDIM and ACS data quality, specifically for the citizenship and year of entry variables. To provide insight into data quality, we compare shared ACS- and IDIM- derived variables for individuals who are matched to both data sets. As discussed previously, these shared variables include sex, age, citizenship status, year of entry, and state of current residence.

30. Results for ACS- and IDIM-derived variables for both sex and age are promising. There is very strong concurrence for individual sex variables derived from both the IDIM and ACS, with over 97% of male respondents and 98% of female respondents reporting the same sex. For age, we would expect to see differences based on how age is defined in each data source. The ACS asks for age at the time of survey (in addition to date of birth), while age on the Numident is based on the mid-point of the year. This appears to be the case, as only 48% of respondents have the same age on both files. However, when we expand the age to plus or minus one year, the concurrence rate increases to 95%. While we cannot determine which data source has more accurate values, age heaping is a known issue for survey-based age responses, and yet the overall age agreement between data sources seems to be high.

Table 2. Citizenship Status Reported on the IDIM and ACS for Matched Respondents

Citizenship Status			
	Reported on ACS		
	Native	Naturalized	Non-Citizen
Reported on IDIM			
Foreign-born Non-Citizen	9%	35%	56%
Naturalized	10%	84%	6%

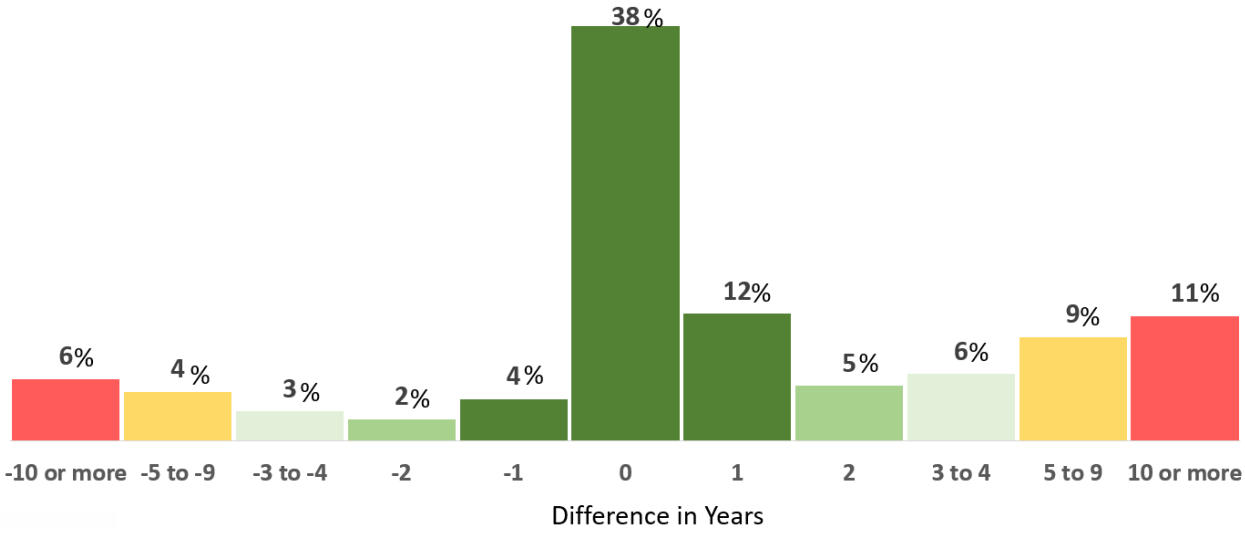
Note: Values are unweighted

Sources: US Census Bureau, Integrated Database for International Migration and 2019 American Community Survey; Social Security Administration; and Internal Revenue Service

31. Citizenship status is measured very similarly on both the ACS and IDIM, with foreign-born respondents being disaggregated as either naturalized or non-citizens. Earlier results suggested a disconnect between these variables as derived by the IDIM and ACS, which is confirmed by this analysis. For the foreign-born on the IDIM who are identified as non-citizens, only 56% of those matched to the ACS report the same non-citizen status with the ACS citizenship variable, while 35% are naturalized and 9% are native born per the ACS citizenship question. For naturalized foreign born on the IDIM matched to the ACS, there is better concurrence, with 84% being naturalized on the ACS question and only 6% being non-citizens, while 10% are native born. The large differences between ACS- and IDIM-derived citizenship status could be indicative of lack of updates to naturalized status on SSA records, but also could reflect inaccurate data reporting on the ACS. Also concerning is that close to 10% of the Numident foreign born are classified as native born on the IDIM. Looking at the foreign born on the ACS, similarly, over 10% are classified as natives on the ACS. Further cross-tabulations by country of birth could help elucidate some of these findings. One partial explanation for this misconnect could be miscategorization of the native born from the "Born Abroad of American Parents" ACS citizenship question, which likely includes those whose parents were not naturalized citizens at the time of respondent's birth, due to confusing question wording on the survey questionnaire. High imputation associated with the citizenship variable and mismatched PIKs are other possible contributing factors, in addition to erroneous ACS and IDIM data responses. More research would be needed to thoroughly investigate this issue though.

32. As discussed earlier, we would expect incongruence between the IDIM and ACS year of entry values, given the different ways this variable is measured on the two data sets, as well as data quality concerns with the ACS variable regarding accurate recall and year heaping in responses. Analysis shown in Figure 3 confirms incongruence between the year of entry variable on the IDIM and ACS for matched individuals. The year of entry values from these two data sets only match 38% of the time. If we expand this range to within one year of each other, this only improves to 54% of cases, while plus or minus two years improves this to 61%. Even with a range of plus or minus 9 years, the year of entry values match just 83% of the time between data sets. Again, to what extent this is mostly due to data reporting issues (for both the ACS or the Numident), high imputation, and/or PIK record linkage error is not known. More research on the year of entry and citizenship status variables is clearly warranted.

Figure 3. Distribution of Differences in Responses to the Year of Entry Question Among Matched Individuals in the ACS and the IDIM



Source: US Census Bureau, Integrated Database for International Migration, 2019 American Community Survey; Social Security Administration; and Internal Revenue Service

33. Finally, for matched ACS-IDIM individuals we look at the reported state of residence on the IDIM and the ACS. The ACS geography variable comes from the location where the survey respondent resided at the time of inclusion in the survey, while IDIM geography comes from where the individual filed their tax return. It is possible that a person made an interstate move during the measurement period, so we would expect some differences on this variable between data sets. Evaluation of this variable shows relatively high congruence with 92% of ACS and IDIM geographies matching at the state level for linked individuals. Differences could easily be caused by interstate moves during the period, though less likely due to ACS imputation since this variable comes from the sampled address list.

34. In summary, differences between some IDIM- and ACS-derived variables for linked individuals were unexpectantly large. As mentioned during the discussion of the citizenship status and year of entry variables, one possible explanation is that data are reported incorrectly on each data set. There is also the possibility that high imputation for foreign-born specific variables like citizenship, place of birth, and year of entry on the ACS further contributes to these differences. Additionally, there may be record linkage errors between IDIM and ACS, and thus they are not the same individuals, which is possible given the probabilistic method used to assign PIKs in the absence of SSN information. This is another area where future research is needed to allay possible concerns about data quality in the IDIM and the ACS.

V. Using the ACS to adjust IDIM

35. The purpose of this exploratory research was not only to evaluate IDIM coverage and data quality, but also to provide us with information about whether the ACS could be used to adjust the IDIM for its confirmed coverage limitations. Despite potential biases in foreign-born unit and item response, it appears the ACS does measure foreign-born populations missing from the IDIM, namely unauthorized migrants and informally employed migrants who do not file taxes, international students, and some dependents of IRS tax filers.

36. The US Census Bureau produces net international migration flow estimates for the nation, state, and county by age, sex, race and Hispanic origin, primarily using ACS data. Development of the IDIM was not done with the

intent of replacing the ACS, but rather is an effort to draw from the strengths of each dataset through data integration, thereby improving our estimates. The IDIM could be particularly useful for improving county-level estimates, for which our survey-based estimates are reliant on 5 years of pooled ACS foreign-born stock data and still have high levels of sampling variability, particularly for smaller counties. At the same time, there is still the potential of using the ACS to adjust IDIM undercoverage for both national and county-level estimates, as well as for national and subnational characteristics.

37. The Census Bureau has previously integrated administrative data at the macro-level to improve national survey-based estimates of migration to and from Puerto Rico after Hurricane Maria, as well as to account for the impact of the COVID-19 pandemic on international migration flows to and from the United States. These methods used historical trends between ACS and administrative data to adjust ACS estimates based on levels seen with administrative data. Informing adjustments to the IDIM with the ACS would be an instance of using survey data to adjust administrative records and could potentially occur at both the macro- and micro-level, given the nature of linking procedures.

38. For example, from a macro-integration perspective, the ACS could be used to adjust for missing international student populations, as well as age distributions at the subnational level, by adding a proportion of students to the national totals, or by applying ACS county-level age distributions for counties with large student populations. It could also be possible to use the levels and characteristics of the ACS-only population to account for missing IDIM populations, either through proportional or modeled estimate adjustments. These methods could help account for some missing unauthorized migrants, as well as other foreign-born populations missing from the IDIM. Further work to better disaggregate the ACS-only population into different categories would also improve the nuances of any adjustments made for this population.

39. From a micro-integration perspective, it could be possible to use linked householders on the IDIM and ACS, for which information about family members is on the ACS but not the IDIM, to adjust for missing dependents who are not included on tax returns. Knowing the size of family on the ACS and to what extent this population is missing on the IDIM could inform some probabilistic estimation methods. All these macro- and micro-data integration methods would still need to be developed, but these initial findings suggest that the ACS could be a useful tool to improve migration estimates produced by the IDIM.

VI. Discussion

40. As this paper illustrates, there is still much work to be done to improve coverage of the IDIM and its estimates. Next steps include adding race and ethnicity data to the IDIM though an established method used by other areas of the Census Bureau—namely, using matched information from the 2010 and 2020 Census on race/ethnicity to assign values, as well as modeling missing information on new migrants from ACS country of origin race distributions. This would provide us with the ability to derive all characteristics needed to produce our migration estimates from the IDIM. The application of the IDIM to produce subnational county-level estimates needs to be further evaluated, even if questions about data coverage persist.

41. Further work on the potential for adding ITINs to the IDIM through a process that does not duplicate individuals could be beneficial and greatly help improve coverage of the unauthorized migrant population. The IDIM also underestimates young children. This may be partially resolved by refining our imputation processes for non-matching dependents but will need further investigation. This paper provided additional insight into this underestimated population, and as discussed earlier, further disaggregation of the ACS-only population would improve our understanding the IDIM. Use of additional data sources could also help in this endeavor.

42. Linking the IDIM to other data sources, such as files provided by the United States Citizenship and Immigration Services or the Department of Health and Human Services, would be very helpful. Data sharing agreements are being developed with these agencies and could provide invaluable information, not just for missing populations, but for verifying and improving data quality on the IDIM.

43. Similarly, though we do not currently have access to data from US Immigration and Customs Enforcement, the Student and Exchange Visitor Program would be ideal for estimating student and exchange visitor flows. Arrival and Departure Information System data from Customs and Border Protection could help us measure unauthorized flows. These are examples of potential data sources which could be incorporated into the IDIM at a future date.

44. In addition to improving IDIM coverage, further work should be conducted to better understand data quality, such as for the citizenship and year of entry questions. In any event, the US Census Bureau will continue to attempt to develop and integrate administrative sources with survey data to improve our net international migration estimates.

References

Abowd, J., William R. Bell, J. David Brown, et al. (2020). Determination of the 2020 US Citizen Voting Age Population (CVAP) Using Administrative Records and Statistical Methodology. Center for Economic Studies Working Paper Series No. 20-23. Washington, DC: US Census Bureau.

Bond, B., J.D, Brown, A. Luque, and A. O'Hara. (2014). The Nature of the Bias when Studying Only Linkable Person Records: Evidence from the American Community Survey. Center for Administrative Records Research and Applications Working Paper Series No. 2014-08. Washington, DC: US Census Bureau.

Brown, J. David, Misty L. Heggeness, Suzanne M. Dorinski, Lawrence Warren and Moises Yi. (2019). Predicting the Effect of Adding a Citizenship Question to the 2020 Census. *Demography* 56:1173–1194.

Jensen, Eric b., Renuka Bhaskar, and Melissa Scopilliti. (2015). Demographic Analysis 2010: Estimates of Coverage of the Foreign-Born Population in the American Community Survey. US Census Bureau Working Paper No. 103. Washington, DC: US Census Bureau.

Luque, A., and R. Bhaskar. (2014). 2010 American Community Survey Match Study. Center for Administrative Records Research and Applications Series Working Paper No. 2014-03. Washington, DC: US Census Bureau.

Rastogi, S., and A. O'Hara. (2012). 2010 Census Match Study Report. 2010 Census Planning Memoranda Series No. 247. Washington, DC: US Census Bureau.

Van Hook, Jennifer and James D. Bachmeier. (2013). How Well Does the American Community Survey Count Naturalized Citizens? *Demographic Research* 29(1): 1–32.

Wagner, D., and M. Layne. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. Center for Administrative Records Research and Applications Working Paper Series No. 2014-01. Washington, DC: US Census Bureau.

Appendix

Table 1. Demographic and Socioeconomic Characteristics for the IDIM and ACS Universes				
Foreign-Born Population				
	ACS as the Base		**IDIM as the Base**	
Demographic Characteristics	**ACS Records Matched to IDIM**	**ACS Records not Matched to IDIM**	**IDIM Administrative Records**	**IDIM-subset for Matched ACS Records**
Sex				
Male	48%	52%	48%	47%
Female	52%	48%	52%	53%
Age				
0-17	6%	10%	5%	5%
18-24	7%	10%	8%	7%
25-34	19%	21%	20%	18%
35-44	25%	25%	24%	24%
45-54	25%	20%	24%	26%
55-64	20%	15%	18%	20%
Race				
White Alone	54%	67%	X	X
Black Alone	12%	10%	X	X
Asian Alone	31%	19%	X	X
Other	3%	4%	X	X
Non-Hispanic	60%	39%	X	X
Hispanic				
Mexican	21%	37%	X	X
Central American/ Dominican Republic	9%	16%	X	X
Other	10%	8%	X	X
Poverty Status				
Not In Poverty	91%	78%	X	X
In poverty	9%	22%	X	X
Employment Status				
Employed	78%	64%	X	X
Unemployed	3%	3%	X	X
Not in Labor Force	19%	33%	X	X
Education				
Less than high school	20%	37%	X	X
High School	21%	26%	X	X
Some college / College graduate	42%	29%	X	X
Post grad	17%	8%	X	X

Citizenship Status				
Non-Citizen	46%	71%	72%	67%
Naturalized	54%	29%	28%	33%
Year of Entry				
Before 1990	23%	17%	17%	19%
1990 to 1999	24%	20%	22%	25%
2000-2009	27%	29%	25%	26%
2010 to 2014	12%	12%	16%	15%
2015 and later	13%	23%	20%	15%
N (in thousands)	26,450	11,040	29,690	283

Note: ACS values are weighted.
Sources: US Census Bureau, Integrated Database for International Migration, 2019 American Community Survey; Social Security Administration; and Internal Revenue Service