Creating a New Data Infrastructure for Foreign-Born Scientist and Engineers: Data, Analysis and Use

Nathan Barrett, Yunie Le,
Ekaterina Levitskaya, and Allison Nunez

Coleridge Initiative
Final Report
Project No. ADC-FBSE-22-05

Creating a New Data Infrastructure for Foreign-Born Scientist and Engineers: Data, Analysis and Use

Project Agreement Holder
Ahu Yildirmaz, Ph.D.

Project Team Technical POC
Nathan Barrett, Ph.D.
1740 Broadway
New York, NY 10019
859-396-0987
nathan.barrett@coleridgeinitiative.org

Submitted:
06/30/2023

Creating a New Data Infrastructure for Foreign-Born Scientist and Engineers: Data, Analysis and Use

Table of contents

Executive Summary

Science and engineering fields are vital to U.S. economic growth, national defense, infrastructure, and overall public and private well-being. While the U.S. has long been a world leader in science and engineering, foreign-born individuals have, and will continue to be, key domestic contributors to this sector. Given the importance of this sector and the contributions of foreign-born individuals, it is imperative that we understand the training they receive, the investments in that training, their workforce outcomes, and the connections between. Yet we do not currently have a robust data infrastructure that can support the evaluation of these connections as the data that currently exist are disparate and often not comparable or linkable. At the same time, there are a host of policies that need to leverage such an infrastructure as they seek to better understand and support this sector and the foreign-born scientists and engineers that are vital to its success.

The long-term objective of this project is to establish the foundations of a national data infrastructure to address unanswered questions about foreign-born scientists and engineers in the United States, beginning with estimating the return on investment (ROI) for U.S. training of foreign-born scientists and engineers (FBSEs). The current report focuses on the feasibility of constructing a linked data infrastructure using existing state administrative data and federal data sources and outlining the complexity of estimating a return on investment for FBSEs in the United States.

This report approaches this need by first documenting the current state of various administrative data systems and surveys, what those data can and can't say about foreign-born individuals and FBSEs. With these gaps and challenges identified, the report then lays out feasible and aspirational data models that provide a comprehensive data infrastructure to study any number of policy related questions including the ROI. The data models must be supported by a comprehensive record linkage strategy that is discussed in detail. Finally, we provide recommendations on a data governance framework that balances the preservation of privacy while supporting a robust research agenda.

While some general patterns and trends emerge from the various surveys included in the report, taken together, the surveys demonstrate several key limitations to understanding the educational outcomes of foreign-born scientists and engineers. First, many of the surveys were not designed to be representative of such a specific group of individuals, FBSEs. Compounding this issue is that, for many states, the surveys are not designed to be representative at the state level. This poses challenges to better understanding the local and regional context that FBSEs may face which are important to designing more effective programs and policies. In some cases, the surveys use slightly different definitions for the variables of interest, such as foreign-born vs. citizenship, which can make comparisons across surveys difficult. In cases where there are fewer concerns about representation, such as the American Community Survey (ACS) data, the results are not able to determine the country in which the respondent earned their credential, and for advanced degrees, cannot account for the field of study. Perhaps one of the more significant limitations with surveys is the ability to link with other sources of data to enhance the ability to comprehensively understand the intersection between investments and outcomes for this important group of individuals.

With those limitations in mind, the report then turns to state-level administrative data. The postsecondary administrative data presents an opportunity to provide a state-level census of foreign-born graduates. Given that the federal government, as well as state governments, have invested significant resources into state longitudinal data systems it is important to evaluate how they can be used to answer questions about such an important group of individuals. At the same time, significant effort would need to be made to integrate information across states to provide a national picture. This effort would need to be supported by a common data model, guidance for record linkage, and a governance structure.

Data models are helpful in clearly defining the necessary information required to answer questions for which current data infrastructure cannot answer. They indicate sources, the attributes required from those sources, how those attributes are defined and constructed, and how the disparate sources will link together. Two data models were developed with input from expert advisory panels: a feasible data model using the current state longitudinal system's data and an aspirational model. The feasible data model has several limitations, including an inability to identify individual occupations, and the view on income is limited to what is included in state administrative wage data. Additionally, there are limitations in using social security as a linkage identifier for the foreign-born population that is less likely to have that specific identifier. The aspirational data model leverages the feasible data model but adds additional sources to attend to the limitations. Namely, some states have developed data systems that incorporate state income tax data. Here, foreign-born taxpayers without a valid social security number, and who would not appear in the state wage data, could have their income included in the data. Other outcomes such as patents, licensure, business ownership, employees, and grants could all be included to provide a more robust outcomes portfolio.

Record linkage for foreign born populations poses some unique challenges, and this work seeks to propose actionable recommendations for assessing and improving record linkage performance and bias for the foreign-born population. This work also has much broader potential benefits to other populations and to administrative record linkage in general. Key findings suggest that, across data systems, there are few common fields available from which to generate linked data. While some data have attributes such as date of birth, others do not, leaving first and last name as primary fields available to develop matches. With few fields available to match, record linkage approaches should employ both deterministic and probabilistic approaches and be transparent as to the performance and possible bias of the linked data. There should be awareness and education efforts to train users of linked administrative data on the existence, impact, measurement, and mitigation of record linkage error and bias as well as how to communicate record linkage methods, performance, and bias. Analysts using the data should be clear about how record linkage bias may influence their results. At the same time, data administrators should strive to augment current data systems to include additional fields so that match rates and potential bias can be improved.

Building a robust data infrastructure is more than addressing the technical issues of identifying and linking disparate data and developing common data standards that support it. Whether developed at the federal level, the state level, or a combination of the two, the data infrastructure must be supported by an equally robust governance framework that facilitates cross-agency data sharing and access. A key first step in developing a robust data infrastructure is to assess the current state of disparate data, the existing gaps, and what data is needed to attend to those gaps. Once the data systems needed are identified, all data stewards that govern each data system must be brought together so that a

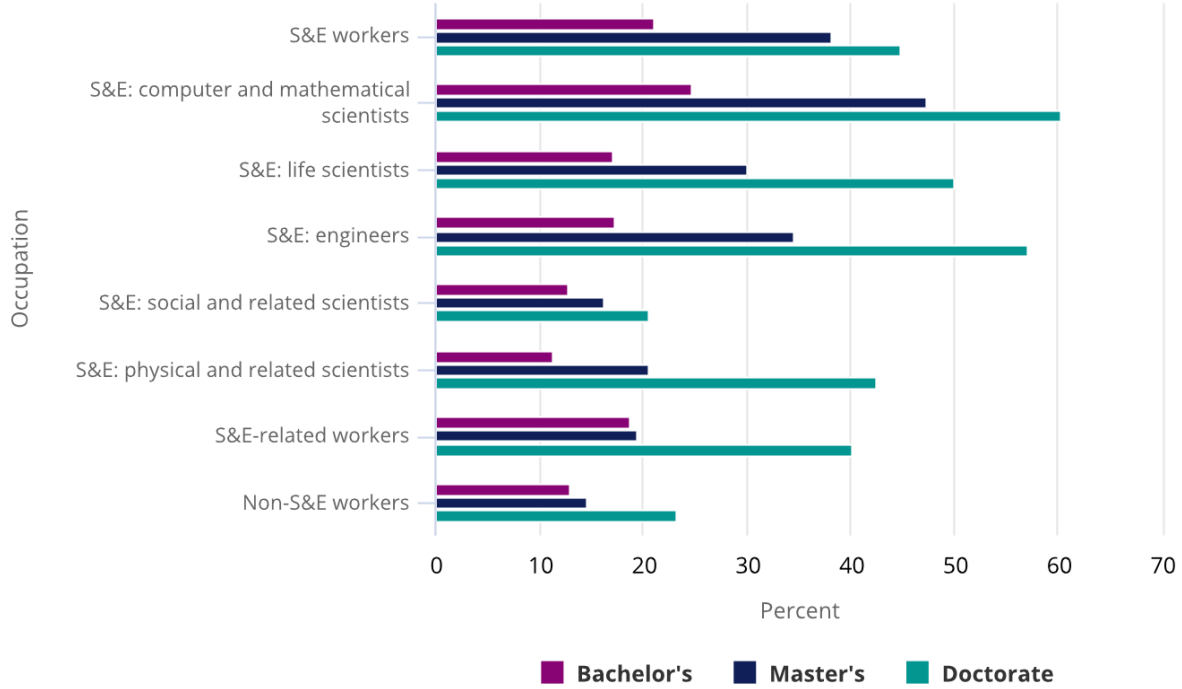
comprehensive multiparty data sharing agreement and a governance structure can be established. While each party typically has a data sharing agreement, best practices suggest that the parties identify the commonalities across each agreement as a starting point. From there, key topics that must be addressed in the data sharing agreement are the record linkage protocols, where the finalized data infrastructure will be housed and how it will be accessed, the protocols for requesting and being granted access, the access modalities available, responsibilities around data disclosure review and release of final data products, and data destruction. Finally, the data sharing agreement should establish a governing body that provides representation for all parties and a set of protocols that allow the governing body to navigate changes to the agreement as needed.

This report is the first step in developing a "playbook" for building a national data infrastructure with a use case focused on an important group of individuals, foreign-born scientists and engineers. Though this use case provides specific guidance for building a data infrastructure focused on this group, the recommendations and lessons learned, in many cases, can apply more broadly to any comprehensive data infrastructure focused on addressing some of our more intractable policy questions.

Introduction

Science and engineering fields are vital to U.S. economic growth, national defense, infrastructure, and overall public and private well-being. Indeed, the United States is recognized as a major leader in global science and engineering (S&E). At the same time, foreign-born individuals serve an important role in S&E fields. The nation has long benefited from the influx of foreign-born scientists and engineers (FBSEs) and the knowledge and skills they bring with them (Burke et al., 2022). Their contributions can be traced back to the 1930s when the political climate in Europe led to the first wave of European scientists and engineers to the United States (Libaers, 2007). As shown in Figure 1, these individuals are often highly trained and hold advanced degrees in their fields (National Science Foundation, 2022).

Figure 1. Foreign-born Individuals in S&E Occupations in the United States, by Level of Degree and Occupation: 2019 (National Science Foundation, 2022)



Source: Science and Engineering Indicator 2022: The State of U.S. Science and Engineering.

Given the importance of the S&E sector and the contributions of foreign-born individuals, it is imperative that we understand the training they receive, the investments in that training, their workforce outcomes, and the connections across these factors. Yet we do not currently have a robust data infrastructure that can support the evaluation of these connections as the data that currently exist are disparate and often not comparable, linkable, or comprehensive. There are a host of policies that would benefit from leveraging a more robust data infrastructure to better understand and support this sector and the foreign-born scientists and engineers that are vital to its success.

Perhaps the most salient policy question is the return on investment associated with FBSEs. In its most straightforward application, one might look at the educational investments made for a foreign-born student in science and engineering fields and then weigh them against a defined set of outcomes. However, the data infrastructure needed to assess this fundamental question is lacking and the policy context is more complex as there are other factors to consider when assessing potential returns to investments. A related and often widely cited policy concern is the possibility of job crowd-out by foreign-born STEM workers. There are some studies that suggest in certain cases immigrants may complement, rather than compete with American workers, because they have different skill sets and educational backgrounds (Wolla, 2014); other studies suggest immigrants may compete for jobs and depress wages (Ottaviano and Peri, 2012.) Another related area of policy interest is the tax incidence experienced by the S&E subset of immigrants compared to their native counterparts. Immigrants have been studied at length with some estimates of the returns. Government expenditure at the federal, state, and local levels on things such as public education is just one of many considerations. For example, immigrants coming to the U.S. as adults are typically net taxpayers (National Academies of Sciences, Engineering, and Medicine. 2017),[1] but with dependents, this may dampen any benefits due to investments in public education for their children. The issue is far more nuanced than considering something like public investments in the education of children of immigrants. One must also consider the potential benefits given FBSE are a growing portion of the STEM workforce and STEM workers are typically strong drivers of productivity in the United States (Abramitzky and Boustan 2017; Kerr and Kerr 2017; Khanna and Lee 2019),[2] and therefore it is important to consider this FBSE group separately from immigrants in other fields.

Existing literature on FBSEs uses a variety of public and private data sources. However, data sources are not always comparable, and in some cases, the quality of available data is less than ideal. One of the widely used public data sources is the National Science Foundation's Science and Engineers Statistical Data System (SESTAT),[3] which defines scientists and engineers as either those who received a college degree (bachelor's or higher) in a science, engineering, or related field, or those who work as a scientist or engineer or related occupation and have a bachelor's degree or higher in any field.

[1] About one in six workers in 2016 was born outside the United States and pay a significant share of the Old Age, Survivors, and Disability Insurance (OASDI) payroll taxes that fund Social Security. Restricting immigration would shrink the labor force, reduce the revenue of the OASDI trust funds, and weaken Social Security's long-term financial position, for example. If current legal immigration levels were cut by 50%, the Social Security fund would lose $1.5 trillion in revenue over the next 75 years.

[2] Immigrants also make an important contribution to the U.S. economy. Most directly, immigration increases potential economic output by increasing the size of the labor force. Immigrants also contribute to increasing productivity. Basso and Peri (2020) find that immigrants are more mobile than natives in response to local economic conditions, perhaps because they have fewer long-standing familial and community ties, helping labor markets to function more efficiently. Hunt and Gauthier-Loiselle (2010) have also shown that immigrants boost innovation, a key factor in generating improvements in living standards. Specifically, they find that a 1 percentage point increase in the population share of immigrant college graduates increases patents per capita by 9 percent to 18 percent (Rouse et al., 2021)

[3] SESTAT is a comprehensive and longitudinal integrated data system of information on the employment, educational, and demographic characteristics of scientists and engineers in the United States.

Data from SESTAT are available for public use and have been used in several studies on FBSEs. For example, Levin et al. (2004) used the Survey of Doctorate Recipients (SDR) from SESTAT to examine differential employment patterns of U.S. doctoral recipients in S&E over the period of 1973-1997 to gauge the extent non-citizen FBSEs may be displacing their citizen counterparts. Besides that, SESTAT's definition of S&E is also used in studies on FBSEs. Espenshade et al. (2001) researchers used SESTAT's classification of S&E occupations to analyze public-use microdata samples of the 1960 and 1990 decennial censuses along with the March 1997 Current Population Survey to explore and compare employment and earnings of FBSEs and their U.S.-born peers. However, little work has been done to assess the relationship of these outcomes to U.S.-based investments.

Given the mixed results in the literature and the importance of this subset of foreign-born individuals, as well as the lack of comparability across data sets in order to study this subpopulation, being able to provide the infrastructure to study and quantify various changes to the U.S. labor market associated with a changing demographic is imperative. An enhanced data infrastructure will allow for the estimation of expenditures and returns to U.S. investments in FBSE and is facilitated by linking data from the sources at all levels of government (federal, state, and local). Indeed, this is precisely what the data hub aims to facilitate. This work leverages test cases using data from states such as New Jersey, where the proportion of foreign-born individuals is large and the state more urban, and Arkansas and Kentucky where the population is smaller but more representative of foreign-born living in more rural environments. Most importantly, however, states in general have the ability to provide rich state and local information on foreign-born, education, and workforce indicators allowing for a robust analysis of investments and outcomes on the population of interest. This report will cover the current U.S. foreign-born landscape using federal survey data and compare that to data from state postsecondary records and will highlight the complexities of benchmarking and creating comparisons between state and federal data sources. In doing so, the report will document that the current infrastructure cannot adequately connect key information about foreign-born scientists and engineers–foreign-born status, education and field, educational investments, and workforce outcomes–to answer key policy questions. There is no existing national data model that streamlines and defines the connections between different data elements and datasets on FBSE. An FBSE infrastructure that includes a clearly defined data model, a robust record linkage approach, and a secure yet available method of access for research purposes can play a key role in bringing together education and workforce data systems, support a variety of use cases, and bolster future research on FBSEs.

Benchmarking Using Federal Survey Data

Three federal education surveys from the National Center for Education Statistics (NCES) were used to provide an estimate for the historical FBSE population counts in the United States. One from the National Center for Science and Engineering Statistics (NCSES) was used to provide an estimate for the historical FBSE doctoral degree earners in the United States. Microdata from the American Community Survey provided the last benchmark for the foreign-born population in the United States over the recent decade.

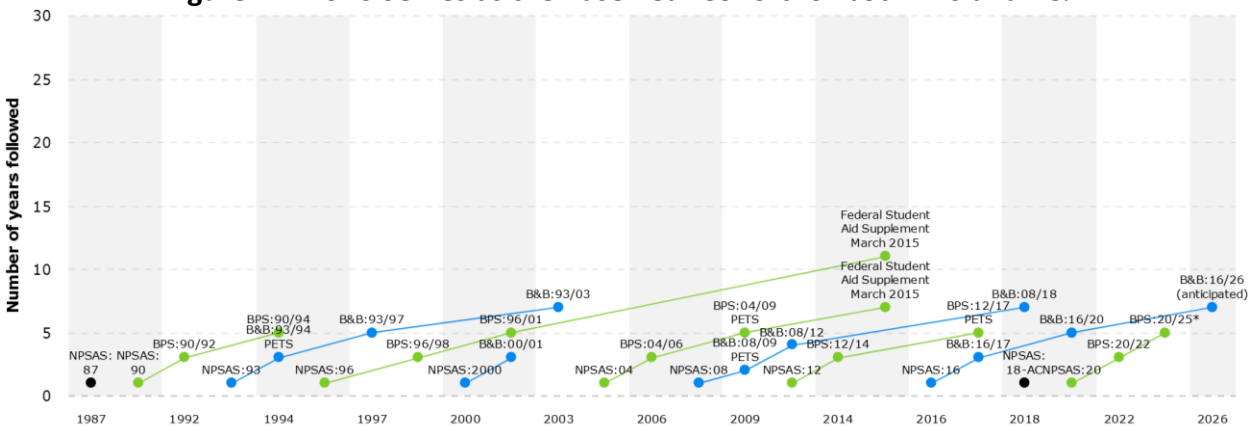
As information about the foreign-born and FBSEs from the individual surveys are presented it is important to note that not all of the surveys were developed with the intention of doing state level

analyses or to focus on this specific population of individuals. However, we present the results to demonstrate what these surveys can say about this population and to provide reference points for the state level administrative data to be presented later. The assessment of the information able to be conveyed from these surveys is an important step to not only identifying what can be said about FBSEs but also to identify gaps and potential opportunities to improve the data infrastructure.

National Center for Education Statistics (NCES)

The three NCES surveys that were assessed and benchmarked are the National Postsecondary Student Aid Study (NPSAS), Beginning Postsecondary Students Longitudinal Study (BPS), and Baccalaureate and Beyond Longitudinal Study (B&B). Among these surveys, the NPSAS serves as the foundation and base year for the BPS and the B&B. That is, the cohort samples of the BPS and B&B are drawn from the NPSAS cohort. Figure 2 demonstrates the cohort selection and relationship between these three surveys. For example, the cohort sample of the B&B:08/09 was drawn from the 2008 NPSAS (NPSAS:08) and B&B:08/12 is the follow-up of this cohort. Similarly, the cohort sample of the BPS:12/14 was drawn from the NPSAS:12 and BPS:12/17 PETS is the follow-up of this cohort. For this project, we use NPSAS data for the years 2008, 2012, 2016, and 2018; and the B&B and BPS samples drawn from the NPSAS:08, NPSAS:16, and NPSAS:12, respectively. We will discuss each of these surveys in more detail in the sections below.

Figure 2. NPSAS Serves as the Base Year Cohort for both BPS and B&B



Source: National Center for Education Statistics - National Postsecondary Student Aid Study (NPSAS)

To identify nativity and citizenship status in these surveys the student's immigrant status (*IMMIGRA*), parents' birthplace (*PARBORN*), and student's birthplace (*USBORN*) for the years 2008, 2013, and 2016 were used. Specifically, to identify naturalized citizens, *IMMIGRA* indicates that a student is foreign-born and *PARBORN* must also indicate that both parents are foreign-born. If *IMMIGRA* indicates that a student is a foreign student with a visa or a resident alien or eligible non-citizen, a student is identified as a non-citizen foreign-born. There are some years (NPSAS 2018, BPS 2014 and 2017) that do not provide information on *IMMIGRA, USBORN,* and *PARBORN*, in these cases we instead used students' citizenship status. Thus, for those years, we cannot identify whether a student is a U.S.-born citizen or a naturalized citizen. To identify a U.S.-born student for the investment table, the *USBORN* variable is used, which indicates whether a student was born in the U.S. or not.

To identify S&E fields, the field of study was used in accordance with the NCSES Survey of Earned Doctorates Field of Study Taxonomy. The following fields are considered S&E based on the National Science Foundation (NSF) definitions: agriculture and related sciences; natural resources and conservation; area, ethnic, and gender studies; computer and information sciences; engineering; engineering technologies/technicians; biological and biomedical sciences; mathematics and statistics; multi/interdisciplinary studies; physical sciences; psychology; health professions and related sciences; anthropology; criminology; economics; geography; international relations and affairs; political science and government; sociology; and social sciences - other. It is worth noting that the major (*MAJORS)* reported in NCES surveys indicates the major or field of study at the time of the survey, not that of the previous degree attained.

The institution state (*INSTSTAT*) was used to identify the state instead of the student's state of legal or permanent residence (*STUSTATE*) to make the statistics comparable with the administrative data from the states. This has the benefit of remaining stable over time for the student (i.e., once graduated from an institution, the institution state is fixed for all longitudinal analyses).

National Postsecondary Student Aid Study (NPSAS)

NPSAS provides information about how students and their families pay for postsecondary education. In this section we will provide degree level, field, and foreign-born estimates and in a following section provide investment estimates. NPSAS surveys are conducted every 4 years, except for the 2018 NPSAS data, which was collected directly from state administrative data. This survey also serves as the base year for two longitudinal surveys that follow: BPS and B&B. While BPS and B&B are longitudinal studies, NPSAS is a cross-sectional study, which means the cohort in each survey year is independent from the cohorts in other years. The distribution of the foreign-born postsecondary population is documented by the level of degree attainment. The survey weight of *WTA000* was used throughout the computation of all reported statistics.[4]

The highest level of educational attainment was categorized into 4 groups: bachelor's only, master's only, doctoral only, and at least a bachelor's degree. Post-baccalaureate certificate earners are counted as bachelor's degree earners. Post-master's certificate earners are counted as master's degree earners. *DEGPRBA* (earned a bachelor's degree since high school) and *DEGPRPTB* (earned a post-baccalaureate certificate since high school) variables were used to identify those who attained a bachelor's degree. Similarly, *DEGPRMS* (earned a master's degree since high school) and *DEGPRPTM* (earned a post-master certificate since high school) variables were used to identify those who obtained a master's degree. Lastly, *DEGPRFP* (earned a professional degree since high school) and *DEGPRDOC* (earned a doctoral degree since high school) variables were used to identify those with a doctoral degree. NPSAS data report the degree level of previously attained degrees (if any) and of the program that a student was enrolling in at the time of the survey. However, there is no indication of whether a student completes

[4] It is important to note that the weighting in the NPSAS:08, NPSAS:12, and NPSAS:16 surveys was not designed to provide state-level estimates for the states included and was not designed to provide estimates for graduate level students. The weighting for the NPSAS:18 allows for state level analysis of Arkansas, Kentucky, and New Jersey and the analysis of undergraduate and graduate students. https://nces.ed.gov/surveys/npsas/state_oversamples.asp

the current program. Thus, to get the count of the students for each degree level, the information associated with the previously attained degrees was used.

Table 1 shows, by state, the percentage of the foreign-born population among all students, the percentage of FBSE among all students, the percentage of foreign-born among S&E majors, and the percentage of S&E majors among the foreign-born population for each degree level. It highlights that although the foreign-born population in the S&E fields is relatively small, a large portion of foreign-born students pursue S&E degrees. In general, New Jersey has the largest foreign-born and FBSE populations; however, in 2018, these percentages are smaller than those of Kentucky. It is important to highlight that NPSAS 2018 data came from state administrative data, not from a survey. That highlights an important question about the comparison between administrative and survey data. Furthermore, it is important to note the limitations with the data when looking at the individual states in 2008, 2012, and 2016, particularly in the advanced degree fields. This will be discussed in further detail in the *Comparison Between Federal and State Data* section below.

Table 1. Distributions of Foreign-born Students using NPSAS Data

		All Degrees				Bachelor				Masters				Doctoral			
State	Year	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB
Arkansas	2008	5	2	4	31	3	1	2	27	0	4	14	-	0	0	0	-
	2012	11	6	21	52	11	4	20	40	10	10	22	94	100	100	100	100
	2016	14	7	24	47	16	9	26	59	13	1	7	5	0	0	-	-
	2018	18	16	22	90	18	16	22	90	23	21	30	90	0	0	0	-
Kentucky	2008	9	6	25	70	9	6	23	70	10	7	38	71	0	0	-	-
	2012	13	7	24	53	11	5	19	47	16	13	35	86	69	16	34	23
	2016	23	14	29	62	20	15	27	74	35	15	50	42	8	4	4	45
	2018	41	35	61	86	41	35	61	86	47	42	74	89	14	4	7	33
New Jersey	2008	25	12	41	49	25	11	39	43	24	17	47	69	0	0	0	-
	2012	28	15	32	55	25	12	27	47	34	24	42	68	70	70	70	100
	2016	33	12	39	36	33	11	35	33	29	21	53	71	100	0	-	0
	2018	21	10	37	46	21	10	37	46	18	7	27	37	39	27	76	68

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero. % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

Baccalaureate and Beyond (B&B)

The B&B survey is used to study bachelor's degree holders in the United States and can be used to determine foreign-born and STEM status. The B&B samples are drawn from the NPSAS. Although various iterations of the survey have been conducted, this work focuses primarily on the B&B:08 cohort in order to longitudinally track the cohort from 2010-2019. Specifically, the survey's target population is postsecondary students in the 50 states, the District of Columbia, and Puerto Rico who have completed a bachelor's degree in the academic year 2007-2008 (July 1, 2007 – June 30, 2008). The B&B includes three follow-up surveys 1-, 4-, and 10-years post-graduation. For more information on the B&B survey design, NCES maintains technical documentation on their website.

The B&B:08/18 is used to estimate the bachelor's, master's, and doctoral populations. Specifically, estimates are generated for those receiving a bachelor's degree in 2008, those from the cohort (bachelor's degree in 2008) who receive a master's degree by 2012 or 2018, and those from the cohort who receive a doctoral degree by 2012 or 2018. Each student is assigned naturalized or non-citizen

status. Naturalized is when both student's parents were born outside of the United States and they themselves are foreign-born but citizens. A non-citizen is assigned when an individual is a foreign student with a visa or a resident alien/eligible non-citizen. Foreign-born is then defined using naturalized or noncitizen status and includes those who are either naturalized citizens or non-citizens.

To define S&E, *MAJORS23* is used to identify the S&E bachelor's degree recipients in 2008. *B2HICMAJ* and *B3HICMAJOR* are used to identify the S&E degree recipients for the highest attained degree (master's or doctoral) in 2012 [*B2HIDEG*] and 2018 [*B3HIDEG*] respectively. For the bachelor's estimates in either 2008 or 2016, the weight *WTA000* is used. For estimates for 2012 and 2018, the weights *WTD000* and *WTG000* are used. They are the weights assigned to the broadest number of respondents, e.g., those who participated in NPSAS 2008 and 2012 or NPSAS 2008 and 2018.[5] Since we only use the 2008 respondents to group individuals and treat those as fixed individual elements, we use the cross-sectional weights. Percentages are then calculated and reported in Table 2 below.

Table 2 shows that for Arkansas and Kentucky, almost none of the students in the bachelor's cohorts of 2008 and 2016 attained higher degree levels in 2012 and 2018, respectively. On the other hand, in New Jersey, the bachelor's cohorts were more likely to pursue a master's degree than in other states. The data shows a 4-year follow-up while the typical length of completing a doctoral degree is 6 years, which contributes to the low percentages of zeros in the doctoral category until 2018. In addition, comparing the percentages for bachelor's students in the B&B samples with the NPSAS samples, a very small number of foreign-born students from the NPSAS sample were selected for the B&B sample. This is an important limitation when considering this use of the B&B survey for understanding the outcomes for FBSEs.

Table 2. Distributions of Foreign-born Students using B&B Data

		Bachelor				Masters				Doctoral			
State	Year	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB
Arkansas	2008	6	0	0	0								
	2012					0	0	0	-	0	0	0	-
	2016	0	0	1	100								
	2018					2	0	0	0	0	0	0	100
Kentucky	2008	1	0	1	19								
	2012					0	0	0	-	0	0	0	-
	2016	1	1	3	75								
	2018					0	0	0	-	0	0	0	-
New Jersey	2008	19	5	27	27								
	2012					22	7	21	33	0	0	0	100
	2016	11	4	11	37								
	2018					24	10	43	40	1	1	1	100

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero. % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

[5] Because the sample for B&B:08 is drawn from the NPSAS:08 sample it is important to note it has the same limitations as the NPSAS:08 estimates in that it is not representative of the three states included.

Beginning Postsecondary Students (BPS)

The BPS survey includes students enrolled at postsecondary institutions who were surveyed at the end of their first year, and then three and six years after first starting in postsecondary education. The sample includes traditional and nontraditional students.

To define S&E fields, the *MAJORS* variable is used, which indicates the student's undergraduate major or field of study in 2011-2012, according to NPSAS records. For 2014 and 2017, for bachelor's recipients, *MAJBA14* and *MAJBA17* variables are used, indicating the field of study the respondent was pursuing when last enrolled in bachelor's as of June 2014 and as of June 2017, respectively. For master's, and doctoral categories, *MAJ14* and *MAJ17* variables are used which indicate the respondent's field of study when last enrolled in any degree. This is the most relevant variable in the absence of a variable that specifically indicates a master's or doctoral field of study, as is indicated with the bachelor's degree.

To define the degree holders, in 2012 the *HIGHLVEX* variable is used which indicates the highest level of education ever expected when interviewed in 2011-2012. This is the most relevant proxy variable to define the level of degree that the person holds. For 2014 and 2017, to define bachelor's degree holders, *ATHTY3Y* and *ATHTY6Y* variables are used which indicate the highest degree attained anywhere though June 2014 or through June 2017. This variable does not include a degree level higher than a bachelor's degree. To define master's, and doctoral degree holders, *DGEVR14* and *DGEVR17* variables are used which indicate the highest level of education the respondent expected to complete when interviewed in 2013-2014 and in 2016-2017. This is the most relevant variable available to define the degree holders above a bachelor's degree. To define bachelor's only, master's only, and doctoral only degrees, the condition of not having a higher level is applied. To note, we found some incidents in the data where students did not indicate that a degree was expected but still attained a degree. Table 3 reflects this issue. For example, in Arkansas in 2017, although the percentage of foreign-born for bachelor's and masters are non-zero, the percentage of foreign-born for all degrees is 0. This raises concern about self-reported information in survey data.

Note, when using *ATHTY3Y* of attained bachelor's degrees in 2014 the number is much lower than when using *DGEVR14* expected degree (e.g., 712 weighted individuals versus 13,431 weighted individuals in Arkansas), because BPS includes beginning postsecondary students who started their studies in 2012 and many have not yet completed the bachelor's degree. When using a later variable of *ATHTY6Y* which indicates the recipients of bachelor's degrees in 2017, the number is much higher, as there have been 4-5 years since the beginning of postsecondary studies in 2012 (e.g., 13,208 weighted individuals using ATHTY6Y variable in Arkansas).

For 2012, the *WTA000* weight variable is used which is a cross-sectional weight, created for all BPS:12/17 interview respondents regardless of their NPSAS:12 or BPS:12/14 response status, to make it consistent with the NPSAS and B&B surveys. NPSAS is a base year survey that uses the cross-sectional weight, and in the B&B the weights are used which do not require a response in all the follow-up years of the survey.[6]

[6] Because the sample for BPS is drawn from the NPSAS survey sample it is important to note it has the same limitations as the NPSAS estimates in that it is not representative of the three states included.

Similar to B&B, BPS is also a longitudinal survey but instead of collecting data from bachelor's graduates, BPS collects the data from the respondents in their first year of enrollment at a post-secondary institution, which is 2012 for this project. The statistics reported in 2012 indicate the highest degree level respondents are ever expected to complete. The statistics in the follow-up years of 2014 and 2017 indicate the actual degree level attained for the bachelor's degrees, while master's and doctoral degrees are defined based on the highest degree expected when interviewed in 2013-2014 and in 2016-2017. For example, in New Jersey in 2012, 14% of students who expected to complete a postsecondary degree in the future were foreign-born. Among those who expected to only achieve a bachelor's degree, 8% were foreign-born, 15% of those aiming for a master's were foreign-born, and 24% of the students expecting a doctoral degree were foreign-born. Data from BPS is also consistent with other NCES surveys that foreign-born students are more likely to pursue a degree in S&E fields.

Table 3. Distributions of Foreign-born Students using BPS Data

		All Degrees				Bachelor				Masters				Doctoral			
State	Year	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB
Arkansas	2012	6	3	5	51	9	4	7	43	7	4	7	63	0	0	0	-
	2014	0	0	-	-	0	0	0	-	1	1	2	100	0	0	0	-
	2017	0	0	0	-	9	0	0	0	9	0	0	0	0	0	0	-
Kentucky	2012	6	3	6	45	2	0	0	4	10	3	8	26	9	9	16	100
	2014	3	3	100	100	3	3	100	100	4	4	9	100	10	10	14	100
	2017	5	3	8	73	0	0	0	100	7	4	11	58	10	5	10	49
New Jersey	2012	14	8	20	54	8	2	9	28	15	13	28	86	24	8	15	32
	2014	19	19	92	100	6	0	0	0	21	10	23	50	16	14	19	89
	2017	11	7	20	67	11	9	22	85	18	11	26	59	5	5	9	95

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero. % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

National Center for Science and Engineering Statistics (NCSES)

Estimates of the foreign-born doctoral population are created using the NCSES Survey of Earned Doctorates (SED). In later sections, descriptions of graduate and undergraduate debt, as investments into doctoral education using this same survey, are provided, as well as the descriptions of the source of support and the postgraduation location.

To approximate the academic year, the *PHDFY* variable is used to subset the data to the years 2010 through 2017. To get the state of the institution, IPEDS location data is used based on the *PHDINST* variable (doctoral institution). Science and engineering (S&E) fields are defined based on the *PHDFIELD* variable (doctoral field). S&E fields include Agriculture (Life Sciences), Biological/Biomedical Sciences (Life Sciences), Health Sciences (Life Sciences), Engineering, Computer and Information Sciences, Mathematics, Physical Sciences, Psychology, Social Sciences, and exclude Humanities, Education, Business Management/Administration, Communication, Fields Not Elsewhere Classified.

Table 4. Distributions of Doctoral Foreign-born Students using SED Data

State	Year	% FB	% FBSE	% FB in SE	% SE in FB	% Non-citizen FBSE
	2011	-	33	41	-	-
	2012	-	36	48	-	-
	2013	-	32	41	-	-
Arkansas	2014	-	23	30	-	-
	2015	-	23	30	-	-
	2016	-	34	44	-	-
	2017	-	30	39	-	-
	2011	35	28	45	82	89
	2012	35	30	44	84	-
	2013	33	28	42	84	93
Kentucky	2014	34	29	43	86	95
	2015	36	30	46	84	94
	2016	35	31	45	87	93
	2017	33	28	44	84	-
	2011	52	42	56	80	94
	2012	53	42	58	81	83
	2013	51	42	56	82	85
New Jersey	2014	49	39	54	79	83
	2015	50	41	54	83	85
	2016	48	39	53	82	88
	2017	50	40	54	80	88

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero. In the case of Arkansas, data were redacted to avoid complementary disclosure. % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

Naturalized citizens are defined with *CITIZ* variable (type of citizenship) with the category called "U.S., naturalized". Non-citizens are defined using the same *CITIZ* variable with the following categories: "Non-U.S., immigrant (permanent resident)", "Non-U.S., non-immigrant (temporary resident)", "Non-U.S., visa status unknown". The foreign-born population is defined as the sum of naturalized and non-citizens. U.S.-born is defined using the *CITIZ* variable with the category "U.S., native-born". Missing category is included where the *CITIZ* value is missing.

To identify the source of support, *SRCEPRIM* variable is used which indicates the primary source of support. *PDUSFOR* variable is used which indicates "postgraduation location: U.S. or foreign". Debt variables are used to indicate the level of undergraduate debt (*UDEBTLVL*) and the level of graduate debt (*GDEBTLVL*). For these variables, the missing category is included where the corresponding values are missing.

Estimates in Table 4 indicate that about a third of doctoral recipients from Kentucky institutions are foreign-born while about half of doctoral recipients from New Jersey institutions are foreign-born. In both states, of those that are foreign-born, more than 80% are non-citizens. In Arkansas, although most degrees are S&E degrees (more than 75%), foreign-born receiving S&E degrees make up about a third of total doctoral recipients. In other words, doctoral degrees in S&E awarded to foreign-born make up less than half of the total S&E degrees awarded.

Census Data

The American Community Survey (ACS) microdata consists of individual records with information about the characteristics of each person and housing unit in the survey. The ACS Public Use Microdata Sample (PUMS) includes a subsample of the ACS microdata, devoid of personalized information. The PUMS represents about two-thirds of the responses collected in the ACS in a specific 1-year or 5-year period. PUMS files for an individual year contain data on approximately one percent of the United States population. The ACS PUMS is a weighted sample, and weighting variables must be used to generate estimates and standard errors that represent the population. *PWGTP* variable was used for the person's weight for generating statistics on individuals.

Similar to the NCES and NCSES surveys, respondents are classified according to their citizenship status, highest degree (or the highest level of school completed), and field of study. The data do not indicate whether the degree was obtained in the United States. The data also do not indicate the state in which the respondent received their degree. S&E fields are based on the field of degree variable where respondents were asked to list the specific major(s) of any bachelor's degree received. This variable does not include the field of any other type of degree earned (such as a master's or doctorate).

Although the population covered in the ACS data is different from other data sources presented earlier, the statistics derived from ACS data (Table 5) are consistent with those derived from those sources in that New Jersey has the largest foreign-born population, a large portion of the foreign-born population majored in S&E across the three states, and the higher the degree level is, the more likely that they were in S&E fields. We provide a more detailed trend analysis below in the comparison of Federal and State Data section.

It is important to note that the review of federal data presented here is not a comprehensive review. There are other sources available such as the National Survey on College Graduates (NSCG) that provide additional information on FBSEs. The NSCG is a particularly useful survey as it provides information on FBSE degree attainment and workforce outcomes. However, there are similar issues with state and subgroup representation.

While some general patterns and trends emerge from the various surveys discussed, taken together, these surveys demonstrate several key limitations to understanding the educational outcomes of foreign-born scientists and engineers. First, many of these surveys were not designed to be representative of such a specific group of individuals, FBSEs. Compounding this issue is that, for many states, the surveys are not designed to be representative at the state level. This poses challenges to better understanding the local and regional context that FBSEs may face which are important to designing more effective programs and policies. In some cases, the surveys use slightly different definitions for the variables of interest, such as foreign-born vs. citizenship, which can make comparisons across surveys difficult. In cases where there are fewer concerns about representation, such as the PUMS, the results are not able to determine the country in which the respondent earned their credential and for advanced degrees cannot account for the field of study. Perhaps one of the more significant limitations with surveys is the ability to link with other sources of data to enhance the ability to comprehensively understand the intersection between investments and outcomes for this important group of individuals. With these issues in mind the report now turns to the assessment of state level administrative data in the states of Arkansas, Kentucky, and New Jersey.

Table 5. Distributions of Foreign-born Students using ACS Data

State	Year	All Degrees				Bachelor				Master				Doctoral			
		% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB
Arkansas	2011	4	3	8	72	4	2	6	60	5	4	12	87	7	7	11	92
	2012	5	3	8	65	4	2	6	57	7	5	14	73	9	6	11	68
	2013	5	3	8	60	4	2	6	49	7	5	14	74	5	4	7	72
	2014	6	3	8	53	5	2	6	50	7	3	11	47	11	8	13	71
	2015	6	3	8	58	4	2	4	39	7	5	14	69	12	11	17	85
	2016	5	4	9	68	4	2	6	59	6	5	14	77	11	8	13	76
	2017	6	3	8	54	5	3	7	49	7	5	13	63	7	4	7	59
	2018	6	3	8	59	5	3	6	54	6	4	11	68	10	6	10	58
	2019	6	4	9	62	4	3	6	58	8	5	12	59	9	8	13	80
Kentucky	2011	6	4	9	63	4	2	6	56	6	4	11	64	11	9	16	77
	2012	5	3	7	58	4	2	5	50	4	3	8	70	10	7	13	65
	2013	6	3	8	60	5	3	7	59	6	4	11	65	9	5	10	55
	2014	6	3	8	54	5	2	5	49	6	3	10	50	9	7	13	78
	2015	6	4	9	61	5	3	6	52	7	4	12	63	10	8	15	80
	2016	5	3	7	57	5	2	6	50	5	3	9	65	10	7	12	67
	2017	6	4	9	59	5	3	7	55	6	3	10	60	11	7	14	67
	2018	6	3	7	54	5	2	6	47	5	3	8	58	10	7	12	72
	2019	7	4	10	55	7	3	8	52	6	3	10	53	12	8	16	67
New Jersey	2011	27	16	35	57	26	13	33	51	30	19	40	63	31	24	36	78
	2012	27	15	34	57	25	13	32	52	29	19	38	63	31	22	35	72
	2013	28	16	35	57	26	13	32	51	31	20	40	64	32	23	36	72
	2014	27	16	34	57	26	14	33	51	29	18	38	63	29	22	33	74
	2015	28	16	35	57	26	13	32	51	31	19	40	62	30	22	32	73
	2016	28	16	35	58	25	13	32	51	32	21	42	64	32	24	35	74
	2017	29	17	36	58	27	14	33	51	32	21	41	65	31	23	35	74
	2018	28	16	35	58	25	13	32	51	31	20	40	65	35	25	37	73
	2019	29	16	35	57	27	13	31	51	32	20	41	63	34	26	38	76

Note: % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

Benchmarking Using State Data

The states of New Jersey, Arkansas, and Kentucky provided estimates of foreign-born and FBSE populations using post-secondary administrative records. While state postsecondary administrative data provides detailed information on enrollment as well as completion, this report focuses on completion to be comparable with other data sources. The predominant source of administrative data for higher education completion is the records collected for the Integrated Postsecondary Education Data System (IPEDS) survey conducted annually by the National Center for Education Statistics (NCES), a part of the Institute for Education Sciences (IES) within the United States Department of Education. The IPEDS survey is required for all institutions that participate in any federal financial assistance program authorized by Title IV of the Higher Education Act of 1965. These data are typically available for all public postsecondary institutions and some private institutions. Among the three states, Kentucky is the only state that collects information on students' country of origin. Thus, they were able to identify whether a foreign-born student is a naturalized citizen or not. In Arkansas and New Jersey, due to the lack of information on country of origin, foreign-born students were identified as those who were not U.S. citizens at the time of graduation. The statistics reported in the state data indicate the number of students who graduated each year with a specific degree level instead of the number of students who were enrolled in a degree program like some of the survey data.

Table 6 presents the overall proportions of foreign-born students. Because these measures are derived from the population level data, we also present the total counts of foreign-born students in Table 7. Over the eleven-year panel, the proportion of foreign-born graduates as a percent of all graduates increases in almost all groups in all states except for doctoral recipients in Kentucky and New Jersey (Table 6). However, the total number of doctoral graduates increases somewhat in Kentucky (Table 7). In all three states, foreign-born bachelor's degree graduates in S&E fields make up about the same yearly proportion as foreign-born graduates across all fields. Foreign-born advanced degree graduates are more likely to be in S&E fields, particularly those receiving master's degrees, than in other fields, and in many cases their proportion of all graduates in S&E is double that of all degree fields (Table 6). Similarly, the proportion of foreign-born graduates in S&E fields of all foreign-born graduates is significant and, in most cases, outpaces the proportion of U.S.-born graduates in S&E fields of all U.S.-born graduates (Table 7).

The panel of population data allows for the analysis of how the cohorts of foreign-born and FBSEs have changed over the eleven years examined. Figure 3 illustrates the proportion of foreign-born and FBSE populations from all students who attained a bachelor's, master's, or doctoral degree. The trend of FBSE is similar to the trend of foreign-born in all three states, suggesting a consistent proportional choice of majors for foreign-born students year over year. Across the three states the proportion of foreign-born and FBSE is increasing. Moreover, this figure also shows a spike in the proportion of foreign-born and FBSE in Arkansas and Kentucky in 2017 and 2020, respectively.

Table 6. Distributions of Foreign-born Students using State Administrative Data

State	Year	All Degrees				Bachelor				Master				Doctoral			
		% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB	% FB	% FBSE	% FB in SE	% SE in FB
Arkansas	2011	5	2	5	46	4	1	3	40	7	4	14	49	-	-	11	91
	2012	5	2	6	47	4	2	4	38	8	4	16	52	10	9	14	85
	2013	6	3	7	49	4	2	4	43	8	4	16	50	9	8	14	88
	2014	5	2	6	44	4	1	3	35	9	4	15	48	10	8	13	83
	2015	6	3	6	45	4	2	4	39	9	4	16	46	10	8	13	86
	2016	6	4	8	58	4	2	4	42	11	8	26	71	-	-	13	91
	2017	9	6	14	69	5	2	4	36	19	16	45	83	8	6	9	81
	2018	8	5	12	67	5	2	5	41	16	13	39	81	13	11	15	81
	2019	6	4	9	57	5	3	5	49	10	6	23	63	11	9	13	76
	2020	6	4	8	57	5	3	6	54	8	5	19	62	12	10	14	79
	2021	6	3	8	57	5	3	5	51	8	4	17	58	12	10	14	85
Kentucky	2011	4	2	6	57	2	1	2	42	8	5	16	60	9	8	13	85
	2012	4	2	6	56	2	1	2	43	8	4	15	56	9	8	13	88
	2013	4	2	6	58	3	1	3	49	7	4	15	57	10	9	13	88
	2014	5	3	6	56	3	1	3	52	7	4	12	49	11	10	15	90
	2015	5	3	7	55	3	2	4	49	8	4	14	52	10	9	14	86
	2016	5	3	7	60	4	2	4	52	9	6	17	61	10	9	13	86
	2017	6	3	7	60	4	2	4	49	9	6	19	65	9	8	12	89
	2018	7	5	11	69	4	2	4	51	15	12	32	78	9	7	11	87
	2019	10	8	16	79	4	2	4	52	23	20	46	88	10	8	12	81
	2020	16	13	26	85	4	2	4	54	39	35	63	90	9	8	11	85
	2021	12	8	16	67	3	2	4	53	28	19	43	69	8	7	10	79
New Jersey	2011	9	6	13	62	5	3	6	58	20	12	33	61	17	14	21	83
	2012	9	6	12	63	5	3	6	57	19	12	32	64	17	14	21	85
	2013	9	5	11	63	5	3	5	55	18	12	30	66	17	14	21	81
	2014	9	6	12	65	5	3	5	56	19	13	33	69	16	13	19	77
	2015	11	7	14	68	6	3	6	57	24	18	40	73	17	14	20	84
	2016	12	8	16	70	6	4	7	60	26	20	45	75	17	13	19	79
	2017	12	8	16	69	7	4	7	58	25	19	43	76	17	14	20	84
	2018	12	8	16	68	7	4	8	60	25	19	44	74	16	13	18	78
	2019	10	7	13	66	7	4	7	57	19	14	35	73	17	15	19	84
	2020	12	8	14	65	8	4	8	55	22	16	38	73	21	17	23	83
	2021	12	8	15	68	8	4	8	59	25	18	41	74	15	12	17	80

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero. % FB = Total foreign-born population/Total population; % FBSE = Total foreign-born in S&E/Total population; % FB in SE = Total foreign-born in S&E/Total S&E population; % SE in FB = Total foreign-born in S&E/Total foreign-born population.

Table 7. Population of Foreign-born Students using State Administrative Data

State	Year	All Degrees				Bachelor				Master				Doctoral			
		FB	FBSE	Pop.	SE Pop.	FB	FBSE	Pop.	SE Pop.	FB	FBSE	Pop.	SE Pop.	FB	FBSE	Pop.	SE Pop.
Arkansas	2011	758	352	16,733	6,596	422	169	12,028	5,172	298	146	4,107	1,069	-	-	678	365
	2012	1,063	499	20,029	7,973	566	217	13,909	6,101	419	216	5,301	1,357	92	78	911	538
	2013	1,141	559	20,327	8,200	625	269	14,156	6,257	441	220	5,290	1,370	93	82	981	600
	2014	1,134	497	21,135	8,923	625	219	15,254	6,966	427	206	5,003	1,363	96	80	980	626
	2015	1,247	559	21,728	9,203	693	270	15,674	7,214	476	218	5,179	1,396	93	80	974	623
	2016	1,378	805	22,364	9,738	678	285	15,855	7,347	616	437	5,593	1,708	-	-	1,041	705
	2017	2,020	1,402	23,521	10,263	866	310	16,065	7,322	1,251	1,042	6,571	2,303	77	62	1,018	666
	2018	1,926	1,287	23,716	10,371	878	364	16,371	7,619	1,024	831	6,542	2,131	135	109	1,035	704
	2019	1,564	890	24,069	10,224	874	426	16,905	7,965	613	389	6,368	1,659	121	92	1,061	700
	2020	1,512	861	23,998	10,299	826	442	16,906	7,965	524	327	6,271	1,682	136	108	1,102	772
	2021	1,483	846	24,626	10,696	874	446	17,228	8,217	493	287	6,431	1,676	163	138	1,345	987
Kentucky	2011	1,132	648	29,553	11,646	375	157	20,208	8,336	602	360	7,669	2,283	155	131	1,676	1,027
	2012	1,316	731	31,166	12,381	487	210	20,709	8,781	651	364	8,562	2,406	178	157	1,895	1,194
	2013	1,367	790	31,607	12,848	544	268	21,058	9,167	652	371	8,786	2,546	171	151	1,763	1,135
	2014	1,440	806	31,695	13,416	617	320	21,557	9,623	620	303	8,281	2,574	203	183	1,857	1,219
	2015	1,632	893	32,036	13,639	743	365	21,805	9,797	701	367	8,433	2,688	188	161	1,798	1,154
	2016	1,797	1,077	32,992	14,430	818	428	22,531	10,235	786	483	8,546	2,920	193	166	1,915	1,275
	2017	1,883	1,124	33,626	15,037	883	435	22,995	10,808	822	531	8,684	2,865	178	158	1,947	1,364
	2018	2,445	1,685	35,003	15,913	859	436	23,415	11,036	1,406	1,092	9,482	3,445	180	157	2,106	1,432
	2019	3,654	2,871	36,689	17,627	902	465	23,475	11,346	2,536	2,231	11,001	4,848	216	175	2,213	1,433
	2020	6,232	5,318	39,399	20,615	832	449	23,696	11,630	5,195	4,695	13,434	7,398	205	174	2,269	1,587
	2021	4,698	3,149	39,027	19,121	816	434	23,412	11,508	3,680	2,556	13,193	6,003	202	159	2,422	1,610
New Jersey	2011	4,482	2,765	47,334	21,786	1,760	1,013	33,270	15,935	2,328	1,426	11,711	4,281	394	326	2,353	1,570
	2012	4,537	2,853	50,636	23,799	1,726	977	35,582	17,438	2,403	1,531	12,646	4,741	408	345	2,408	1,620
	2013	4,329	2,730	50,856	24,272	1,724	952	36,220	17,901	2,195	1,447	12,226	4,777	410	331	2,410	1,594
	2014	4,576	2,961	51,934	25,609	1,752	984	36,898	18,948	2,419	1,664	12,566	5,037	405	313	2,470	1,624
	2015	5,748	3,899	53,130	26,992	2,118	1,201	37,520	19,444	3,235	2,366	13,284	5,926	395	332	2,326	1,622
	2016	6,391	4,476	54,299	27,817	2,285	1,367	37,901	19,891	3,700	2,788	13,991	6,223	406	321	2,407	1,703
	2017	6,423	4,443	55,382	28,028	2,587	1,500	38,996	20,295	3,399	2,578	13,816	5,938	437	365	2,570	1,795
	2018	6,546	4,480	55,202	28,479	2,709	1,629	39,064	21,036	3,447	2,546	13,707	5,732	390	305	2,431	1,711
	2019	5,289	3,480	50,967	26,239	2,556	1,445	36,648	19,531	2,308	1,676	11,868	4,852	425	359	2,451	1,856
	2020	7,194	4,704	61,155	32,445	3,315	1,819	43,566	23,945	3,240	2,353	14,540	6,231	639	532	3,049	2,269
	2021	5,675	3,834	47,567	25,488	2,565	1,524	34,055	18,852	2,750	2,023	11,150	4,919	360	287	2,362	1,717

Note: a hyphen (-) means the statistic is either redacted due to disclosure guidelines or the population of interest in the denominator is zero.

Looking more closely at the number of foreign-born graduates, Figure 4 demonstrates that the spikes in both Arkansas and Kentucky were driven by those who graduated with a master's degree. The overall number of graduates with a master's degree is similar to those with a bachelor's degree in Arkansas except for the spike in 2017. The number of graduates with a bachelor's degree is similar to those with a master's degree in Kentucky until 2018 at which point there is significant growth in the number of graduates with master's degree though that number begins to trend downward in 2021. In discussions with the state, it is likely that this number was influenced, in part, by the response of international students to the pandemic, with many expediting their degree completion before returning to their home country. The number of bachelor's graduates increases over the years for all three states but remains stable for Kentucky post 2016. However, New Jersey bachelor's degree graduates experienced rapid growth since 2014, with a slight drop in 2019, then hit its peak in 2020 and back to its normal level. The number of graduates with a doctoral degree remains relatively stable throughout the panel across the three states with New Jersey graduating more than Arkansas and Kentucky.

Looking more closely at science and engineering fields, Figure 5 demonstrates that foreign-born master's degree holders make up the highest proportion of foreign-born students. It could be that a master's degree in S&E attracts more foreign-born students, or a U.S.-born student is less likely to pursue a master's degree in S&E. Though among the foreign-born population, the higher the degree level that a student pursued, the more likely it was that they chose an S&E major. Figure 6 shows that while the proportion of foreign-born master's level S&E graduates has steadily increased, the proportion of doctoral level foreign-born is consistently higher across the years studied. This raises the question of whether it is due to the abundant funding for S&E at the doctoral level.

The postsecondary administrative data presents an opportunity to provide a state-level census of foreign-born graduates. Given that the federal government, as well as state governments, have invested significant resources into state longitudinal data systems it is important to evaluate how they can be used to answer questions about such an important group of individuals. At the same time, significant effort would need to be made to integrate information across states to provide a national picture. Here, the ACS data may be able to provide similar information, though other issues such as linkage to other data systems to attend to information gaps may be difficult to address. Those issues aside, the next section focuses on the comparison between federal and state level data.

Figure 3. Distribution of Foreign-born Graduates across Arkansas, Kentucky, and New Jersey

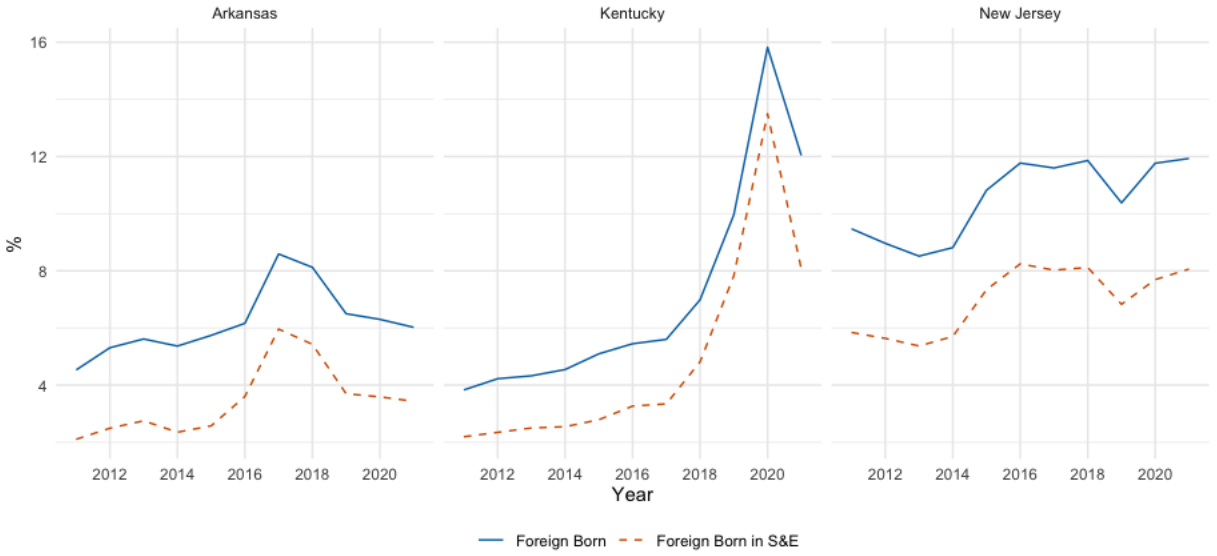


Figure 4. Trends of Foreign-born Graduates by Degree Level

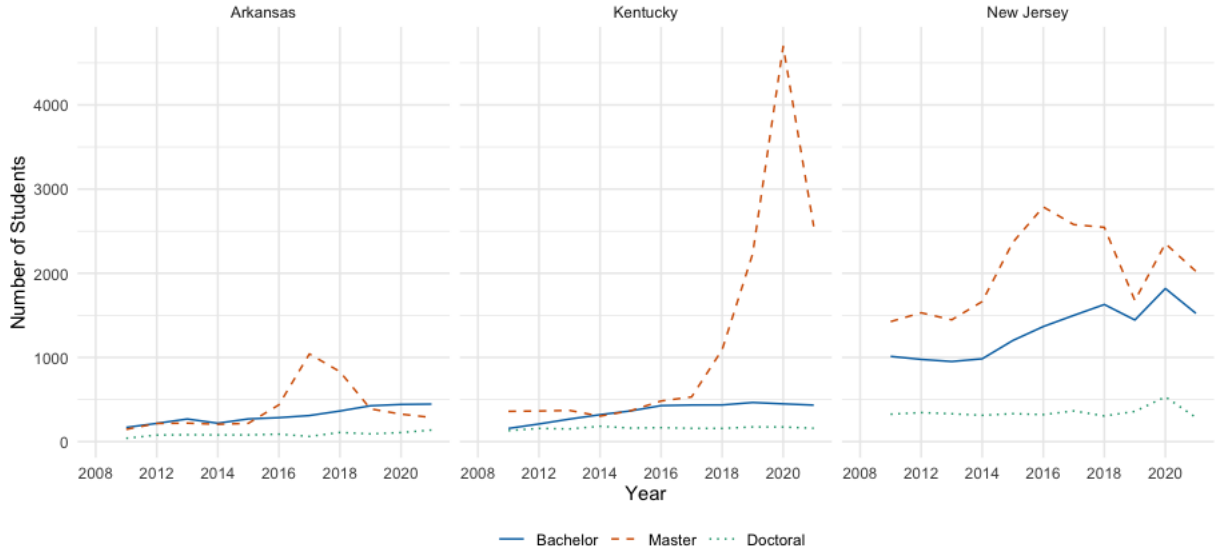


Figure 5. Proportion of Foreign-born Graduates in S&E Fields by Degree Level

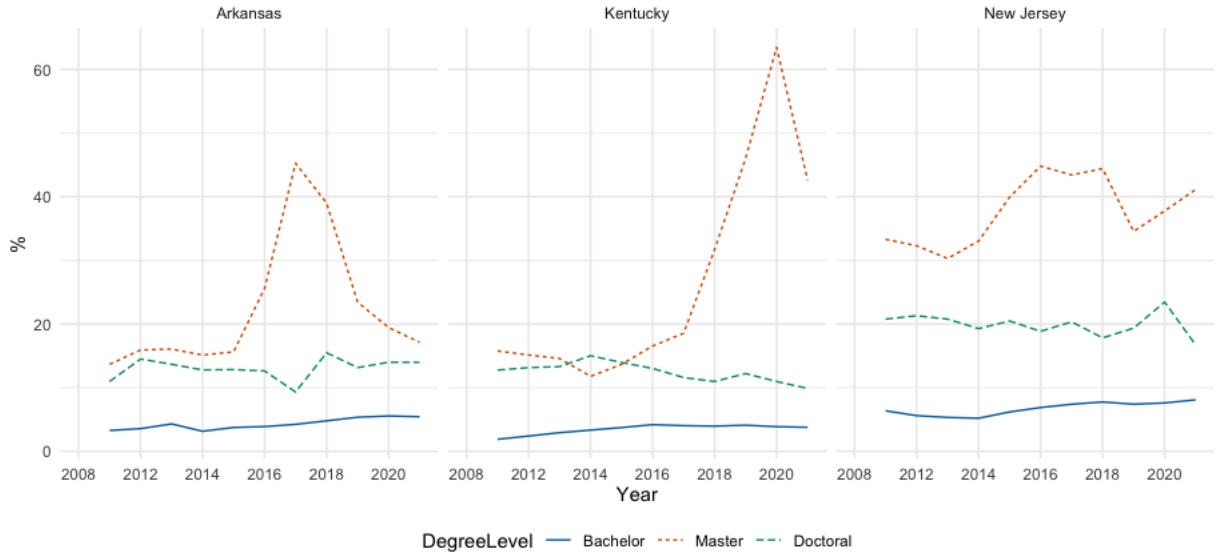
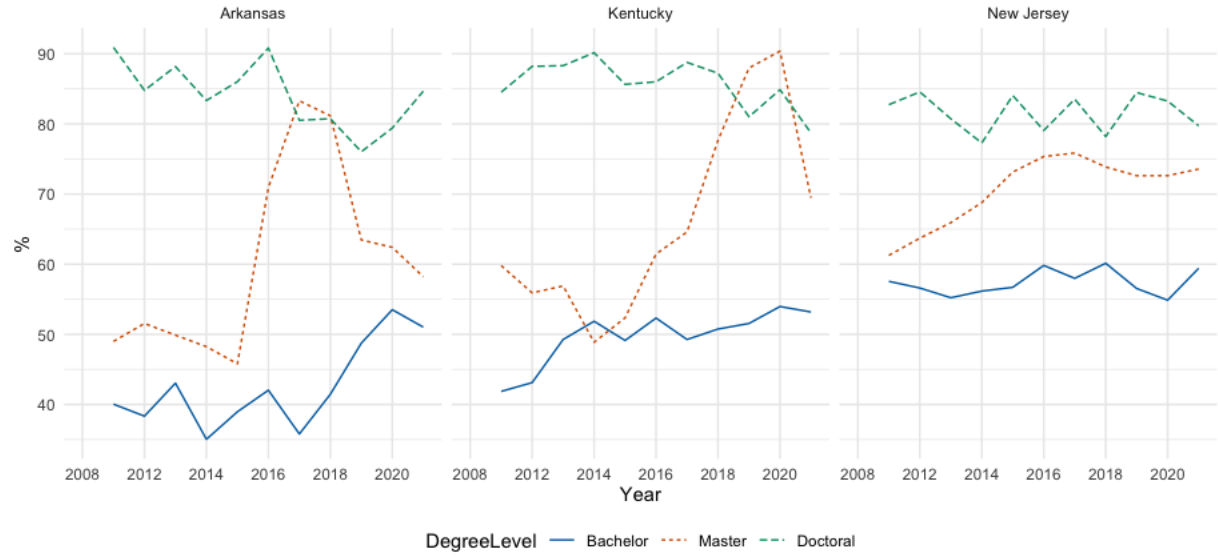


Figure 6. Proportion of S&E Majors among the Foreign-born Graduates by Degree Level



Comparison of Federal and State Data

Both the federal surveys and the state-level administrative data provide information on foreign-born and FBSE graduates. The preceding sections have discussed what these data say, and perhaps more importantly, what they cannot say. The NCES surveys are limited in their ability to be representative at the state and group level, while the NCSES data focuses on doctoral recipients only. Accordingly, this section will focus mainly on the comparison between the state-level data and the ACS Census data.

Benchmarking must begin with addressing the population that each data system includes and the definition of each key indicator. Table 6 summarizes the discrepancy in the definitions of measures across data sources. First, the state administrative data, as well as the NPSAS and SED data, attribute graduates to their institution and are accordingly assigned to the state in which the institution is located. Using the state that the institution is in gives us not only more comprehensive data coverage and consistency across all sources, but also, more accurate measures of investments made by states and institutions. Census data, on the other hand, report the respondent's legal or permanent residence. While this is useful information it is important to note that comparisons between the two sources may be influenced by this difference. It is also important to note that these individuals may have also received their degree in another country.

A foreign-born person is defined as an individual that was born outside of the continental U.S. and its territories and of non-U.S. citizen parents. Those who were born in the U.S. territories are considered U.S.-born because they are either U.S. nationals or citizens and receive similar social benefits/investments as the US-born. However, this definition can only be applied to the federal survey data because the state data only provide information on citizenship status. Arkansas and Kentucky record the student's country of origin, but this data is not complete. As U.S. citizenship can be obtained through naturalization, using citizenship status to identify foreign-born status will understate the size of the foreign-born population. On the other hand, federal survey data records detailed information on the nativity of both students and their parents, citizenship status, and immigration status. Therefore, using federal survey data, the number of foreign-born broken down by their citizenship status (naturalized vs non-citizen) is benchmarked.

Degree field is treated somewhat differently depending on the source of the data. Because the state administrative data designated the Classification of Instructional Programs (CIP) codes states were able to develop a master crosswalk informed by NCES which standardized how they defined S&E degrees. While the Census data does not use CIP codes the field of degree variable was coded to align with the established definitions of S&E.

Table 6. Definition Discrepancy across Data Sources

Measure	State Definitions	NCES Definition	NCSES Definition	Census Definition
Population Covered	Students who graduated with a U.S. degree	Students who enrolled in a U.S. degree program at the time of the survey	Students who graduated with a U.S. doctoral degree	Individuals who attained a degree
Time	Academic year	Academic year	Academic year	Calendar year
Location	State of institution	State of institution	State of institution	State of residence
Degree Origin	U.S.	Unknown	U.S.	Unknown
Science and Engineering Fields	CIP code of each degree level attained	23 NCES categories of majors of the program enrolled at the time of the survey	PhD field by NSF definition	Bachelor's field of degree (master's, doctoral N/A)
Degree Level	Level of the degree attained	Highest degree level ever **expected** (no attainment status) or the level of the **previously attained** degree	PhD level	Highest level of the degree attained
Foreign-born	Citizenship status	Nativity (immigrant status and parent's nativity)	Citizenship status	Citizenship status

Figure 7 compares the distribution of foreign-born and FBSE between state and Census data. It is consistent across all states and data sources in that the proportion of foreign-born students who majored in S&E fields versus all other fields is relatively stable over time. The percentages and the gaps between foreign-born and FBSE in Arkansas and Kentucky across state and census data are similar; however, in New Jersey, the percentages and the gap between foreign-born and FBSE in Census data are very different from those of state data. Despite the size of the gaps, in a relative sense, the FBSE is always about half of the foreign-born, regardless of the data source. It is important to recall that Census data reports the number of state residents while state data reports the number of graduates from in-state institutions. This may suggest that in Arkansas and Kentucky, most state residents graduated from in-state institutions, or at the least there are a similar number of those that exit and enter the state, while in New Jersey, a large portion of their foreign-born and FBSE residents graduated from out-of-state institutions in the U.S. or received their degree in another country. Finally, the census data does not account for the upturn in graduates in Arkansas and Kentucky in 2017 and 2019, respectively. This likely can be attributed to the sampling done for the ACS survey and the population covered.

Similarly, as shown in Figure 8, when breaking the data down by degree levels, the difference between state and Census data is the largest in New Jersey. Here the proportion of foreign-born bachelor's and doctoral degree holders is much larger, while the proportion of foreign-born master's degree holders is somewhat comparable post 2015. In Arkansas and Kentucky, the percentages of bachelor's and doctoral FBSE between state and Census data are largely comparable, however, the percentages of FBSE graduates with a master's degree are higher than that of FBSE residents. In addition, Census data does not exhibit the peaks for the master's degree like state data, likely due to the discrepancy in the population covered by the two data sources.

When looking at the proportions of foreign-born who graduated in S&E fields broken down by degree level in Figure 9, both state and Census data agree: the higher the degree level is, the more likely that a foreign-born student will choose an S&E field. Despite that similarity the data are otherwise quite dissimilar, particularly in Arkansas and Kentucky. Though beyond the scope of the current work this would be an interesting extension and possible use case for an expanded data infrastructure. It is plausible that the disparity here could be a result of FBSE graduates, specifically those with a master's or doctoral degree, exiting the state for employment elsewhere. Without a data system that can follow individuals across borders, or account for where they received their degree, there will remain a limited understanding of how attainment, and the investments that support it, translate into outcomes for individuals and states.

Figure 10 exhibits the distribution of doctoral graduates in the state versus the SED data. Recall that state administrative data cover all doctoral graduates including those who obtained professional degrees while the SED only covers those with research degrees. Also, state data do not necessarily cover all private institutions while the SED covers all accredited Ph.D. granting institutions. Looking at the FBSE population in panel (a), the series in Arkansas and Kentucky are very closely aligned, which implies that most FBSE doctoral graduates in these states obtained degrees in research rather than professional degree or state data has a good coverage of private in-state institutions. However, in New Jersey, the SED numbers are higher than state numbers. This could be due to the fact that state data does not cover all private institutions and there could be a considerable number of private institutions that offer professional degrees in New Jersey.

Looking at the total doctoral graduates in panel (b), the gaps between state and SED are much larger than in panel (a). This poses questions on whether foreign-born students are less likely to pursue professional degrees or if there are any constraints for foreign-born students to obtain professional degrees and/or pursue a career in those fields.

Figure 11 compares the number of FBSE with at least a bachelor's degree in the state administrative data versus in the NPSAS data. Recall that NPSAS 2012 and 2016 were collected through surveys while 2018 data was collected from state administrative data. Across the three states in 2012 and 2016, survey data shows larger counts of FBSE population than administrative data. Furthermore, although NPSAS 2018 data were collected from state administrative data, the NPSAS numbers in 2018 do not align with the administrative data provided by Kentucky and New Jersey. While the totals in Arkansas are similar, the NPSAS number is much higher in Kentucky and much lower in New Jersey compared to state data. This is likely a result of the fact that in Kentucky, 50% of the universe of schools reported data, and in New Jersey only 40% of schools reported data. More specifically, 40 and 90 private institutions in Kentucky and New Jersey, respectively, did not report data.

While there are some similarities between federal and state data, the disparities documented suggest that the administrative data is likely more representative, but it needs to be linked to additional data sources so that we can better understand the investments and outcomes of FBSEs.

Figure 7. Comparison of the Distribution of Foreign-born between State and Census Data

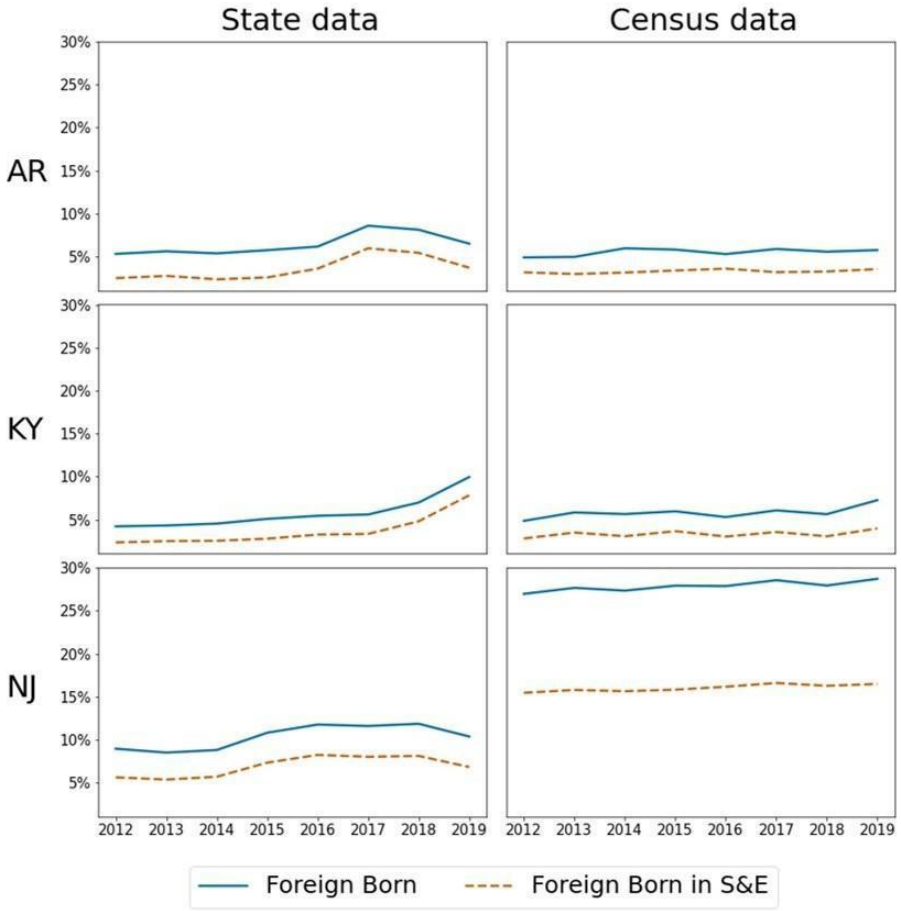


Figure 8. Comparison of the Proportion of Foreign-born in S&E between State and Census Data

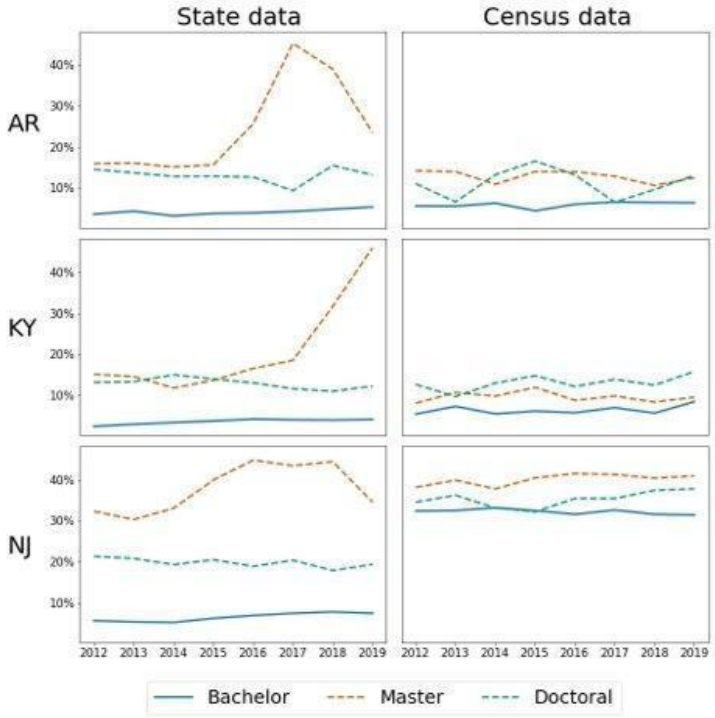


Figure 9. Comparison of the Proportion of S&E Majors among Foreign-born Students between State and Census Data

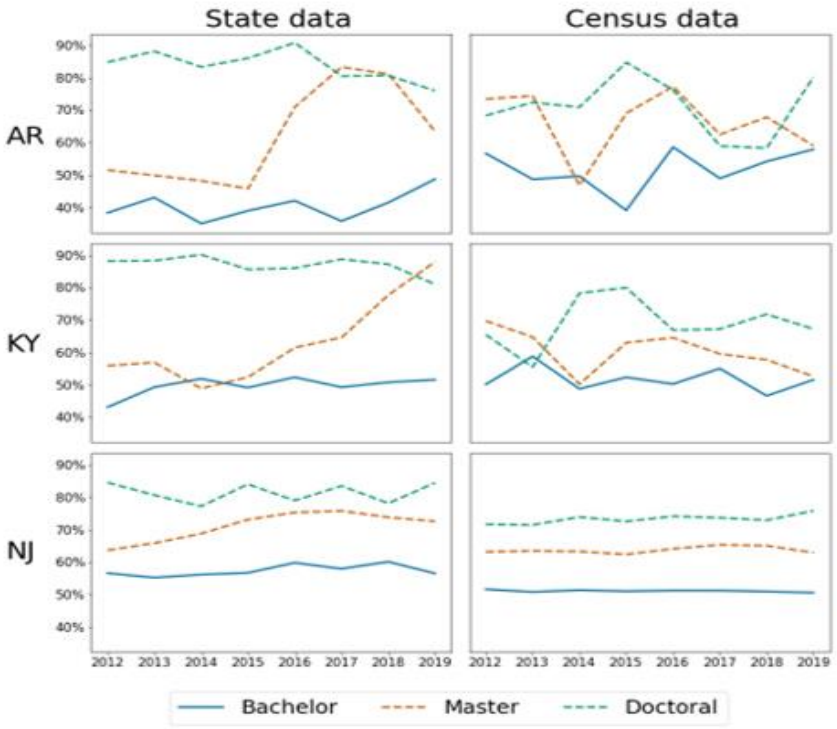
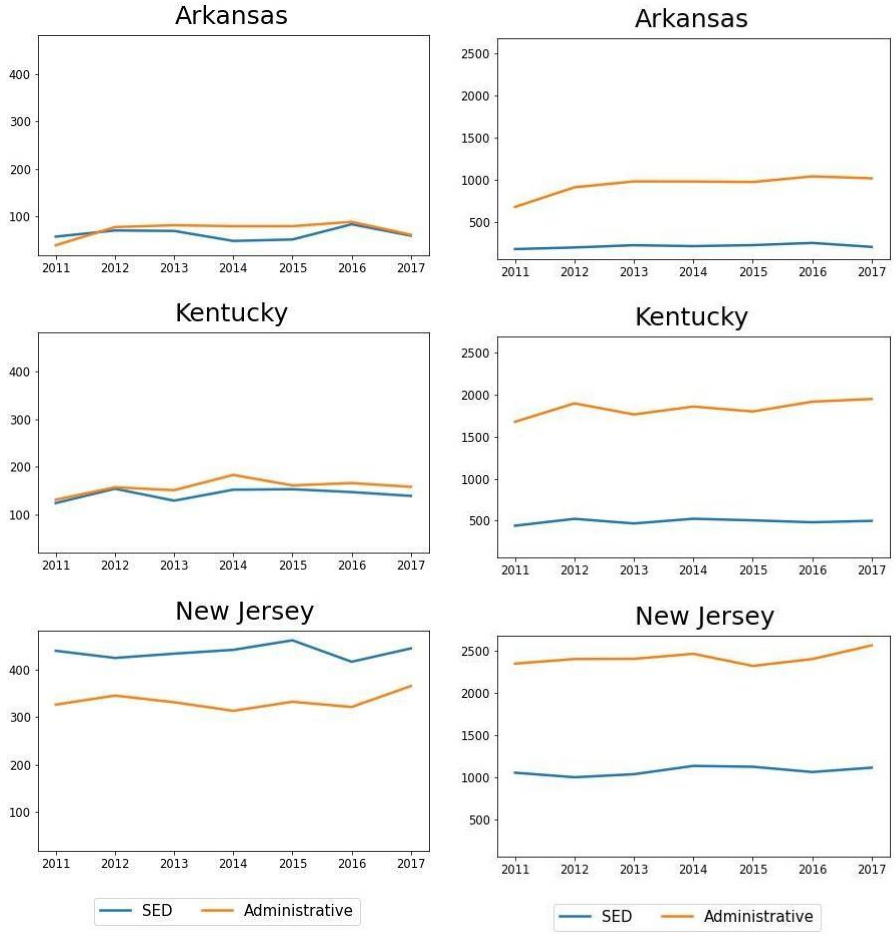
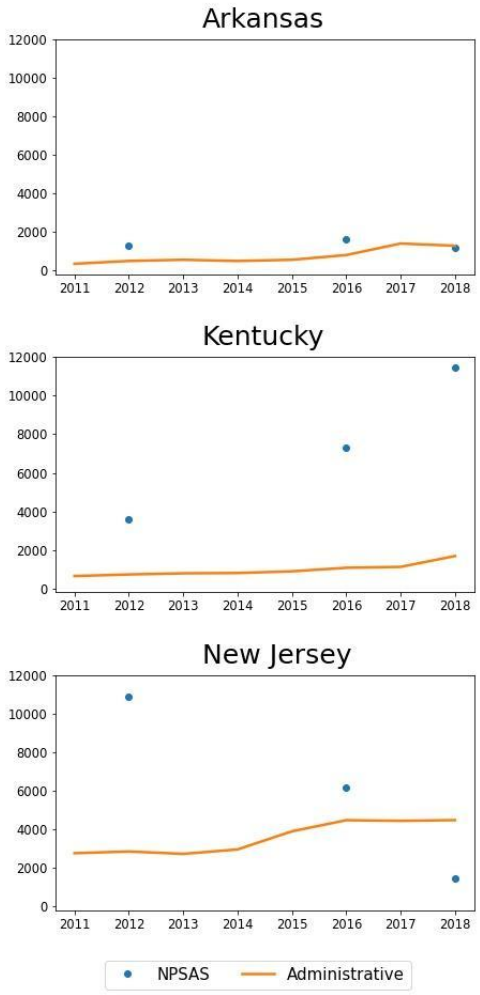


Figure 10. Distribution of Doctoral Graduates in the State Administrative versus the SED Data



(a) FBSE Doctoral Graduates (b) Total Doctoral Graduates

Figure 11. Counts of Foreign-born in Science and Engineering with at least Bachelor's Degree in the State Administrative Data and in the NPSAS Data.



Identification of Investments and Outcomes

Census level information about the number and proportion of foreign-born graduates in S&E fields as well as other fields is an important first step in creating the data infrastructure necessary to answer pressing policy questions. However, to fully realize the potential of the infrastructure we must also understand investment and outcome data, what is currently available, what are the gaps and limitations, and what are the opportunities.

Investments

The United States, across all levels of government, invests significant resources into higher education. In 2020, state and local investments in higher education amounted to $321 billion and accounted for roughly 17% of all state-level expenditures (Urban Institute, 2023.) Much of this funding supports the general operations of public institutions. On the other hand, federal investments in higher education make up a small percentage of total federal expenditures and are typically directed to individuals and research projects. There are also tax expenditure considerations such as deductions for interest paid on student loans. Given the complexity of the numerous investment streams and the various levels to which they are directed, this section focuses only on funding at the individual level.

Figure 12 demonstrates the undergraduate and graduate grants as a percentage of the total grants received by a student. We categorize grants into 4 sources: federal, state, institution, and outside. Federal grants include grants from the Department of Defense and Veteran benefits. The outside category includes private and employer grants. The statistics were computed using NPSAS data for the years 2012, 2016, and 2018. The percentages were calculated for each source in a given year, then averaged across the 3 years of data. Therefore, the total grant percentage (height of each column in Figure 12) of each group does not add up to 1.

While most of the undergraduate grants are from institutions and federal, most of the graduate grants are from institutions and outside sources. Among the three states, New Jersey has the smallest portion of grants that come from outside sources for both undergraduates and graduates. In general, across all grant sources, the percentages of grants received by foreign-born students are higher than that of their U.S.-born counterparts.

Figure 13 below depicts the distribution of the graduate and undergraduate debt of doctoral students by citizenship status[7] collected in the SED. Non-citizens are less likely to have any graduate or undergraduate debt. Notably, U.S. born, and naturalized citizens are more likely to have higher amounts of debt. Individuals with missing citizenship and missing graduate debt information are included as separate categories.

[7] Naturalized citizens are defined with CITIZ variable (type of citizenship) with the category called “U.S., naturalized”. Non-citizens are defined using the same CITIZ variable with the following categories: “Non-U.S., immigrant (permanent resident)”, “Non-U.S., non-immigrant (temporary resident)”, “Non-U.S., visa status unknown”. Foreign-born population is defined as the sum of naturalized and non-citizens. U.S. born is defined using the CITIZ variable with the category “U.S., native born”. Missing category is included where the CITIZ value is missing.

In terms of respondents' primary source of support, Figure 14 below depicts the distribution of the primary source of support of doctoral students by citizenship status[8] in the SED. Non-citizens are more likely to have research assistantship, teaching assistantship, foreign support as the primary source of funding. Notably, U.S. born, and naturalized citizens are more likely to have loans, personal earnings during graduate school, fellowship/scholarship, employer reimbursement, dissertation grant as the primary source of support.

The current data sources available to examine the investments made for FBSEs are helpful but largely inadequate. The survey data are not representative and do not provide connections to workforce and other post education outcomes so that a full ROI assessment can be done. There are a few options available to develop this infrastructure. The first is for states to link financial aid data to their postsecondary enrollment and graduation administrative data. While some states do this, the way in which many states interpret legal restrictions preclude them from creating the linked data. Another possible avenue is for states to request borrower and repayment data from the Department of Education. However, to date, this has not been done and questions remain as to whether policy would allow such a request. Finally, current surveys, such as the ACS, could be expanded to include such information.

Outcomes

Outcome data is important in providing a complete picture of the return on investments made to support FBSEs. Typically, outcomes are defined in terms of wage earnings, stable employment, career trajectory, etc. While these are important outcomes, they are largely drawn from existing UI wage data. More importantly, as will be discussed in more detail in the sections below, there are considerable challenges with linking data across systems for this group of individuals. Still, linkage issues aside, there is a significant opportunity to build a robust set of outcome data for this group.

Individual workforce outcomes are important as many foreign-born individuals fund their own education. For this group, it is reasonable to think about their individual ROI. However, because of eligibility requirement, many of these individuals do not appear in state Unemployment Insurance wage data. One possible way for states to address this issue is to explore the possibility of connecting postsecondary data with state income tax records. There are legal barriers to work through and some states have had to make changes to existing laws to allow for linkage. Accordingly, there are only a handful of states with this capability on a limited basis.

We know FBSEs do more than earn wages. They start businesses and employ people, they develop and register patents, they apply for grants and conduct research, and they invest in development themselves. Other work funded by NCSES has already begun to work through some of the linkage with

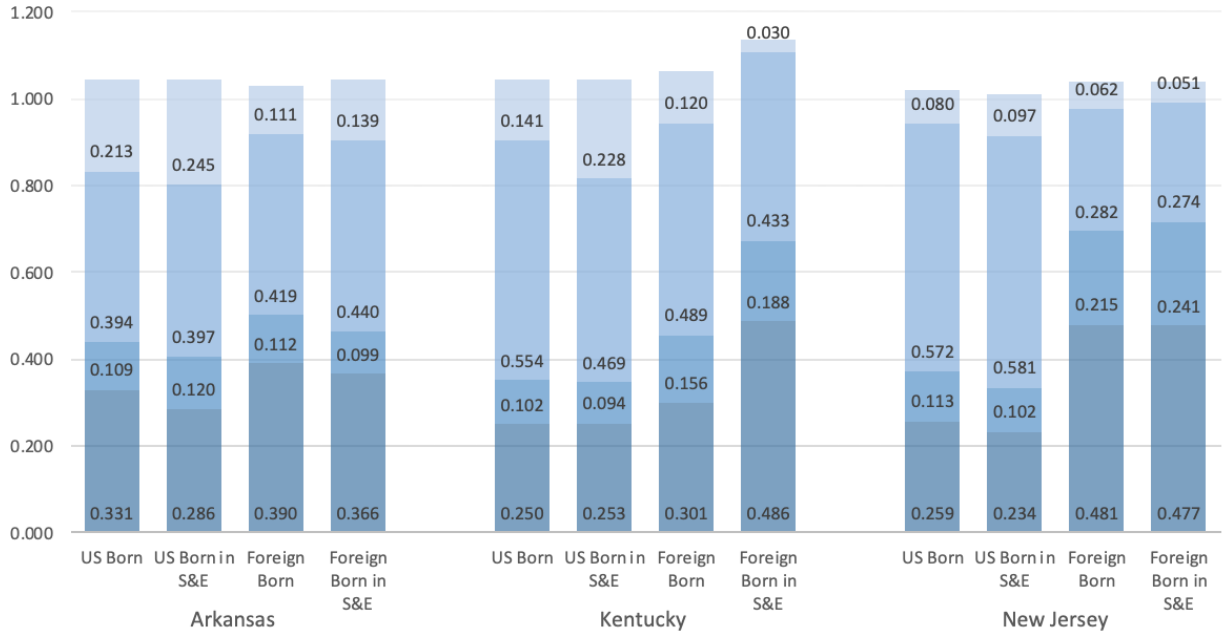
[8] Naturalized citizens are defined with CITIZ variable (type of citizenship) with the category called "U.S., naturalized". Non-citizens are defined using the same CITIZ variable with the following categories: "Non-U.S., immigrant (permanent resident)", "Non-U.S., non-immigrant (temporary resident)", "Non-U.S., visa status unknown". Foreign-born population is defined as the sum of naturalized and non-citizens. U.S. born is defined using the CITIZ variable with the category "U.S., native born". Missing category is included where the CITIZ value is missing.

Patents View data and federal grant award data. Though beyond the scope of this current work, there are promising opportunities to build a robust set of outcome indicators.

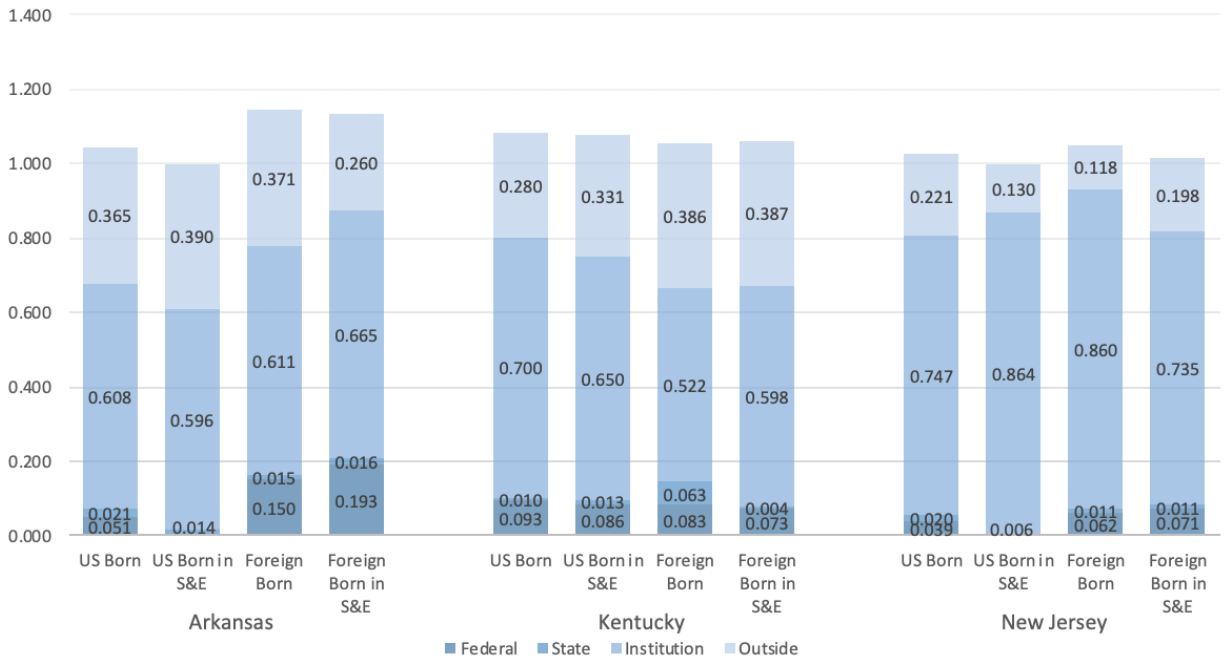
Because there are limitations to linking the data the assessment of current outcomes is limited. However, one core question that can be addressed is the proportion of FBSEs that remain in the U.S. post-graduation. In term of the postgraduation location of respondents, Figure 15 below depicts the distribution of the postgraduation location of doctoral students by citizenship status[9] in SED. U.S. born and naturalized citizens are more likely to have U.S. as the postgraduation location.

[9] Naturalized citizens are defined with CITIZ variable (type of citizenship) with the category called "U.S., naturalized". Non-citizens are defined using the same CITIZ variable with the following categories: "Non-U.S., immigrant (permanent resident)", "Non-U.S., non-immigrant (temporary resident)", "Non-U.S., visa status unknown". Foreign-born population is defined as the sum of naturalized and non-citizens. U.S. born is defined using the CITIZ variable with the category "U.S., native born". Missing category is included where the CITIZ value is missing.

Figure 12. Undergraduate and Graduate Grant as Percentage of Total Grant by Source

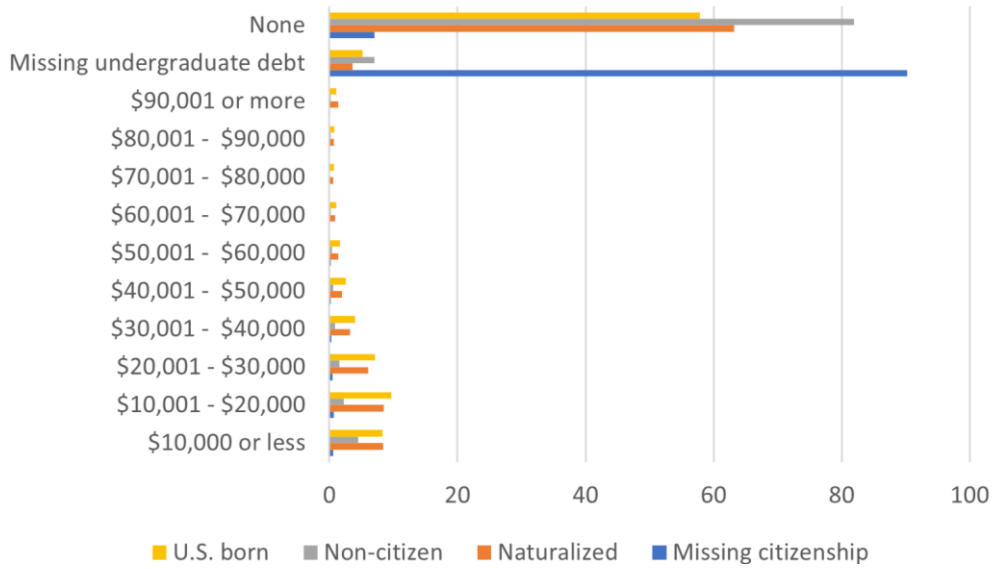


(a) Undergraduate

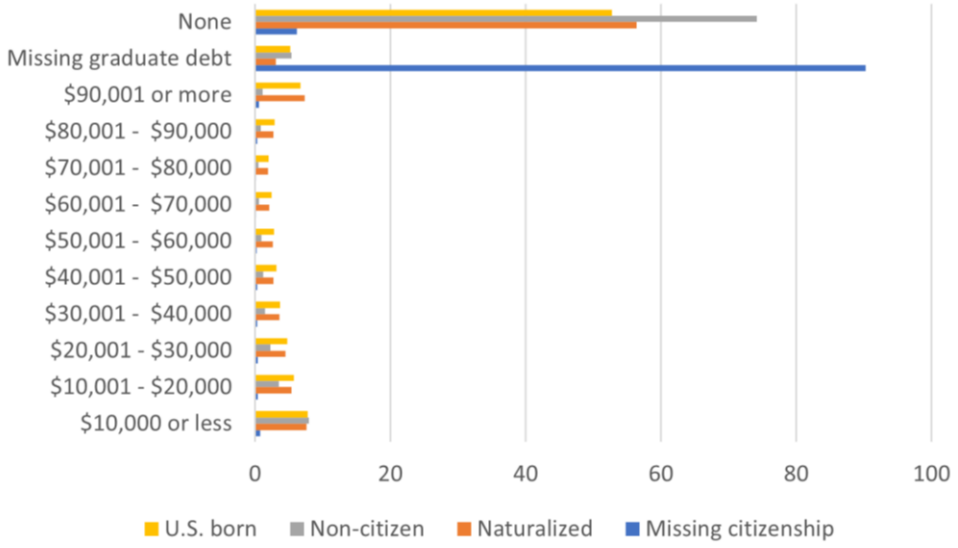


(b) Graduate

Figure 13. Non-citizens are Less Likely to Have any Graduate or Undergraduate Debt.



(a) Undergraduate



(b) Graduate

Figure 14. Non-citizens are More Likely to Have Research Assistantship, Teaching Assistantship, and Foreign Support as the Primary Source of Funding

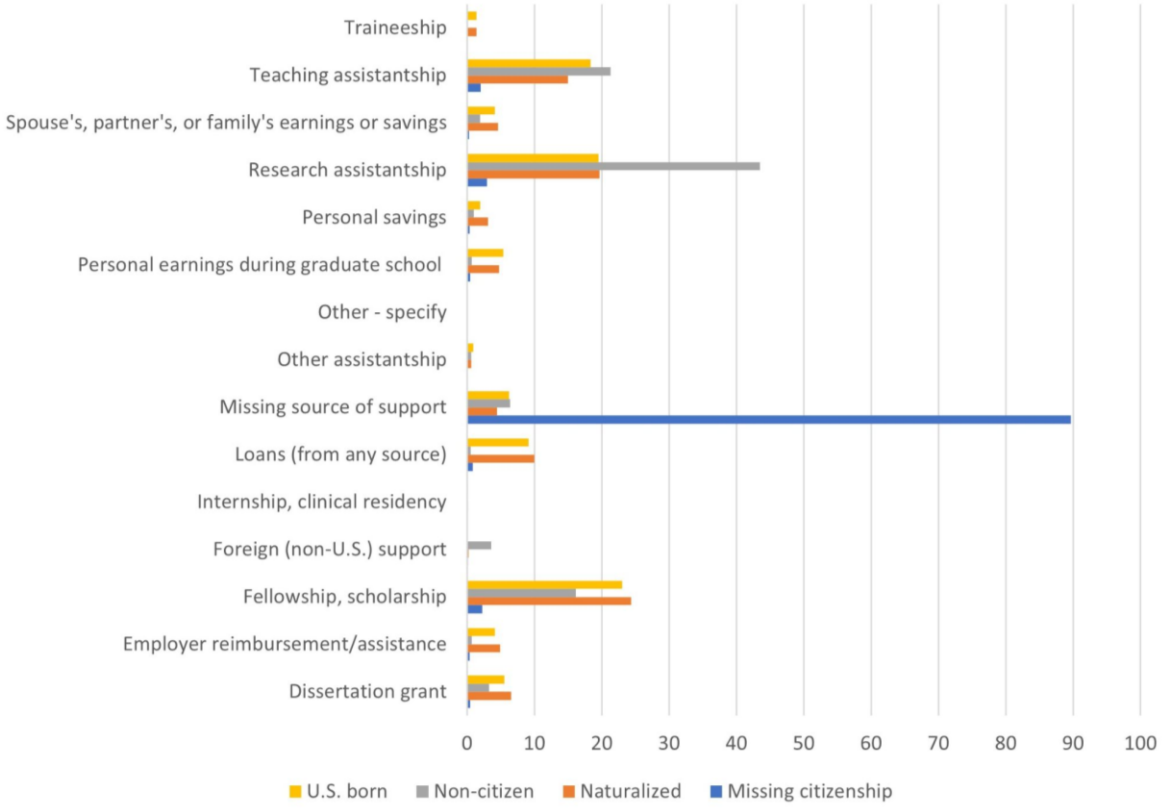
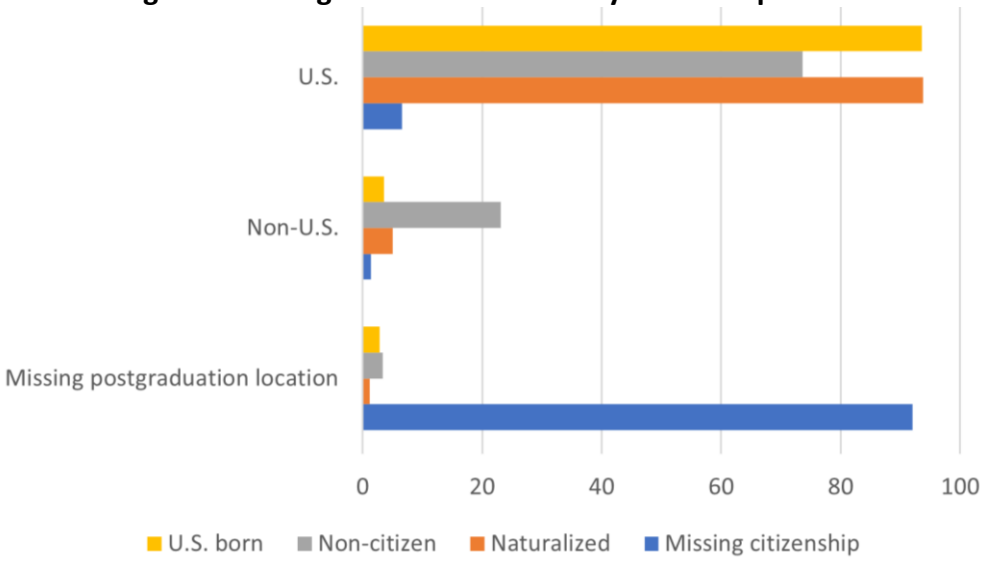


Figure 15. Postgraduation Location by Citizenship Status



Building a Data Infrastructure with Linked Educational Data and Outcomes

Given that federal investments contribute to the recruitment and retention of foreign-born scientists and engineers, linking these investments to outcomes at the individual level would provide researchers the opportunity not only to quantify the degree to which an individual is employed or the associated wages, but also to identify the level of investment supplied by the federal government. This would provide an important input in measuring the ROI for FBSE in the U.S.

Currently, surveys such as the NPSAS and SED provide a view into the complex world of federal funding and loans but have no way of being linked to wage records (via SSN) or tax records. Surveys may also be limited in scope in terms of what is considered a source of investment. NPSAS and SED cover various loans and grants, but there may also be other factors considered investment that are unaccounted for by a survey. Discussion on the data model to follow will cover current and suggested data elements for linking postsecondary data to outcomes, but not on investments. Measuring investments at the individual level and the ability to link those investments to postsecondary data and outcomes will prove a necessary component for the FBSE infrastructure going forward.

Feasible Data Model

Data models are helpful in clearly defining the necessary information required to answer questions for which current data infrastructure cannot answer. They indicate sources, the attributes required from those sources, how those attributes are defined and constructed, and how the disparate sources will link together. Two data models were developed with input from expert advisory panels: a feasible data model using the current state longitudinal system's data and an aspirational model. This section focuses on the feasible data model.

Prior to developing the data models, the team first identified and received feedback on research questions and use cases to inform the development of the required data elements. Below, the most important use cases that researchers are likely to use the data in a secondary data analysis are presented:

- What is the supply and demand for FBSEs? What are the in-demand occupations they may be trained in?
- What is the economic impact on the United States of various immigration policies?
- What are the effects of U.S. industry investment and higher education investment in bringing in and training students who are foreign-born?
- How do we measure return on government, academic, or industry investment beyond FBSE earnings?

This information was used to develop the feasible data model (which can be implemented with existing data) and the aspirational data model (which can be implemented with further data system development and/or data sharing/linkages). The feasible data model, presented in Table 7 and Figure

Table 7. Entity Attribute Table for Feasible Data Model

Table Name	Attribute	Type	Definition
Foreign-Born	Citizenship	Numeric	U.S. citizenship indicator
Secondary Education	Student ID	Alphanumeric	Unique hashed ID for individual students
	LEA Enrollment	Alphanumeric	Local Education Agency at which student is enrolled
	District ID	Numeric	School district code
	CTE Enrollment	Alphanumeric	Completion of CTE certification
	Graduation Year	Numeric	Award year derived from award date (YYYY)
Postsecondary Education	Social Security Number[1]	Alphanumeric	Student social security number (hashed)
	Race	Numeric	Series of binary variables with race categories: American Indian/Alaska Native; Asian; Black/African American; Hispanic; Native Hawaiian/Pacific Islander; White
	Year of Birth	Numeric	Birth year
	Institution ID	Numeric	Unique institution code
	Financial Aid	Numeric	Term award amount from state and (where eligible) federal financial aid programs
	Classification of Instructional Program (CIP)[3]	Numeric	Code for degree majors that are approved degree/formal award programs, and are categorized and coded according to the CIP manual
	Award Type	Alphanumeric	An indication of the degree/ certificate conferred during the fiscal year reporting period. For multiple awards in the same reporting period, multiple records must be submitted for the student (SURE Code: D11).
	Award Date	Numeric	Award month and year derived from award date (MMYYYY)
Workforce – Employee	Social Security Number	Alphanumeric	Individual social security number (hashed)
	Year	Numeric	Year (YYYY) the wage applies
	Quarter	Numeric	Quarter the wage applies
	Quarterly Wages[2]	Numeric	Quarterly wages
	Employer ID	Alphanumeric	Federal Employer Identification Number (hashed)
Workforce – Employer	Year	Numeric	Code indicating the year the file was updated
	Quarter	Numeric	Code indicating the quarter the file was updated
	Employer ID	Alphanumeric	Federal Employer Identification Number (hashed)
	Industry Code[3]	Alphanumeric	The North American Industry Classification

[1]Where SSN is not present for students who are not citizens and have not obtained an SSN, we will use a persistent identifier and use probabilistic matching to wage records.
[2]Partner states involved in the initial pilot use of this data model will develop a shared approach for addressing records with $0 wages regarding inclusion/exclusion rules.
[3]Due to the longitudinal nature of data in these analyses, the most recent CIP and NAICS codes will be used as reference for program of study and industry of employment. Where possible, historical records will be updated with a crosswalk.

16, can be used immediately to link data across state longitudinal data systems to measure the presence and educational and employment outcomes of FBSEs.

However, there are several limitations to state data, including an inability to identify individual occupations and a limited view on income to what is included in state administrative wage data. Additionally, there are limitations in using social security as a linkage identifier for a population that is less likely to have that specific identifier. The first task for the Foreign-Born Scientists and Engineer portion of America's Data Hub is to analyze the availability of and demand for scientists and engineers on a national scale. That includes building evidence to fully understand the public value of recruiting scientists and engineers from other countries and training them in U.S. universities and labs. Record linkage for foreign born populations poses some unique challenges, and this research, conducted as part of a multi-state collaborative effort coordinated by the Coleridge Initiative, seeks to propose actionable recommendations for assessing and improving record linkage performance and bias for the foreign-born population[10], which also has much broader potential benefits to other populations and to administrative record linkage in general.

In Arkansas, individuals attending postsecondary institutions were identified as U.S.-born or foreign-born and subsequent analysis of SSN validity was conducted. Current linkage methods (and subsequent policy recommendations) rely heavily on the presence of SSN to identify individuals. In Arkansas Higher Education records, at baseline, the rate of having an invalid SSN was substantially higher for the foreign-born than their U.S.-born counterparts.[11] This implies that the matching rate using only SSN as an identifier will be much lower among the foreign-born and may require additional identifiers to create a successful linkage from postsecondary information to future earnings.

Presented below are findings from the Arkansas linkage work:[12]

Assessment of individual identifiers for Higher Education and UI Wage administrative data found that:

- SSN is available on all UI Wage records but only valid on 66% of Higher Education records for FBSE.
- First Name and Last Name are available across both sources with a high level of completeness.
- Middle Name is only complete across 55% of UI Wage and 48.2% of Higher Education records.
- Date of Birth is available on Higher Education records but not UI Wage records.
- Additional demographic identifiers (gender, race/ethnicity) are available on Higher Education records, but not on UI Wage records. The only individual identifiers present across both sources

[10] For Arkansas higher education administrative records, inclusion in the population of interest (foreign-born) was determined by a Non-US Resident value of "Yes" or a County or Origin other than "USA".

[11] To assess SSN completeness in Higher Education records, the Social Security Administration's validation criteria were applied to the Unique Identification Code values in Arkansas administrative data. While 98.62% of total Arkansas Higher Education records were found to have Unique Identification Code values with a valid SSN format, only 66% of foreign-born postsecondary graduates were found to have a valid SSN format. Valid SSN format does not guarantee that the value present is actually an SSN, but invalid SSN format does indicate that the value present is not an SSN and not a candidate for SSN match.

[12] <<<Link to full linkage report here>>>

are SSN, First Name, Last Name, and Middle Name. These are the only candidate attributes currently available for use in record linkage.

There are multiple possible approaches for record linkage, each with different advantages depending on characteristics of the source data (in this case, the abovementioned information may exist with varying completeness) and the intended use of the linked data. To identify the most relevant record linkage approaches for the population of interest:

- The predominant record linkage approaches and their respective benefits and applicability were surveyed through a comprehensive literature review. Deterministic record linkage, probabilistic record linkage, and machine learning approaches are covered.
- The performance (accuracy) of representative algorithms for each type of record linkage was assessed through testing with synthetic truth data sets.
- Record linkage approaches supported by the available identifying attributes were assessed for performance (accuracy) against a curated truth set constructed using actual administrative data.

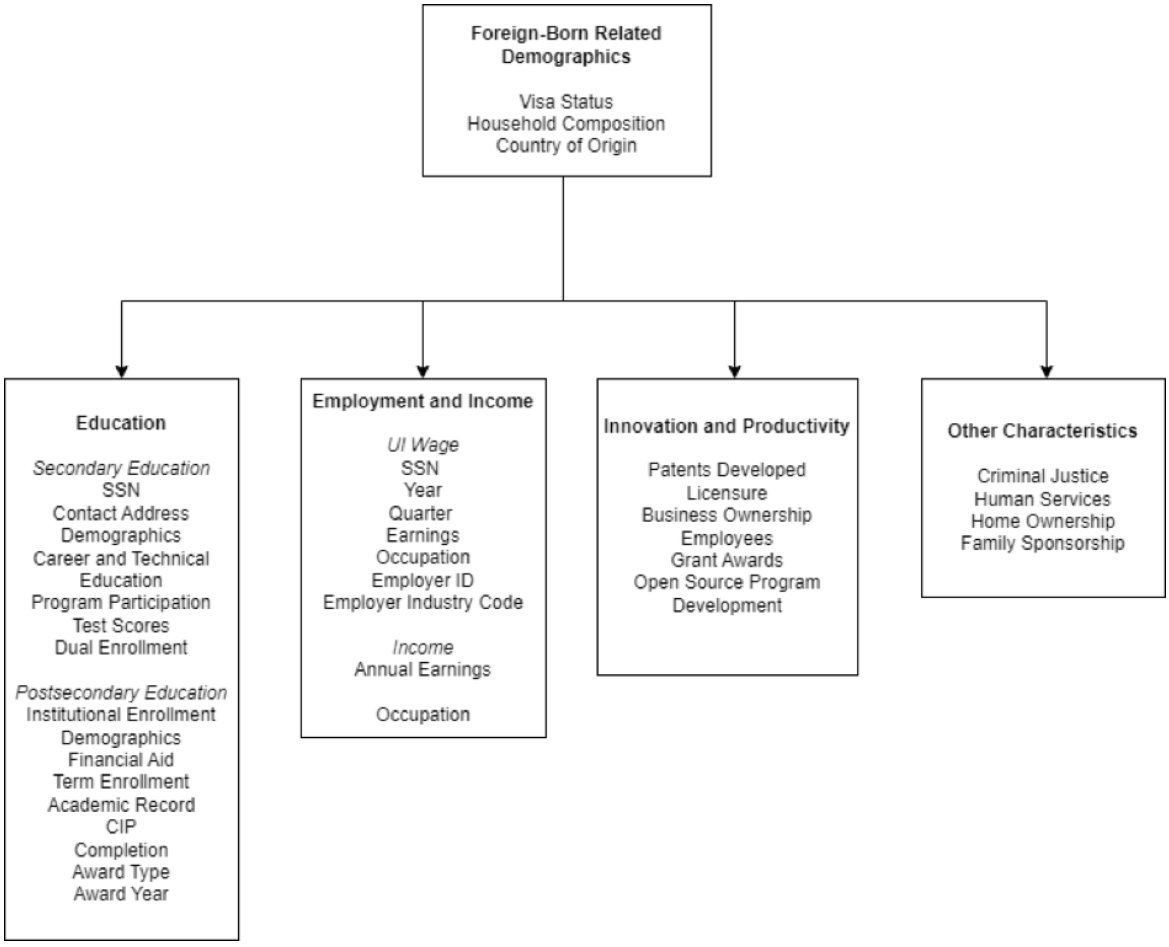
Findings:

- The incumbent record linkage solution for the population of interest, deterministic linkage on Social Security Number, cannot yield higher than 66% successful record linkage on the representative administrative data due to the lack of valid SSNs for 34% of the population of interest.
- Deterministic record linkage on Name alone was not found to be a viable record linkage approach.
- Deterministic record linkage on Name and Date of Birth was found to be a viable record linkage approach in absence of SSN if Date of Birth were available on UI Wage records.
- Probabilistic record linkage was assessed for the population of interest but was not found to result in significant improvement in accuracy to warrant the additional complexity of implementation given available identifying attributes.
- Machine learning approaches (neural networks and transfer learning) were assessed, and while these approaches demonstrated impressive accuracy and computational performance on synthetic truth data sets, there was insufficient data available for the population of interest for these approaches to be immediately applicable or beneficial.
- Facilitation of both deterministic and probabilistic record linkage approaches is recommended for achieving a balance of record linkage fidelity and computational efficiency while affording data analysts more versatile linkage options based on data use and tolerance for false negatives, false positives, and overall predictive performance.

Ideal Data Model

The aspirational data model illustrated in Figures 16 includes the many other topics identified in the use cases, expanding what we know about FBSEs. The proposed aspirational data would include the data tables and elements presented in Figure 16.

Figure 16. Proposed Entity Attributes for Aspirational Data Model



Discussions with panel participants and project partners identified several nuances to data model implementation, including the desire to track individuals as they engage with the higher education system such that if someone who is foreign-born graduates with a bachelor's in a non-S&E field but a master's in an S&E field, they are still eventually categorized as FBSE. In addition, a future topic of interest is those who are working for U.S. companies from abroad.

It is also important to think through sources of data for the aspirational data model. A key question is whether or not this data already exists in an administrative system or what effort would need to be made to create the data. Foreign-born related demographics (top box in Figure 16) could be obtained from the SAVE data system administered by the U.S. Citizenship and Immigration Service (USCIS). This may also serve to standardize definitions across the various current data systems. The education section can largely be addressed by current state longitudinal data systems, though states would need to work through the legal framework necessary to access K-12 data since state interpretation of FERPA requirements varies. Employment and income could be handled through Unemployment Insurance wage data but as discussed it is likely that many of foreign-born workers are not eligible for unemployment insurance and thus are not covered. To address this gap states could augment their

income data by including state income tax records. Innovation and productivity data could come from any number of data systems. At a minimum the infrastructure should include data from the registry of patents from the U.S. Patent and Trademark Office (USPTO), data on business ownership from state level departments of state, licensure data available from various agencies at the state level, and grant awards from various federal and state agencies.

Panel participants discussed the range of support they would need to implement such a data model. Regardless of the data model type, participants listed data dictionaries and model documentation as necessary resources. In addition, participants noted the importance of a template data sharing agreement to allow them to implement the data model, and a proof-of-concept report that would highlight the importance of this work to make the argument in favor of the effort necessary to implement data sharing agreements.

Without the necessary data elements (as in the aspirational data model) to follow individuals throughout their U.S. career, or without appropriate identifiers, data linkage approaches must be used. As mentioned above, the foreign-born population is especially susceptible to being excluded from analyses involving SSN as an identifier. This can result in biased results that have implications for policies targeting this population. The section below covers the work done by collaborators (Arkansas) on their assessment of record linkage bias for FBSEs and ways to mitigate it.

Assessing and Mitigating Record Linkage Bias

Due to the identification of unsuccessful record linkage for the population of interest, a literature review was conducted on approaches and best practices for assessing, mitigating, and communicating record linkage bias.
Key findings from the literature review include:

- There should be awareness and education efforts to train users of linked administrative data on the existence, impact, measurement, and mitigation of record linkage error and bias as well as how to communicate record linkage methods, performance, and bias. (Wiegand et al., 2019)
- Data analysts should assess and report on the quality of linked data used for analysis, including how analyses took linkage error and bias into account.
- Measuring and mitigating the presence and impact of record linkage errors and bias requires infrastructure design considerations to allow for more transparency into record linkage processing and performance. (Ruth et al., 2018)
- Data providers should make details available on the population included in the data set, the coverage, and the data generation or collection mechanism. (Harron et al., 2020)

Recommended approaches for assessing and mitigating record linkage bias were tested with statewide administrative data to determine feasibility, implementation requirements, and impact on results.

One of the most common metrics included in federal reporting and consumer information products leveraging linked higher education and UI wage administrative data is the percentage of graduates who are employed one year (or other intervals) post completion. This metric is calculated as the number of higher education completers found in UI wage records at the interval of interest divided by the total number of higher education completers in the period being assessed. Failed record linkage due to

insufficient identifiers essentially removes completers from the numerator, artificially lowering post completion employment due to record linkage bias.

Testing was performed to mitigate this bias in analyses by:

- Adding an SSN validity indicator to the source data prior to deidentification. This is important because validity rules cannot be applied to hashed identifiers.
- Making the full population of source records available to analysts with transparency into which records were successfully or unsuccessfully linked.
- Incorporating additional data quality and linkage metadata into the analysis in order to remove records that could not be linked due to invalid SSNs from the denominator since they are removed from the numerator. Removal of these records essentially treats this as a sample statistic versus a population statistic, and the increased transparency allows for communication of the confidence in the statistic.

Mitigation of record linkage bias through improved record linkage transparency led to a 47% change in post-completion employment statistics for the population of interest, suggesting a material impact to data and evidence on the foreign-born population, the programs from which they graduate, and the science and engineering and STEM workforce of which they constitute a significant percentage.

Recommendations

Recommendations for improving the fidelity of administrative data linkage in support of evidence-based policy and practice include:

Awareness, Measurement and Mitigation

- There should be awareness and education efforts to train users of linked administrative data on the existence, impact, measurement, and mitigation of record linkage error and bias as well as how to communicate record linkage methods, performance, and bias.
- Data analysts should assess and report on the quality of linked data used for analysis, including how analyses took linkage error and bias into account.

Record Linkage Approaches

- Measuring and mitigating the presence and impact of record linkage errors and bias requires infrastructure design considerations to allow for more transparency into record linkage processing and performance.
- Facilitation of both deterministic and probabilistic record linkage approaches is recommended for achieving a balance of record linkage fidelity and computational efficiency while affording data analysts more versatile linkage options based on data use and tolerance for false negatives, false positives, and overall predictive performance.

Data Collection and Preparation

- Data providers should make details available on the population included in the data set, the coverage, and the data generation or collection mechanism.
- A key limiting factor to record linkage fidelity is the lack of identifying attributes on some key administrative data sources. The lack of identifying attributes beyond Social Security Number

and Name on UI Wage data is particularly limiting due to the broad use and relevance of administrative data on employment and earnings.

- Efforts to enhance the collection of individual labor market information data should consider not only information gain from additional observational attributes (occupation, hours worked) but also enhanced collection of individual identifiers to reduce information loss from record linkage error.
- Government and employer participation in the Jobs and Employment Data Exchange (JEDx) initiative has the potential to not only provide more timely, detailed, and relevant administrative data, but also improved record linkage fidelity through the inclusion of additional individual attributes.

Governance and Privacy

Building a robust data infrastructure is more than addressing the technical issues of identifying and linking disparate data and developing common data standards that support it. Whether developed at the federal level, the state level, or a combination of the two, the data infrastructure must be supported by an equally robust governance framework that facilitates cross-agency data sharing and access. The following recommendations were developed through the work focused on building a data infrastructure for foreign-born scientists and engineers but can apply broadly to any effort focused on building large-scale data systems across any number of agencies across state and federal levels.

A key first step in developing a robust data infrastructure is to assess the current state of disparate data, the existing gaps, and what data is needed to attend to those gaps. This work can be done without directly linking the data which has the benefit of avoiding the need to navigate individualized data sharing agreements on a case-by-case basis. This is an important feature as all necessary data can be determined without investing resources into establishing data sharing agreements in an ad hoc fashion.

Once all the data assets are determined it is important to identify all data stewards that govern each data set so that a comprehensive multiparty data sharing agreement and a governance structure can be established. While each party typically has a data sharing agreement, best practices suggest that the parties identify the commonalities across each agreement as a starting point. From there, key topics that must be addressed in the data sharing agreement are the record linkage protocols, where the finalized data infrastructure will be housed and how it will be accessed, the protocols for requesting and being granted access, the access modalities available, responsibilities around data disclosure review and release of final data products, and data destruction. Finally, the data sharing agreement should establish a governing body that provides representation for all parties and a set of protocols that allow the governing body to navigate changes to the agreement as needed.

The keystone of building a robust data infrastructure is the ability to link data from various sources. This often requires the availability of personally identifiable information (PII). Here, the data sharing agreement should determine what elements will be available for linking, where the linkage will occur, how the data will be transferred, and who will have access to the PII. The agreement should also address how often the linkage will be updated. Depending on the sources available, this will typically happen every one to two years.

Once the data infrastructure is linked, the data agreement should identify where it will be hosted. Given the nature of the data, particularly in the case of foreign-born individuals, it is recommended that all PII be hashed, and that the infrastructure resides in a FedRAMP certified secure cloud-based facility. The hashing provides additional safeguards against potentially improper use, and the secure cloud-based platform ensures that the data is accessible to a broad community without sacrificing security.

Access to the facility should be strictly managed by the governing body and operate similarly to requests for restricted-use federal data. Here, a potential principal investigator will propose a research project, what data the project will need, the researchers that will be a part of the project, and the term of the project. Each researcher should complete a notarized nondisclosure agreement (NDA) as determined by

the governing body. The NDA should address usage and access protocols e.g., use is strictly for statistical purposes, access from a secure U.S. location. If approved for access the governing body will submit the request for a project workspace to the secure cloud-based facility where only the approved researchers will be given access to only the approved data for the specified project term. In the case of the data infrastructure for foreign-born scientists and engineers, it is imperative that the infrastructure is used solely for statistical purposes. Again, given the population in question for this work, it is important to safeguard against unapproved uses.

The governing body and the data sharing agreement implemented should also consider disclosure rules for aggregated data and statistics coming out of the secure facility. They should consider issues such as cell size suppression, a minimum number of contributing entities to each cell, and secondary disclosure. The agreement should also designate the parties responsible for the review and where the review will take place. Given the nature of the data infrastructure in question, the responsible parties should be approved to see the raw data from each contributing agency to simplify the review process. At a minimum, preliminary disclosure reviews should occur within the secure facility to mitigate the potential for disclosive information being exported for reviews outside the secure facility.

In sum, the core issue in building the type of infrastructure necessary to better understand the investments and outcomes of foreign-born scientists and engineers is the tradeoff between the risk to privacy and the utility of the data. As more and more data are brought together the risk to privacy increases, but so does the potential to better inform policies and programs that can support this important group of individuals. The governing body and the work that they support must consider this balance and safeguard against improper use and maintain privacy preservation while supporting a robust portfolio of research.

References

Abramitzky R, Boustan L. 2017. Immigration in American Economic History. *Journal of Economic Literature* 55(4):1311–45.

Basso, G., & Peri, G. (2020). Internal mobility: The greater responsiveness of foreign-born to economic conditions. *Journal of Economic Perspectives*, *34*(3), 77-98. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.34.3.77

Chair Cecilia Rouse, Lisa Barrow, Kevin Rinz, and Evan Soltas (2021). The Economic Benefits of Extending Permanent Legal Status to Unauthorized Immigrants. *The White House*. https://www.whitehouse.gov/cea/written-materials/2021/09/17/the-economic-benefits-of-extending-permanent-legal-status-to-unauthorized-immigrants/

Espenshade, T. J., Usdansky, M. L., & Chung, C. Y. (2001). Employment and Earnings of Foreign-Born Scientists and Engineers. *Population Research and Policy Review*, *20*(1/2), 81–105. http://www.jstor.org/stable/40230299

Hunt, J., & Gauthier-Loiselle, M. (2010). How much does immigration boost innovation?. *American Economic Journal: Macroeconomics*, *2*(2), 31-56. https://pubs.aeaweb.org/doi/pdf/10.1257/mac.2.2.31

Khanna G, Lee M. 2019. High-Skill Immigration, Innovation, and Creative Destruction. In Ganguli I, Kahn S, MacGarvie M, editors, *The Roles of Immigrants and Foreign Students in U.S. Science, Innovation, and Entrepreneurship*. National Bureau of Economic Research Conference Report, pp. 73–98. Chicago: University of Chicago Press.

Kerr SP, Kerr WR. 2017. Immigrant Entrepreneurship. *In* Haltiwanger J, Hurst E, Miranda J, Schoar A, editors, *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges.* National Bureau of Economic Research Studies in Income and Wealth, Vol. 75, pp. 187–249. Chicago: University of Chicago Press.

Levin, S. G., Black, G. C., Winkler, A. E., & Stephan, P. E. (2004). Differential employment patterns for citizens and non-citizens in science and engineering in the United States: Minting and competitive effects. *Growth and Change*, *35*(4), 456-475.

National Academies of Sciences, Engineering, and Medicine. 2017. *The Economic and Fiscal Consequences of Immigration*. Washington, DC: The National Academies Press. https://doi.org/10.17226/23550.

National Science Board, National Science Foundation. 2022. *Science and Engineering Indicators 2022: The State of U.S. Science and Engineering*. NSB-2022-1. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsb20221

Ottaviano, G.I.P. and Peri, G. (2012), Rethinking the Effect of Immigration on Wages. *Journal of the European Economic Association*, 10: 152-197. https://doi.org/10.1111/j.1542-4774.2011.01052.x
Scott A. Wolla. (2014) The Economics of Immigration: A Story of Substitutes and Complements. *Federal Reserve Bank of St. Louis*. Available at https://research.stlouisfed.org/publications/page1-econ/2014/05/01/the-economics-of-immigration-a-story-of-substitutes-and-complements/

Urban Institute. 2023. *State and Local Backgrounder: State and Local Expenditures.* Available at https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/state-and-local-backgrounders/higher-education-expenditures