National Center for Science and
Engineering Statistics

Matching SDR Respondents to Investigators of NSF Awards

Working Papers | NCSES 22-211 | August 12, 2022

Contents

Disclaimer 3

Abstract 4

Introduction 5

Data Sources and Preparation 6

Approach 9

Results 12

Data and Methodological Limitations 16

Use Cases 19

Conclusion and Future Research 25

References 26

Notes 27

Acknowledgments and Suggested Citation 28

Contact 29

Appendix A: NSF Awards Summary Data 30

Appendix B: Supplemental Data of NSF Support in the SEH Doctorate Population 42

Disclaimer

Working papers (WPs) are intended to report exploratory results of research and analysis undertaken by the National Center for Science and Engineering Statistics at the National Science Foundation (NSF). Any opinions, findings, conclusions, or recommendations expressed in this WP do not necessarily reflect the views of NSF. This WP has been released to inform interested parties of ongoing research or activities and to encourage further discussion of the topic.

Abstract

Developing linkages between administrative data and National Center for Science and Engineering Statistics (NCSES) survey data provides new opportunities to examine the impacts of federal funding on the U.S. science and engineering enterprise. This working paper describes a novel effort to match respondents of the NCSES Survey of Doctorate Recipients (SDR) to investigators in the National Science Foundation (NSF) awards database. Specific inputs to the matching process included data from the 2015 SDR—which consisted of 78,320 respondents, representing 920,050 U.S.-residing and 127,800 non-U.S.-residing doctoral scientists and engineers—and roughly 700,000 NSF award-investigator records spanning the last 60 years. Matching was performed sequentially using deterministic methods, primarily relying on survey respondent and award investigator names and e-mail addresses, as well as institution information. The matching process resulted in 7,363 SDR respondents matched to NSF awards, with an average of 4.4 awards matched to each individual who was matched to any award. Data limitations of this matched set, as well as illustrative use examples, are discussed. This proof-of-concept exploration indicates the potential of data linkages to improve the utility of NCSES survey data and to open a wide range of applications for research and program evaluation.

Introduction

The National Center for Science and Engineering Statistics (NCSES) conducts a broad range of surveys on topics, including science and engineering (S&E) education, workforce, research facilities, and research and development funding and expenditures. The results of these surveys inform policymakers, educators, and the public on topics ranging from U.S. competitiveness to science, technology, engineering, and mathematics (STEM) education. As a statistical agency within the National Science Foundation (NSF), NCSES is in a unique position to provide statistical expertise per Title I of the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act), and to offer rich and highly relevant data for studying the effects of federal funding on the U.S. S&E enterprise.

Linking of federal awards, survey, and administrative data has been conducted to assess the effects of federal research funding on workforce, innovation, and productivity and to analyze educational and career outcomes of individuals supported by federal science training and research projects. Recent work in this area of research has suggested that “the richness and complexity of the research enterprise is such that linking data from a variety of sources is likely to be an important way to get new understanding about the multiple facets of scientific activity” (Chang et al. 2019:1487). Large-scale bibliometric linkages have been used to derive insights on the effects of federal research on patenting, including the significant and increasing reliance of U.S. patents on federally supported research (Fleming et al. 2019).

The UMETRICS data set, containing wage and vendor transactions by major U.S. research universities supported by federal awards, has been a particularly fruitful resource for scholarship in this field (Lane et al. 2015), including data linking efforts. Buffington et al. (2016) linked personnel supported by federal research grants in UMETRICS to U.S. Census Bureau earnings records and then utilized the linked data to explore differences in STEM training environments and labor market placement outcomes across gender and household characteristics. More recently, research personnel in UMETRICS have been linked to individuals in the NCSES Survey of Earned Doctorates (SED), a major source of data on U.S. doctoral training and plans for employment. The authors demonstrated how the linked information could increase knowledge of federal funding flows and the importance of federal funding by doctoral field (Chang et al. 2019).

For this project, a matching process was developed to link award and investigator data in the public NSF awards database to restricted-use data in the Survey of Doctorate Recipients (SDR). The aims of this exploratory analysis were to determine the feasibility of linking publicly available NSF administrative data with NCSES survey response data and to assess new avenues for evaluating the impacts of NSF investments. Specifically, three goals were identified, including to (1) explore the potential of utilizing the public NSF awards database and expand knowledge and use of this data set, (2) identify best practices for matching federal award data with other data sets, and (3) enhance the analytical potential of NCSES data for science policy research. As this was exploratory, no inferences should be made with data presented in this paper.

The project described in this working paper establishes that individual-level records can be linked between the two data sources. It identifies a new way to significantly improve the utility of the NCSES survey data and open a wide range of applications for research and program evaluation. This paper outlines the matching process and identifies potential opportunities for future research to utilize the matched set. Ultimately, such data resources can aid NSF in fulfilling its goals under the Evidence Act and in building a culture of evidence-based planning and policymaking.

Data Sources and Preparation

NSF funds research and education in all fields of S&E, through grants, contracts, and cooperative agreements. The foundation accounts for nearly a quarter of federal obligations for basic research performed by U.S. academic institutions (see NCSES report *Federal Funds for Research and Development: Fiscal Years 2019–20*: table 27). To conduct this analysis, several variables were obtained from the NSF awards database and the NCSES SDR.

National Science Foundation Awards Database

The public NSF awards database includes data on all active and expired awards funded by the foundation. The complete database is available in bulk via API and XML downloads. Differences in field availability exist between the API and XML. Specifically, the XML files reflect what is available on the NSF website, with a total of 464,647 award records at the time of access (spring 2020). The NSF awards database classifies awards by award instrument, funding directorate or office and division, and several field and program reference attributes. Summary data on how NSF awards are classified on these dimensions, as well as two new classifications developed for this project, are provided in Appendix A: NSF Awards Summary Data.

In the XML file, two substructures are utilized in this matching project: Investigators, and Institutions. The Investigator block contains the following elements for each investigator on an award as follows:

- First name
- Last name
- E-mail address
- Start date
- End date
- Role code—Principal Investigator (PI), Co-Principal Investigator (co-PI), Former PI, or Former Co-PI.

All NSF awards have an assigned PI and may have one or more co-PIs. Awards may be transferred to new PIs (or co-PIs), in which case the existing PI in the record will be reclassified into the role code former PI, and the new PI added to the record. Each PI and co-PI had an associated institution.

The Institution block contained the following elements for each awardee institution:

- Name
- City
- ZIP code
- Phone number
- Street address
- Country
- State name
- State abbreviation

The resulting data set included 700,488 records in which each award may have had multiple rows—one for each PI or co-PI that was extracted from the XML files. Fifty-two awards with malformed XMLs were removed, as were 82 duplicate records that matched on all the above data elements, leaving a total of 700,354 unique records.

Data Cleaning

The NSF awards database required extensive cleaning prior to matching. A small number of award records included the name of a PI twice—one record marked as PI, and another record marked as former PI (about 4%). Therefore, it was necessary for matching to create a unique list of award number, first name, last name, and e-mail address field to be used for matching.

After removing the role code, start date, and end date columns, deduplication was performed on the award, name, and e-mail entries. This resulted in 695,261 award-investigator pairs representing 455,552 unique awards.

The NSF awards database has coverage of awards dating back to 1959. Data quality issues plagued some of the pre-1973 award dates. A combination of manual coding based on the award number, the "minimum letter amended date" field, and "award effective date" was used to determine award year. For example, award 7723272 is listed with an award effective date of 10 January 1917, so its initial amendment date of 14 December 1977 was used instead. Thirteen remaining awards were assigned a start date based on the first two letters in the award number. For the vast majority of the data, the year value in the "award effective date" field was used.

A full e-mail address (meaning at least X@Y.Z) was used for 85.5% of records, 14.4% had no e-mail information, and very few (0.1%) had partial information or server information. However, noise in the e-mail address field necessitated extensive cleaning. For example, some award-investigator pairs had two e-mail addresses in the e-mail address field for one award, while other award-investigator pairs had multiple entries to account for multiple e-mail addresses. Therefore, observations with multiple e-mail addresses in the e-mail address field were split into as many rows as e-mails, retaining the same award-investigator ID. In addition, the duplicate award and name entries were examined and assigned consistent award-investigator IDs so that the same person on the award was always assigned the same ID, even if multiple e-mail addresses (and, hence, observations) existed. The resulting data set was deduplicated, leaving 695,320 entries (because some award-investigator pairs have multiple e-mail addresses, they have multiple rows) representing 695,190 unique award-PI pairs and 455,552 unique awards.

Further cleaning was required on a small subset of records. Because some awards date back to before the adoption of modern e-mail, some e-mail address fields were bitnet or omninet addresses. Where possible with bitnet, the username and server formation were retained for use in matching. Some records were assigned a nonpersonal e-mail address, which is not useful for the person-person matching performed here; however, the server information was still retained for matching purposes. Other records required removal of spaces or unallowed special characters, as well as changing of erroneous commas into periods. These steps only applied to a small number of records, with over 99% requiring no cleaning.

A new derived data field, called "e-mail stem," was created to match the NSF investigator e-mail field with the SDR respondent e-mail field. An email stem defined here refers to the organization plus domain type portion of the e-mail. E-mail stems were created using the value before the "@" and the last two segments of the text after the "@." E-mails were stemmed manually using Excel functions, but this can also be done algorithmically.

Finally, special non-English language characters in investigator names were replaced with their English equivalents for standardization. Beyond this step, names were not cleaned.

Survey of Doctorate Recipients

The SDR provides demographic, education, and career history information from individuals with a U.S. research doctoral degree in a science, engineering, or health (SEH) field. The SDR is sponsored by NCSES and the National Institutes of Health (NIH). Conducted since 1973, the SDR is a unique source of information about the educational and occupational achievements and career movement of U.S.-trained doctoral scientists and engineers in the United States and abroad.

The SDR restricted-use data reside on a secure data enclave, accessed through secure credentials. The personally identifiable information (PII) used in this project is only available to a very small number of individuals and is only accessed through contractual mechanisms and a restricted-use data license. This project used the results from the 2015 SDR, which consists of 78,320 respondents, representing 920,050 U.S.-residing and 127,800 non-U.S.-residing doctoral scientists and engineers that meet the following criteria:

- Earned an SEH research doctorate degree from a U.S. academic institution prior to 1 July 2013.
- Are not institutionalized or terminally ill on 1 February 2015.
- Are less than 76 years of age as of 1 February 2015.

Most SDR records (REFID) contain multiple e-mail address fields. The SDR data set was manipulated to create a new row for each e-mail address listed within a particular record. To do this, an e-mail matching list was created by flattening this structure to give each e-mail address its own observation associated with the REFID. Not every SDR respondent had an entry for e-mail; therefore, this action resulted in a data set with 135,493 rows representing 73,985 unique REFIDs.

Data Cleaning

Only minor special character replacement was necessary on the SDR data set. E-mail addresses were stemmed in the same way as the e-mail addresses in the NSF awards data set. E-mail addresses that were found to be erroneous during quality assurance work were recorded in a separate file, and the e-mail addresses were removed by the code during matching. Additionally, all data were turned to uppercase.

Approach

Matching was performed sequentially using deterministic methods in the following steps. The outcome of the matching process is a list of NSF awards linked to an SDR respondent (REFID). At the end of each matching step, the matched award-investigator IDs were used to subset the universe of potential matches, and the subsequent matching algorithms were applied to that subset of unmatched award-PI pairs. Probabilistic matching, such as fuzzy matching, was explored, but true differences in e-mails can be one letter. Names in both data sources were relatively clean, and experiments with fuzzy matching resulted in significantly more false positives than true positives. A list of the elements used to match between the data sets is shown below in table 1.

TABLE 1

Data elements used for matching

(Variable)

Data source	Variable name	Description	Match step used in
Person information			
SDR	E-mail (up to 15 values per respondent)	E-mail addresses associated with the respondent; the data quality is unknown[a]	1, 2, 3, 11
SDR	DRFLNAME	Last name of respondent recorded in the doctoral records file (i.e., when the respondent received the PhD)	All
SDR	LNAME15	Last name of respondent recorded in the 2015 survey file	All
SDR	DRFFNAME	First name of respondent recorded in the doctoral records file (i.e., when the respondent received the PhD)	2 - 9
SDR	FNAME15	First name of respondent recorded in the 2015 survey file	2 - 9
NSF grants database	E-mailAddress	E-mail address of investigator	1, 2, 3, 11
NSF grants database	LastName	Last name of investigator	All
NSF grants database	FirstName	First name of investigator	2 - 11
Location information			
SDR	EMZIP	Employer ZIP code	4
NSF grants database	ZIP5	Institution ZIP code	4, 5, 6
SDR	ZIP15	Individual ZIP code in 2015	5
SDR	DRFZIP	Individual ZIP code at time of PhD	6
SDR	EMCITY	Employer city	7
NSF grants database	CityName	Institution city	7, 8, 9
SDR	CITY15	Individual city in 2015	8
NSF grants database	StateCode	Institution state	8
SDR	STATE15	Individual state in 2015	8
SDR	DRFadd	Individual city at time of PhD	9
SDR	EMNAME	Employer name	10
NSF	NAME	Institution name	10, 11

NSF = National Science Foundation; SDR = Survey of Doctorate recipients.

[a] Initially, it was assumed that these e-mail addresses were the e-mail addresses the survey contractor used to communicate with the respondent, making it a very high-quality data element. However, during the matching process, it was observed that some e-mail addresses were clearly for other individuals besides the respondent. These e-mails may be connected to the alternate contacts in the record or co-principal investigators on grants.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Match Step

Match Step 1: Whole E-mails and Last Name

The first step in the matching process compared cleaned e-mail addresses in the SDR and NSF data sets. A direct match was considered to be a true positive, as it is highly unlikely that two separate individuals in this small subset would have the exact same e-mail address over time. In the process of this first match step, incorrect e-mails were identified in both the SDR and NSF data sets through manual review. These e-mails were manually removed to prevent inaccurate matches.

The presence of incorrect e-mails indicated that e-mail address itself cannot be a sole link between two records. Therefore, an exact match on the last name was also required. The remaining 10 steps in the matching process utilize both the Doctorate Records File (DRF) last name and the 2015 SDR last name. These fields were cycled through each step that used last name as a matching variable. The same process was undertaken for first names in the steps that utilize first names.

Match Step 2: Stemmed E-mails, First Name, Last Name

E-mail addresses were reduced to their stems and matched. A last name match and first name match were also required.

Matching on first names required additional actions. The SDR respondent data contain both the first name(s) and middle name(s) or initial(s) in the field. It cannot be assumed that two names in the field are first or middle names because some individuals have hyphenated names or two first names. Therefore, each first name match was performed first using the text before the space in the first name column and then matched using the whole text of the field, including spaces.

Match Step 3: First Name, Last Name, and E-mail Server

The e-mail institution information (e.g., sri.com) can provide high-quality information about an individual’s institution and does not suffer from the variety of ways an institution can be listed. The following high-frequency e-mail servers were filtered out: gmail.com, hotmail.com, yahoo.com, aol.com, verizon.net, and comcast.net.

Match Step 4: First Name, Last Name, and Employer ZIP to Award Institution ZIP

This step performed an exact match on the first name, last name, and the SDR employer ZIP to the award institution ZIP.

Match Step 5: First Name, Last Name, and Respondent ZIP to Award Institution ZIP

This step performed an exact match on the first name, last name, and the SDR respondent ZIP to the award institution ZIP.

Match Step 6: First Name, Last Name, and Respondent DRF ZIP to Award Institution ZIP

This step is the same as step 5 but also used respondents’ DRF ZIP.

Match Step 7: First Name, Last Name, and Employer City to Award Institution City

This step performed an exact match on the first name, last name, and the SDR respondent employer city and the award institution city.

Match Step 8: First Name, Last Name, and Respondent City to Award Institution City and State

This step performed an exact match on the first name, last name, and the state and city of the SDR respondent and the grant institution.

Match Step 9: First Name, Last Name, and Respondent DRF City to Award Institution City

This step performed an exact match on the first name, last name, and the SDR respondent DRF city and the award institution city.

Match Step 10: First Name, Last Name, and Employer to Award Institution

This step performed an exact match on the first name, last name, and the SDR respondent employer to the award institution.

Match Step 11: First Name, Last Name, and Respondent E-mail Institution to Award Institution

The SDR PII record contains historical e-mail addresses but not historical employers. It is likely that a subset of respondents obtained NSF awards while employed at other institutions. Therefore, these historical e-mail addresses provide information about past employers. To create an “after the @” (such as sri.com) to institution name (such as SRI International) crosswalk, institution names and investigator addresses from the NSF awards database were extracted and manually cleaned and filtered to 2,564 high-quality links between e-mail servers and institutions. This information was then appended to the SDR records using the “after the @” part of the e-mail stem. This final match step performed an exact match on the first name, last name, and the SDR appended institution affiliation name and the award institution name.

Quality Assurance

A review was completed for the underlying code used to conduct this analysis as well as an examination of the matched file in the restricted data environment. No issues were identified in the code. It was confirmed that the award-investigator pairs matched to REFIDs were appropriately flagged by match step, ensuring the number of matches reported in the results accurately reflect the matching steps that are described.

Results

Collated Matching Results by Match Step

Table 2 shows the collated results of the matching process by match step. Seven percent of unique awards in the NSF awards database were matched to an SDR respondent, and 9% of the 2015 SDR respondents were matched to an NSF award. The vast majority of matches occurred in the initial e-mail-based portion (steps 1 and 2) of the process. The presence of e-mail addresses increased the likelihood of a match. Only 1% of respondents without at least one e-mail address were matched to an NSF award, compared to 11% of respondents with e-mail addresses.

TABLE 2

Collated matching results

(Number)

Match step	Number of REFID-award matches	Unique REFIDs (per step)	Number of unique awards matched	Filters
1	29,067	6,327	28,110	None
2	1,601	348	1,596	None
3	385	162	385	Name commonality
4	421	208	421	Name commonality
5	115	69	115	Name commonality
6	84	70	84	Name commonality
7	134	86	134	Name commonality
8	133	79	133	Name commonality
9	65	52	65	Name commonality
10, 11	46	31	46	Name commonality
Unmatched	663,142			

NSF = National Science Foundation.

Note(s):
Match steps 10 and 11 are combined for disclosure purposes.

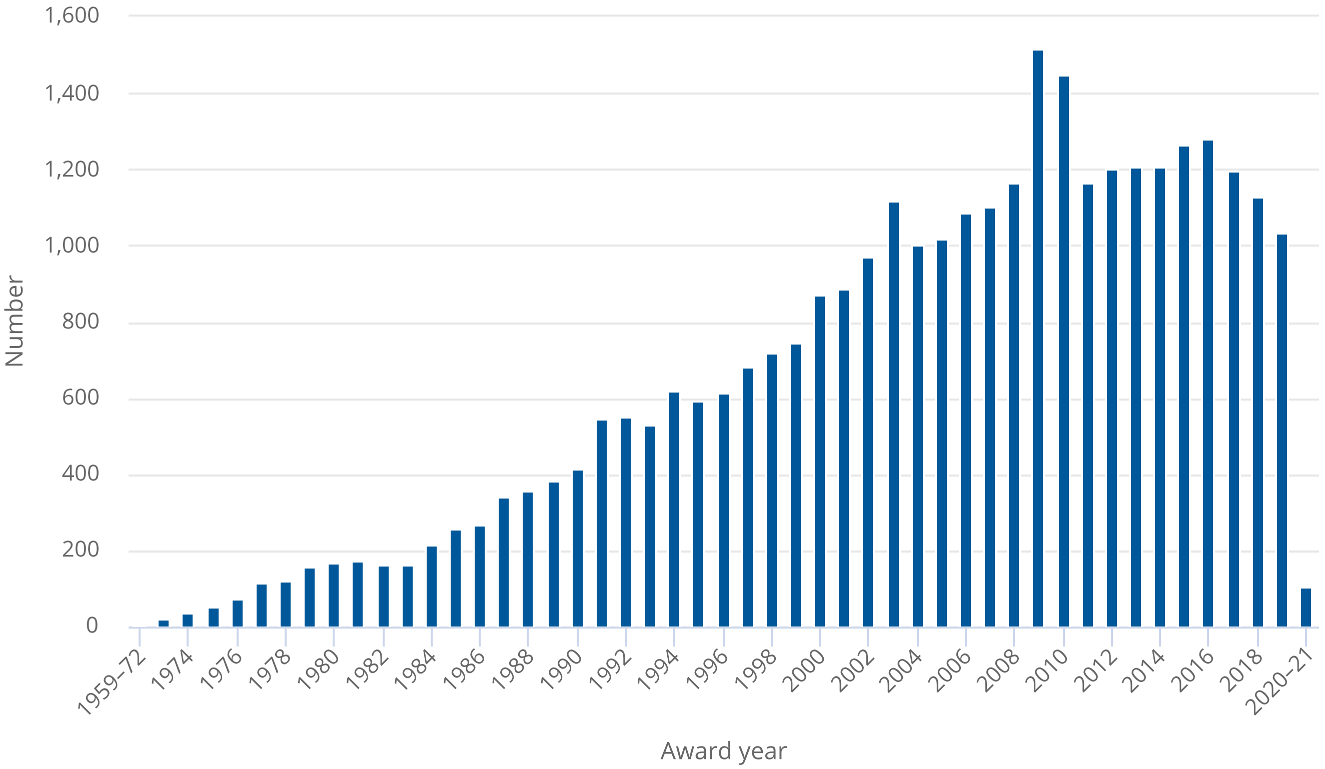
Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Distribution of Matched NSF Award-Investigator Pairs by Award Year

Figure 1 shows the number of matched award-investigator pairs by award year. The vast majority of matched awards are from the last 30 years. This distribution reflects the general increase in the number of grants but also can be attributed to data quality related to e-mail addresses.

FIGURE 1

Distribution of matched NSF award-investigator pairs, by year: 1959–2021



NSF = National Science Foundation.

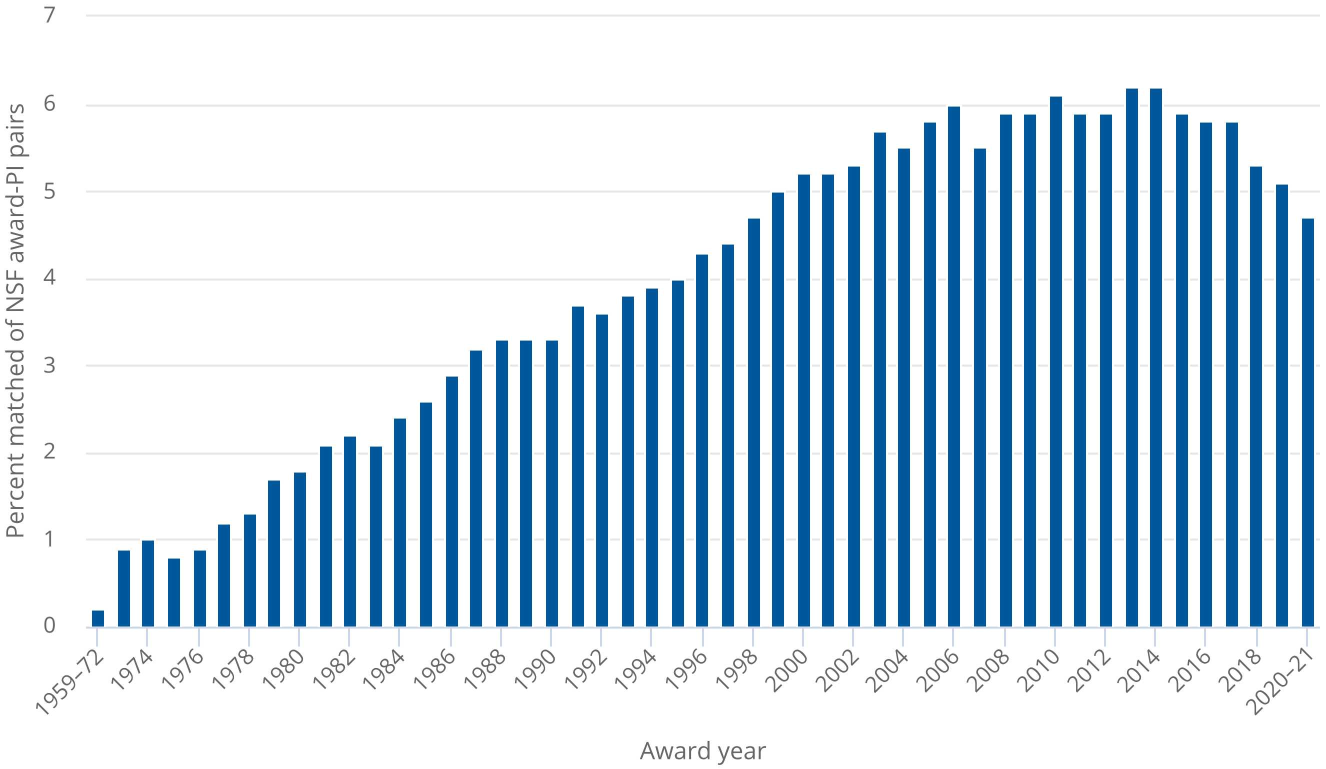
Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Percent of Records Matched by Award Year

The match rate by year ranged from a low of 0% in early years of the NSF awards data to a high of 6.2% of award-investigators pairs in 2013 and 2014 (figure 2). Overall, 4.6% of the award-investigators pairs were matched to an SDR respondent.

FIGURE 2

Records matched to SDR respondents, by award year: 1959–2021



NSF = National Science Foundation; PI = principal investigator; SDR = Survey of Doctorate Recipients.

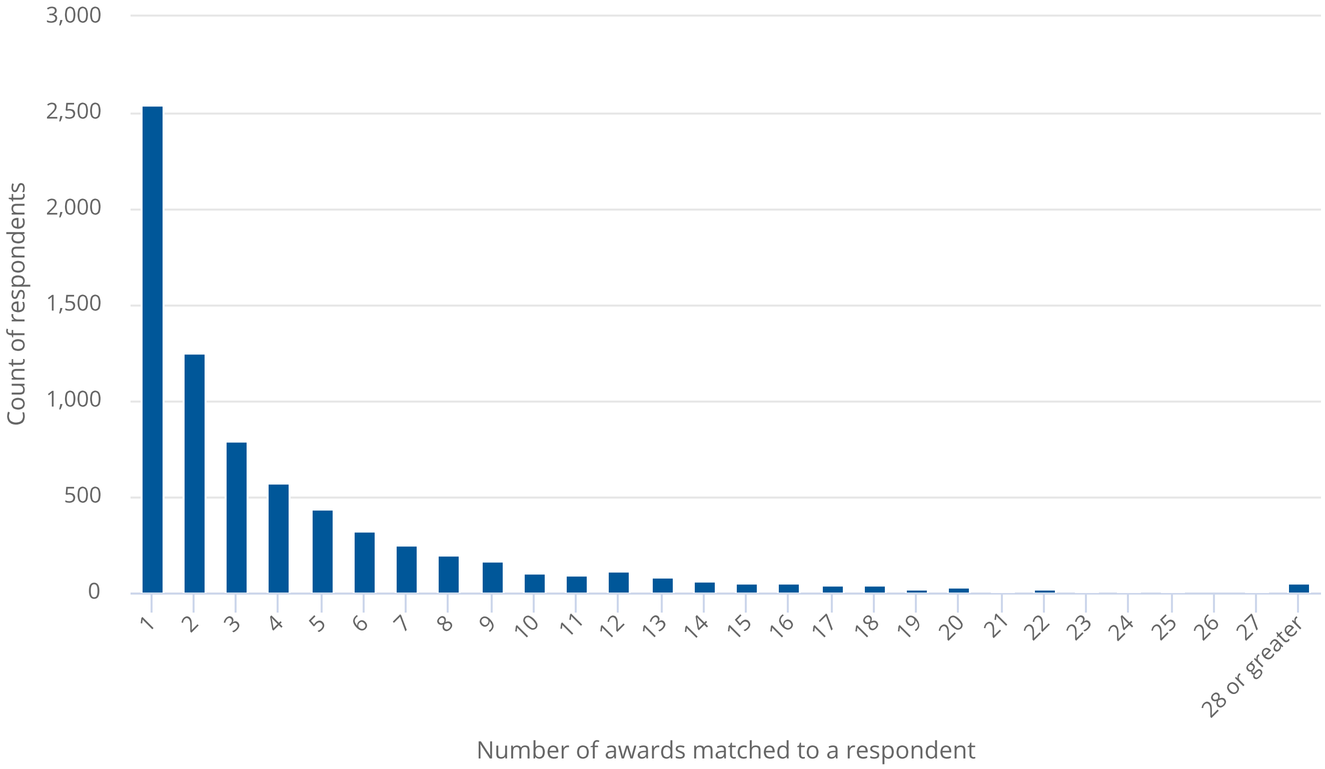
Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Distribution of Awards Matched to SDR Respondents

A total of 7,363 SDR respondents were matched to awards in the public NSF awards database. Figure 3 shows the distribution of the number of awards matched to individuals. Because of disclosure requirements, the exact number of awards matched in the tail of the distribution is not provided. These individuals were rolled up to 28 awards or more. An average of 4.4 awards were matched to each individual that was matched to any awards, though the median count is 2.0 awards.

FIGURE 3

Distribution of NSF awards matched to SDR respondents: 2015



NSF = National Science Foundation; SDR = Survey of Doctorate Recipients.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

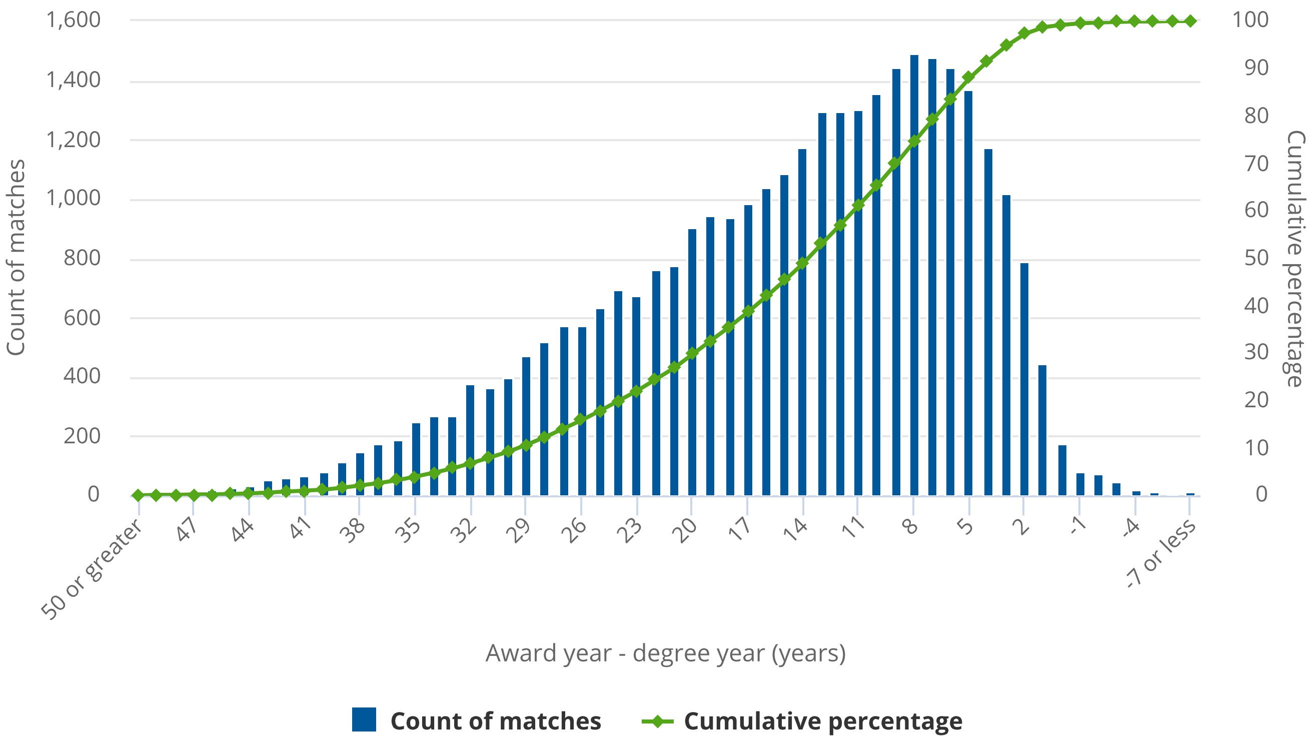
Data and Methodological Limitations

This project was a proof-of-concept exploration of matching. Although matching was performed deterministically, match quality was not determined via a gold standard set to calculate precision and recall. However, some factors were considered as checks of match quality.

The difference between birth year and award year was examined and factored into the name commonality filter previously discussed that was applied to the initial results. The difference between the award start year and the matched SDR respondent's doctorate year was also investigated. As shown in figure 4, the vast majority of the matches (99%) had a difference between the award year and doctorate calendar year of 0 or greater, meaning that the individual received an NSF award no earlier than the same year he or she received the doctorate. An ad hoc review of the negative values shown in the figure indicated that some individuals receive NSF awards prior to receiving their doctorate. A very small number of matches had award years more than 5 years prior to their degree years, which is likely the result of these individuals obtaining multiple doctorates.

FIGURE 4

Difference between award year and doctorate year



SDR = Survey of Doctorate Recipients.

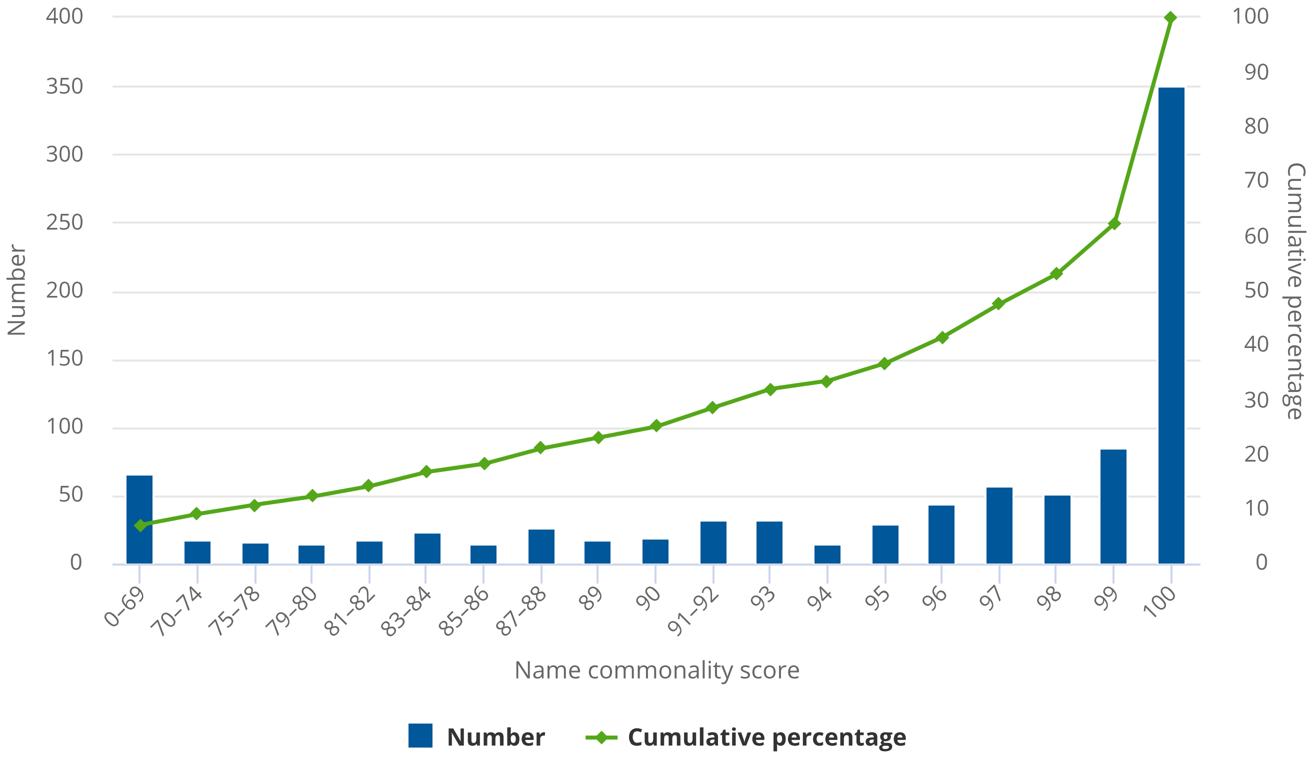
Note(s):
Data in this figure are calculated from 32,051 total matches.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Analysis of initial results revealed a subpopulation that appeared to be receiving NSF awards in their 20s or earlier. These anomalies were concentrated where names were more common, as indicated by a name commonality score. The name commonality score is defined here as the natural log of the frequency of a last name contained in the NSF awards database divided by the count of all names. This score was then divided into 100 levels, with higher levels for more common names. After initial matching was performed, potentially problematic matches were reviewed if the difference between the award year and the birth year was less than 20 or greater than 66. The number of these pairs is plotted in figure 5 as a function of the name commonality score.

FIGURE 5

Distribution of name commonality scores for anomalous award-investigator matches



Note(s):

Anomalous matches are defined as matched pairs where the difference between the year of award and birth year of investigator was less than 20 years or greater than 66 years. Bars represent the count of name commonality among problematic matched pairs. Dots represent the cumulative percentage of problematic matches, showing that 90% of this set are excluded from the results if a name commonality score threshold of 77 is used. Because of disclosure issues, some name commonality scores are binned.

Source(s):

National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

As shown in the figure, many of the problematic matches have a name commonality score of above 70. A name commonality threshold of 77 was chosen to remove 90% of the problematic matches. This filter was applied to all matching steps except in steps where an e-mail address was used, as an e-mail match is of high quality. Results were ultimately filtered to those individuals with a difference between their NSF award year and birth year of greater than 20 (three false-positive matches were filtered out at this step).

The NSF public awards database contains information on PIs or co-PIs at the moment the data are collected. Similarly, data of the SDR respondents reflect employment information reported for the 2015 survey reference period. Without having complete longitudinal affiliation data, the matching algorithm will miss cases where a respondent obtains an NSF award at an institution different from his or her PhD institution or his or her employer institution at time of response to the SDR. This recall error may vary by career stage because level of mobility of research doctorate holders can change over time. For example, early career doctorate holders and those taking postdoctoral researcher (postdoc) positions are likely to experience job changes more often than tenured professors.

Use Cases

Here, we provide two use cases of the matched data set produced for this project. First, the NSF administrative data on awardees are used as a benchmark for survey data quality—specifically, reporting of NSF support by SDR respondents. Second, the linked data set is used to make inferences on patterns of NSF support among the larger SEH doctorate holder population (with the caveats stated in the previous section).

Comparison of Survey and Administrative Data on NSF Support

The matching procedure utilized data elements in the SDR and NSF awards database to match SDR respondents to NSF investigators. SDR respondents' reporting of NSF support for their research in the survey allows for some comparisons between the survey data and the administrative data.[1] This comparison can be used to generate reference precision and recall values for the matched data set.

The SDR asks respondents if they had been supported by NSF funding in the previous year. The reference year for this data set is 2015, and the standard NSF award is 3 years; therefore, awards that started in 2012 could reasonably be expected to be supporting the respondent.

Precision

Between 78% and 81% of individuals matched to NSF awards in each year from 2012 to 2014 reported NSF support of their work on the 2015 SDR. An ad hoc review of matched individuals who answered that they did not receive federal support for their work or did not select NSF as a specific federal source did not show a conclusive pattern to explain the discrepancy.[2] The reviewed matches are correct matches, and the NSF projects ranged from research work to scholarship programs.

There are many plausible reasons why the precision is not higher; without cognitive testing of how individuals interpret this survey question, however, it is difficult to draw any specific conclusions. Individuals can be named as the PI on grants but may not work on the project on a level for the respondent to consider their work to "be supported." This may be more common for award types focused on training and education of graduate and undergraduate students. For example, 80% of individuals matched to NSF awards classified under the research funding program reported NSF support on the SDR, but only 70% of those matched to training awards did so. (See appendix table A-7 for more details on funding program categories.) In these cases, PIs may be professors, and their students and research staff would be supported by the grant. It is not clear how professors would answer this question on the SDR in cases that they are the PI on a grant but do not charge time to the grant.

Correspondence between the survey and administrative data may be increased if survey respondents with imputed values on the SDR question of NSF support are excluded from the comparison. Approximately 10% of SDR respondents' values for NSF support are imputed; among those individuals matched to investigators of NSF awards from 2012 to 2014, the imputation rate on this question is 9%. Limiting the comparison to only individuals without imputed values results in slightly higher precision: 83% of respondents matched to NSF awards in this 3-year period reported NSF support on the survey. Including all individuals regardless of imputation, 79% of those who were matched to NSF awards reported NSF support on the SDR. For the relatively small number of individuals with missing data who were also matched to an NSF award, 33% were imputed as having received NSF support.

Recall

For the reference year 2015, there were 4,915 individuals who responded that they had received NSF support for their work. Within this group, 59% were matched to an investigator on an NSF award, and 41% were not. These low values are not surprising because many researchers work on awards on which they are not the PI or the co-PI. In addition, there is some imputation in the data set for this flag.

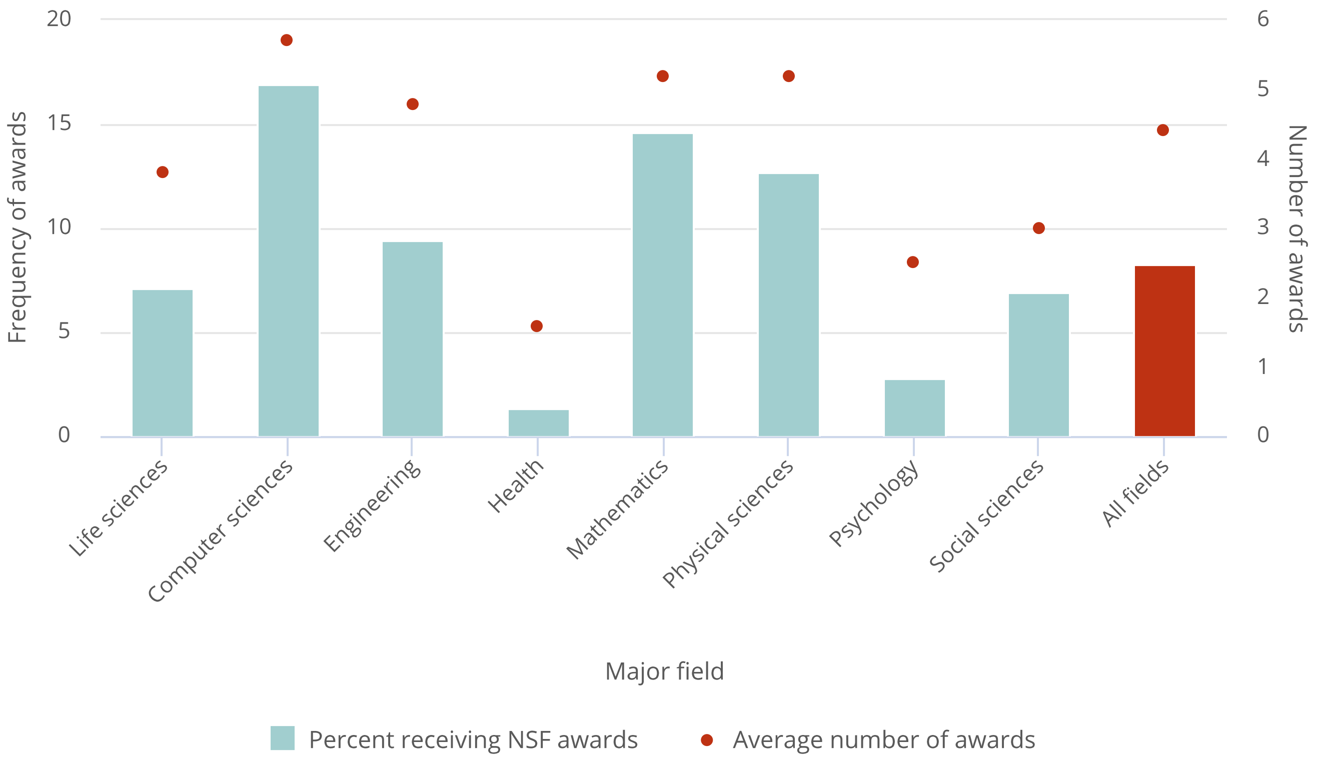
Inspection of the academic job classification of those that reported NSF support but were not matched to an award revealed that teaching faculty were most likely to be matched to at least one award (75% of those who indicated NSF support were matched). Those who indicated they are a research assistant, teaching assistant, or postdoc had a much lower match rate—only 18% of those individuals were matched.

Patterns of NSF Support in the SEH Doctorate Holder Population

This project used data from the 2015 SDR, which consists of 78,320 respondents, representing 920,050 U.S.-residing and 127,800 non-U.S.-residing doctoral scientists and engineers. By examining attributes of the 7,363 SDR respondents matched to investigators of NSF awards, inferences can be made on patterns of NSF support among the SDR target population of SEH doctorate holders. The estimates below are produced using the final survey-specific weights. Note that references to receipt of NSF support in this section include only those reported as PIs and co-PIs in the NSF awards database. Estimates presented in figure 6 through figure 9 and in table 3 are expected to underestimate the true population characteristics, such as the share of doctorate holders receiving NSF awards and the average number of NSF awards received, largely due to the recall errors discussed in Data and Methodological Limitations. However, we believe the patterns of NSF support in relative scales across the major doctorate field of study remain informative. We present the analysis below as illustrative examples of potential utility of this type of research data and hope these examples motivate future research on improving the source data for matching and for the matching methodology.

FIGURE 6

NSF awards in the SEH doctorate population, by major field of first doctorate



NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

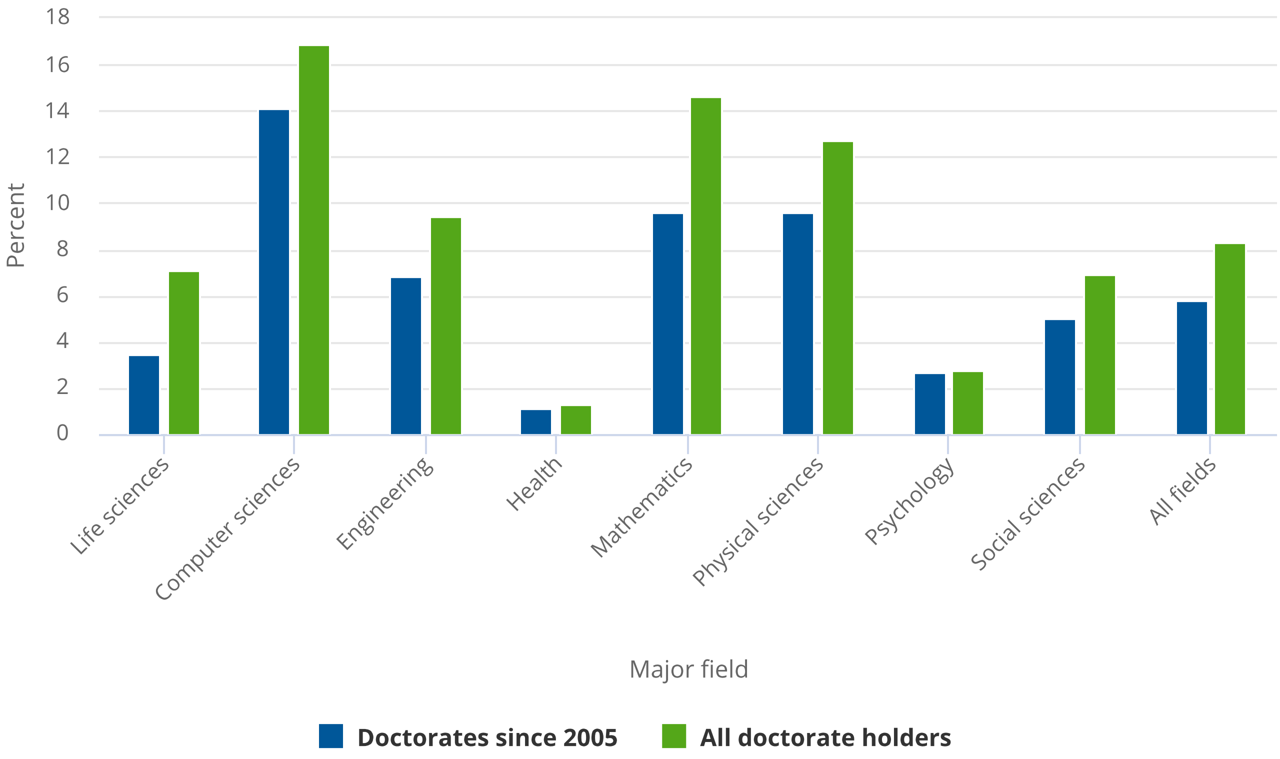
NSF Support by Doctorate Field of Study

An estimated 8.3% of SEH doctorate holders have received NSF awards as PIs and co-PIs, for whom the average award count is 4.4. As indicated in figure 6, levels of NSF support vary significantly by major field of doctorate, with individuals with computer and information science doctorates having the largest estimated share receiving an award (16.9%) and the highest average number of awards (5.7). In addition to computer sciences, mathematics and statistics and physical sciences doctorates also had higher than average levels of NSF support. Engineering doctorate holders experienced higher rates of NSF support than those in the social sciences (9.4% vs. 6.9%), as well as higher average award counts (4.8 vs. 3.0). Individuals with doctorates in psychology and health had much lower than average levels of NSF support, with those in health fields having the lowest rate of support (1.3%) and average number of awards (1.6). The particularly low level of NSF support among health doctorate holders is plausible, given the primary role of NIH rather than NSF in funding research in this field (see NCSES report *Federal Funds for Research and Development: Fiscal Years 2019–20*: table 23). Standard errors for estimated shares receiving NSF awards by doctorate field of study are provided in appendix table B-1.

Limiting the analysis to SDR respondents receiving doctorates in more recent years may provide more timely measures of NSF support. When only considering individuals receiving doctorates in 2005 or later, a few key observations are made. As expected, both the share of doctorate holders receiving awards (from 8.3% to 5.8%) and average number of awards (from 4.4 to 2.5) are lower. These two measures decline for all major fields of doctorate. Although the gap between fields is reduced, the same overall pattern by field is observed. Namely, the estimated shares of doctorate holders in computer and information sciences, mathematics and statistics, and physical sciences are highest, and shares for psychology and health are lowest. Rates of NSF support for doctorate holders in other fields are closer to the average of 5.8% (appendix table B-1).

FIGURE 7

NSF awards in the SEH doctorate population, by doctorate cohort and major field of first doctorate



NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

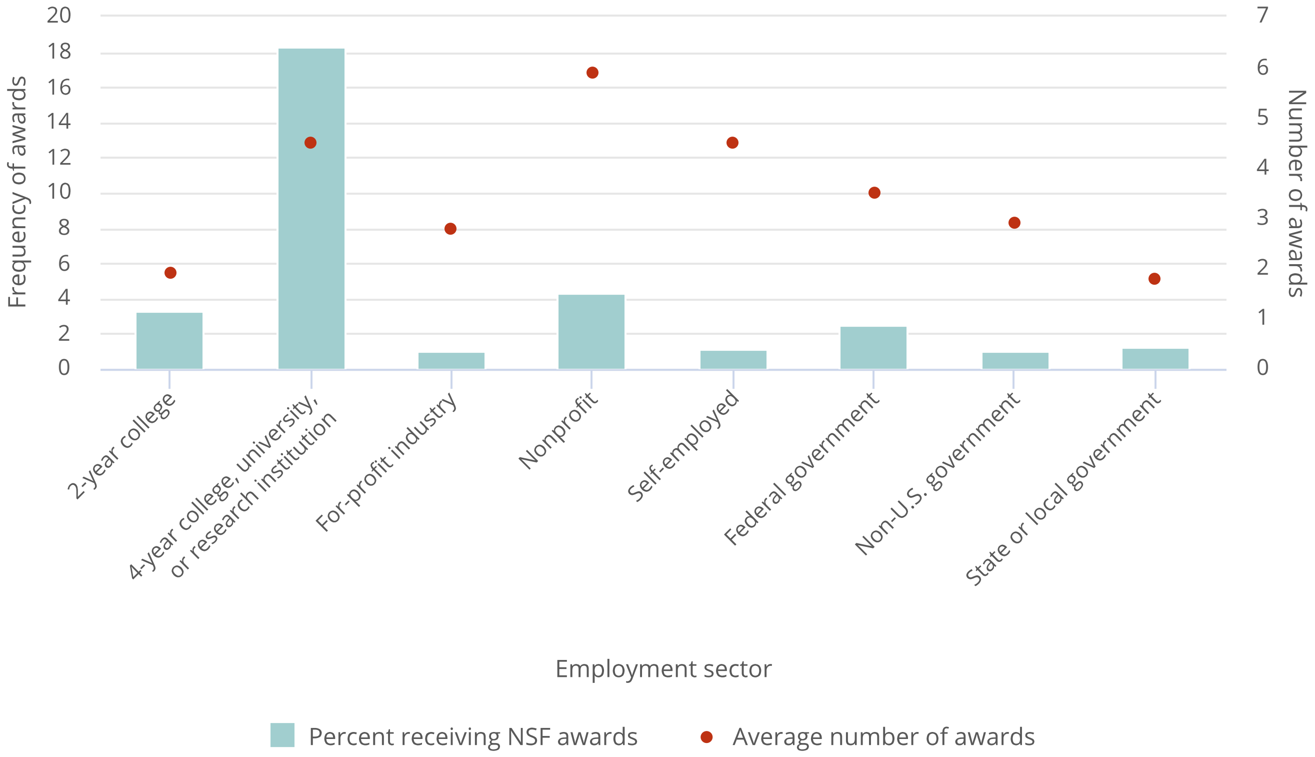
NSF Support by Employment Sector

Levels of NSF support have a strong relationship to doctorate holders' employment sector (figure 8; appendix table B-2).[3] As expected, individuals employed by a 4-year college, university, or university-affiliated research institution are vastly more likely to receive NSF support (18.3%) than those employed in any other sector. The sector with the next highest share, private nonprofit organizations, was less than a quarter (4.3%) of this level. Self-employed individuals, and those working in the for-profit industry and in state, local, or non-U.S. governments had the lowest rates of NSF support, with shares between 1.0% and 1.3%.

Notably, the share of individuals in a sector receiving NSF support was not predictive of the average award count of those that did receive an NSF award in that sector. For example, average award count per matched respondent was higher for those employed in the nonprofit sector (5.9) than for those employed by 4-year colleges (4.5), despite the latter being more likely to receive an NSF award. Likewise, the average number of NSF awards for self-employed individuals receiving NSF support (4.5) was far greater than the average for those employed in state or local government (1.8), even though the share receiving support was very similar for these sectors.

FIGURE 8

NSF awards in the SEH doctorate population, by employment sector



NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

NSF Support by Demographic Characteristics

The estimated share of doctorate holders who have served as PIs and co-PIs on NSF awards differs by demographic characteristics (table 3 and appendix table B-3). Men were more likely (9.3%) than women (6.2%) to receive NSF support overall and in the 4-year colleges and universities employment sector (21.0% vs. 13.1%). Within the university sector, white men were the specific group with highest shares receiving an NSF award (24.5%). The estimated share for individuals of Other race (inclusive of American Indian or Alaska Native, Native Hawaiian or Pacific Islander, and multiple race),[4] was 9.4% across all sectors and 19.8% for the university sector, roughly similar to the shares (8.3% and 18.3%, respectively) for the overall doctorate population. However, Asians had the lowest rates of NSF support, particularly Asian females, with 3.6% across all sectors and 7.9% in 4-year colleges and universities receiving NSF support.

The share of individuals receiving NSF support is estimated to be much higher (at least double) for those employed by 4-year colleges and universities for nearly every sex, race, or ethnicity group. As shown in Figure 9, in this sector men had higher rates of NSF support than women across all races.

TABLE 3

NSF awards in the SEH doctorate population, by employment sector, sex, race, and ethnicity

(Number and percent)

	All doctorate holders			Employed by 4-year educational institution		
Sex, race, and ethnicity	NSF awardees	Total	Share (%)	NSF awardees	Total	Share (%)
Both sexes	86,850	1,047,900	8.3	73,150	399,800	18.3
Asian	13,600	258,800	5.3	12,000	97,400	12.3
Black	2,150	33,950	6.4	1,900	14,250	13.4
Hispanic	3,650	48,000	7.6	3,150	22,400	14.1
White	66,100	693,150	9.5	54,950	260,150	21.1
Other	1,300	14,000	9.4	1,100	5,550	19.8
Female	21,050	341,400	6.2	18,050	137,550	13.1
Asian	2,550	72,000	3.6	2,250	28,150	7.9
Black	700	15,150	4.7	600	6,200	9.9
Hispanic	1,100	17,500	6.3	950	8,050	11.8
White	16,250	230,900	7.0	13,900	92,800	15.0
Other	450	5,800	7.4	350	2,350	15.1
Male	65,750	706,500	9.3	55,150	262,300	21.0
Asian	11,050	186,800	5.9	9,800	69,300	14.1
Black	1,450	18,750	7.8	1,300	8,100	16.1
Hispanic	2,550	30,500	8.4	2,200	14,350	15.4
White	49,850	462,200	10.8	41,100	167,400	24.5
Other	900	8,250	10.9	750	3,200	23.3

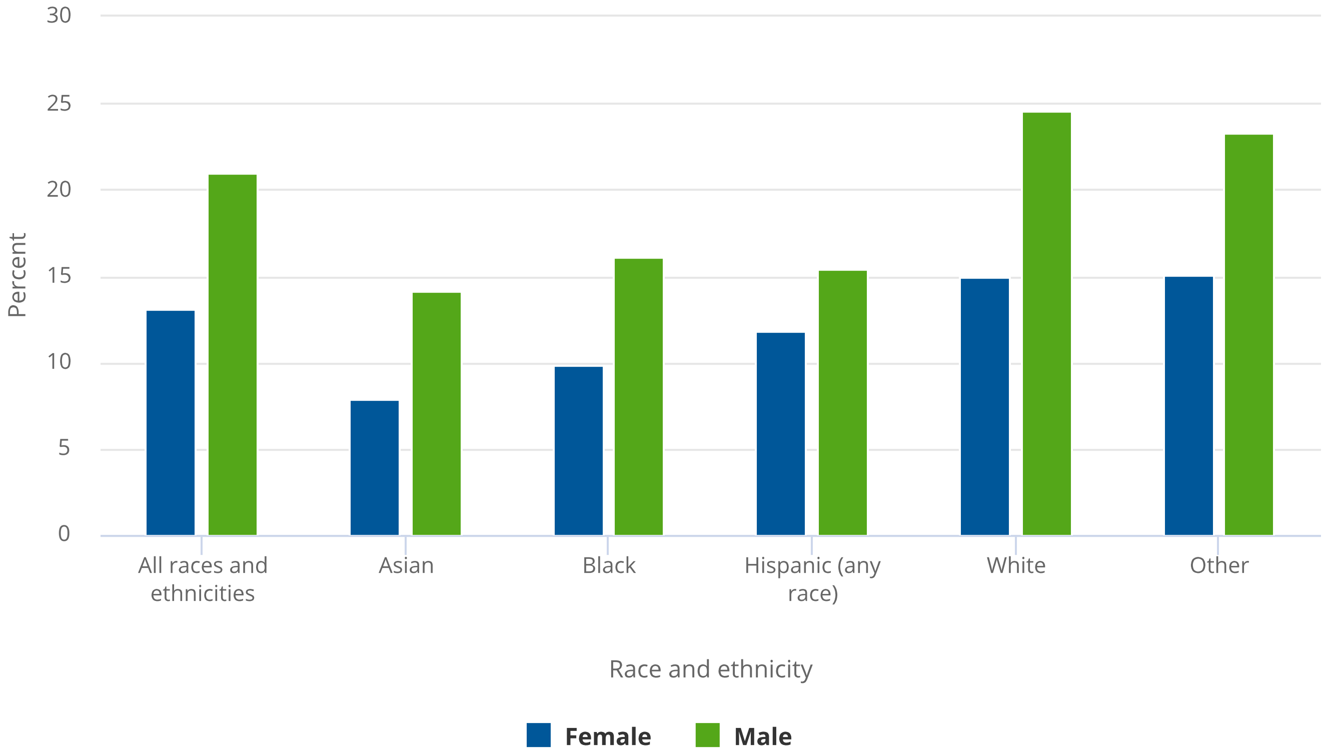
NSF = National Science Foundation; SEH = science, engineering, and health.

Note(s):
Other includes American Indian or Alaska Native, Native Hawaiian or Pacific Islander, and multiple race. Hispanic may be any race; Asian, Black, and White exclude Hispanic origin. The 4-year educational institution sector includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), and university-affiliated research institutions.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

FIGURE 9

Share of SEH doctorate holders employed by 4-year educational institutions receiving NSF support, by sex, race, and ethnicity



NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

Conclusion and Future Research

This project was a proof-of-concept exploration that matching can be executed to append the public NSF awards information to the SDR. In total, 7,363 SDR respondents were matched to awards in the public NSF awards database, thus expanding the analytical potential and range of applications for the NCSES survey data.

To recall, three goals were articulated for this project: (1) explore the potential of utilizing the public NSF awards database and expand knowledge of this data set, (2) identify best practices for matching federal award data with other data sets, and (3) enhance the analytical potential of NCSES data for science policy research. Extensive cleaning was carried out on the NSF data, including investigator e-mail addresses, which proved to be a vital linkage point to the survey data. Next, a reproducible matching procedure was documented and performed, establishing that individual-level records can be linked between the two data sources. Finally, use cases were provided to demonstrate the utility of the matched data in assessing patterns of NSF support in the doctorate population.

By utilizing this novel approach, researchers could replicate the steps outlined in this working paper to explore other linkages between NCSES surveys and awards data. As this was an initial exploration into the possibilities, we manually cleaned much of the data instead of automating the process. An opportunity exists for researchers to automate this process to replicate results faster—specifically, with regard to e-mail stemming and character removal.

Further opportunities exist to expand the usability of the NSF program reference and element codes. Creating a standard ontology applied via a crosswalk on research award data could open new possibilities to explore how research funding trends have changed over time and are distributed across institutions. More research might identify opportunities to incorporate the new grantee type field could also be explored and further refined. Our research team developed nine categories; however, further refining these groups, such as separating public and private schools, could shine light on how research awards are being dispersed. Finally, additional analysis could focus on different segments of the surveyed populations, such as undergraduate researchers supported by federal awards.

New possibilities to build on the matching procedures outlined in this working paper exist, and a few use cases are listed in this section to provoke further research questions and opportunities. Specifically, this matching procedure might be applied to other databases. As noted in past research and through this paper, it is possible to find novel information from data sets that do not individually give the same level of depth. When combining multiple data sources, richer information can be derived to advance our understanding of federal funding, doctoral candidates, and the state of research grants in the country. A potential use case could be to explore how the matching procedure could be applied to NIH's Research Portfolio Online Reporting Tools (RePORT) Expenditures and Results (RePORTER) database. The RePORTER database allows for users to explore all past and present grants funded through NIH. Much like the NSF awards database, RePORTER includes PI names. Some awards also include the e-mail information for the PI. Potentially connecting SDR data with this data set could enrich NIH's knowledge of the individuals receiving funding through the agency.

References

Buffington C, Cerf B, Jones C, Weinberg B. 2016. STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census. *American Economic Review* 106(5):333–38.

Chang W-Y, Cheng W, Lane J, Weinberg B. 2019. Federal Funding of Doctoral Recipients: What Can Be Learned from Linked Data. *Research Policy* 48(6):1487–92.

Fleming L, Greene H, Lil G, Marx M, Yao D. 2019. Government-Funded Research Increasingly Fuels Innovation. *Science* 324(6446):1139–41.

Lane J, Owen-Smith J, Rosen R, Weinberg B. 2015. New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value. *Research Policy* 44(9):1659–71.

National Center for Science and Engineering Statistics (NCSES). 2021. *Federal Funds for Research and Development: Fiscal Years 2019–20*. NSF 21-329. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21329/.

Notes

[1] See questions A42 and A43 of the 2015 SDR questionnaire at https://www.nsf.gov/statistics/srvydoctoratework/surveys/srvydoctoratework_2015.pdf.

[2] This question is subject to missing data, which are replaced with imputed values. The imputation rates are 10.15% for question A42 (reporting of federal support) and 10.28% for question A43 (specific federal source of support) on the 2015 SDR.

[3] Employment sector reported as of the week of 1 February 2015 for those that were employed.

[4] These race categories were combined into a composite group for the analysis due to low counts in the individual categories.

Acknowledgments and Suggested Citation

Acknowledgments

Kelsey Cooper and Paul Liu (SRI International) assisted with data checking and code review.

Suggested Citation

Freyman C, Chang W-Y, Cooper K, Deitz S, Liu P; National Center for Science and Engineering Statistics (NCSES). 2022. *Matching SDR Respondents to Investigators of NSF Awards*. Working Paper NCSES 22-211. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/ncses22211/.

Contact

Report Authors

Christina Freyman
Evaluator
Evaluation and Assessment Capability, NSF, under contract to NCSES

Steven Deitz
Analyst
SRI International, under contract to NCSES

Wan-Ying Chang
Mathematical Statistician
Statistics and Methods Program, NCSES
E-mail: **wchang@nsf.gov**
Tel: (703) 292-2310

NCSES

National Center for Science and Engineering Statistics
Directorate for Social, Behavioral and Economic Sciences
National Science Foundation
2415 Eisenhower Avenue, Suite W14200
Alexandria, VA 22314
Tel: (703) 292-8780
FIRS: (800) 877-8339
TDD: (800) 281-8749
E-mail: **ncsesweb@nsf.gov**

Appendix A: NSF Awards Summary Data

This appendix provides summary data on the National Science Foundation (NSF) awards data set used in the matching procedure. The distribution of NSF awards across the award attributes available in the database provides information on the funding patterns of the foundation. Labels are shown as they appear in the database (see appendix table A-1 through appendix table A-3).

The XML data include a value for field of application, which is not displayed on the NSF awards website. Roughly half of the awards in the database (260,236) include a field of application; of these, about half are labeled as “other applications nec.” NSF administration has indicated that these labels have not been used in many years, which is supported by the recent lack of coverage in the data. Appendix table A-4 shows the top 10 values of field of application. Each award may contain multiple labels.

NSF awards are assigned program reference codes, of which there are 2,096. Because awards can be assigned more than one program reference code, there are 876,418 award-program reference code pairs. The top 10 most frequent values after “unassigned” are shown in appendix table A-5. These 10 categories represent about 35% of the total award-program reference code pairs, indicating a highly skewed distribution.

NSF awards are also assigned program element codes. There are 2,054 unique program element codes in the database. An award can have multiple codes, resulting in 532,851 award-program element pairs. Appendix table A-6 shows the top 10 codes by award count. These 10 codes represent about 9% of the pairs, also indicating a highly skewed distribution. For the program reference and program element codes to be useful for award classification, a standard ontology would need to be developed and applied.

Additional Classifications

Two new classification variables for NSF awards were developed in the course of this project: funding program, and grantee type. First, based on the existing “primary program” field in the award records, NSF awards were classified into four funding programs: research, training, facilities, and other. Appendix table A-7 displays counts and percentages by funding program and primary program. Approximately 4% of records, dating mainly from 1993 or earlier, do not have an entry for primary program.

Second, the “types of grantees” field was developed to identify grantee types across all awards. NSF provides grants to various kinds of institutions, but because it does not publish a grantee classification type, a rule-based classification process was followed in order to categorize each award recipient. A total of 19,694 unique institutions are represented in the data. Using the institution name to institution type crosswalk from the National Center for Science and Engineering Statistics Taxonomy of Institution Names, about 5,000 institutions were mapped to categories using a direct name or standardized name match to the NSF listings. The remaining institutions were manually mapped according to rules applied to the contents of the name. For example, records containing “school” were classified as K–12 institutions, and those with “LLC” were mapped to the private sector (see appendix table A-8).

Table	Title
A-1	NSF awards, by award type
A-2	NSF awards, by directorate or office label: 1959–2020
A-3	NSF awards, by funding division: 1959–2020

Table	Title
A-4	Top 10 fields of application for NSF awards: 1959–2020
A-5	Top 10 program reference codes for NSF awards: 1959–2020
A-6	Top 10 program element codes for NSF awards: 1959–2020
A-7	Primary program and funding program mapping: 1959–2020
A-8	NSF awards, by type of grantee: 1959–2020

TABLE A-1

NSF awards, by award type

(Number)

Type of award	Number
BOA/task order	801
Continuing grant	141,199
Contract	1,259
Contract interagency agreement	943
Cooperative agreement	2,553
Fellowship	11,628
Fellowship award	1,242
Fixed amount award	12
Fixed price award	186
GAA	110
Interagency agreement	2,162
Standard grant	302,552
Total	464,647

BOA = Basic Ordering Agreement; GAA = General Agency Agreement; NSF = National Science Foundation.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

TABLE A-2

NSF awards, by directorate or office label: 1959–2020

(Number and percent)

Directorate or office	Number	Percentage
Directorate for Biological Sciences	64,960	13.98
Directorate for Computer and Information Science and Engineering	57,743	12.43
Directorate for Education and Human Resources	42,228	9.09
Directorate for Mathematical and Physical Sciences	96,224	20.71
Directorate for Social, Behavioral and Economic Sciences	37,328	8.03
Directorate for Biological Sciences	47	0.01
Directorate for Computer and Information Science and Engineering	93	0.02
Directorate for Education and Human Resources	17	0.00
Directorate for Engineering	76,305	16.42
Directorate for Geosciences	65,243	14.04
Directorate for Mathematical and Physical Sciences	16	0.00
Directorate for Social, Behavioral and Economic Sciences	64	0.01
National Coordination Office	17	0.00
National Nanotechnology Coordinating office	5	0.00
Office of Budget, Finance, and Award Management	341	0.07
Office of Information and Resource Management	245	0.05
Office of Polar Programs	3	0.00
Office of the Director	23,625	5.08
Missing	143	0.03
Total	464,647	100.00

NSF = National Science Foundation.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-3

NSF awards, by funding division: 1959–2020

(Number and percent)

Funding division	Number	Percentage
Division of Mathematical Sciences	35,727	7.69
Division of Earth Sciences	22,800	4.91
Division of Undergraduate Education	22,208	4.78
Division of Civil, Mechanical, and Manufacturing Innovation	21,136	4.55
Division of Chemistry, Bioengineering, Environment, and Transportation Systems	21,063	4.53
Division of Chemistry	20,661	4.45
Division of Ocean Sciences	19,964	4.30
Division of Materials Research	18,921	4.07
Division of Environmental Biology	18,809	4.05
Office of International Science and Engineering	18,728	4.03
Division of Behavioral and Cognitive Science	18,610	4.01
Division of Industrial Innovation and Partnerships	17,173	3.70
Division of Social and Economic Sciences	16,902	3.64
Division of Molecular and Cellular Bioscience	16,193	3.49
Division of Integrative Organismal Systems	15,344	3.30
Division of Computing and Communication Foundations	15,130	3.26
Division of Atmospheric and Geospace Sciences	13,961	3.00
Division of Computer and Network Systems	13,532	2.91
Division of Information and Intelligent Systems	12,529	2.70
Division of Physics	12,415	2.67
Electrical, Communications and Cyber Systems	11,035	2.37
Division of Biological Infrastructure	10,263	2.21
Division of Research on Learning	9,387	2.02
Division of Astronomical Sciences	7,832	1.69
Division of Experimental and Integrative Activities	6,825	1.47
Office of Polar Programs (OPP)	6,311	1.36
Missing	5,857	1.26
Division of Graduate Education	5,503	1.18
Division of Human Resource Development	4,776	1.03
Office of Advanced Cyberinfrastructure (OAC)	4,544	0.98
Division of Engineering Education and Centers	3,149	0.68
Office of Polar Programs	2,188	0.47
Division of Integrative Organismal Sys	1,650	0.36
Directorate for Engineering	1,402	0.30
Office of Integrative Activities	1,082	0.23
SBE Office of Multidisciplinary Activities	962	0.21
Division of Behavioral and Neural Sciences	947	0.20
Emerging Frontiers	917	0.20
Division of Polar Programs	870	0.19
Division of Microelectronic Information Processing Systems	814	0.18
ICER	728	0.16
Emerging Frontiers and Multidisciplinary Activities	715	0.15
National Center for S&E Statistics	670	0.14
Directorate for Geosciences	574	0.12
Office of Planning and Assessment	493	0.11
Direct for Computer and Info Science and Engineering	473	0.10
Division of Cellular Biosciences	400	0.09
Direct for Biological Sciences	393	0.08
Division of Educational System Reform	310	0.07
EPSCoR	306	0.07

TABLE A-3

NSF awards, by funding division: 1959–2020

(Number and percent)

Funding division	Number	Percentage
Division of Grants and Agreements	199	0.04
Directorate for Social, Behavioral and Economic Sciences	169	0.04
Division of Human Resource Management	119	0.03
Office of Diversity and Inclusion	100	0.02
Office of Inspector General	66	0.01
Division of Industrial Innovation and Partnerships	62	0.01
Office of Legislative and Public Affairs	56	0.01
Division of Administrative Services	55	0.01
Division of Information Systems	55	0.01
National Center for Science and Engineering Statistics	52	0.01
Directorate for Education and Human Resources	48	0.01
MPS Multidisciplinary Activities	48	0.01
Directorate for Mathematical and Physical Sciences	38	0.01
Division of Financial Management	38	0.01
Division of Institution and Award Support	30	0.01
Office of Budget, Finance, and Award Management	26	0.01
Division of Experimental and Integrative Activities	25	0.01
Budget Division	24	0.01
National Science Board	21	0.00
Division of Materials Development	20	0.00
Division of Acquisition and Cooperative Support	19	0.00
Division of Civil, Mechanical, and Manufacturing Innovation	18	0.00
Office of the Director	18	0.00
National Coordination Office	17	0.00
General Counsel	16	0.00
Division of Information and Intelligent Systems	15	0.00
Division of Computer and Communication Foundations	13	0.00
Office of Information and Resource Management	13	0.00
Division of Molecular and Cellular Biosciences	10	0.00
Division of Atmospheric and Geospace Sciences	8	0.00
Division of Behavioral and Cognitive Sciences	8	0.00
Division of Research on Learning in Formal and Informal Settings (DRL)	8	0.00
Division of Chemical, Bioengineering, Environmental, and Transport Systems	6	0.00
CISE Information Technology Research	5	0.00
Directorate for Computer and Information Science and Engineering	5	0.00
Division of Biological Infrastructure	5	0.00
Large Facilities Office	5	0.00
National Nanotechnology Coordinating Office	5	0.00
Arctic Sciences Division	3	0.00
Directorate for Biological Sciences	3	0.00
Division of Elementary, Secondary and Informal Science Education	3	0.00
Division of Social and Economic Sciences	3	0.00
Office of Science and Technology Infrastructure	3	0.00
Division of Research, Evaluation and Communication	2	0.00
Division of Design and Manufacturing Innovation	1	0.00
Division of Polar Programs	1	0.00
SBE Office of Multidisciplinary Activities	1	0.00
Total	464,647	100.00

CISE = Computer and Information Science and Engineering; EPSCoR = Established Program to Stimulate Competitive Research; ICER = Integrative and Collaborative Education and Research; MPS = Mathematical and Physical Sciences; NSF = National Science Foundation; SBC = small business concern; SBE = Social, Behavioral and Economic Sciences.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-4

Top 10 fields of application for NSF awards: 1959–2020

(Number)

Field of application	Number
Other applications nec	118,753
Life science biological	22,941
Industrial technology	20,551
Materials research	17,644
Oceanography	14,425
Mathematics	13,676
Other sciences nec	13,157
Geological sciences	12,460
Human subjects	12,237
Chemistry	11,623

nec = not elsewhere classified; NSF = National Science Foundation.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-5

Top 10 program reference codes for NSF awards: 1959–2020

(Number)

Program reference code	Number
Other Research or Education	72,295
Science, Math, Engineering & Tech Education	46,601
Undergraduate Education	43,055
Exp Prog to Stim Comp Res	28,020
Reu Supp-Res Exp For Undergrad Supp	21,140
Environment and Global Change	20,997
High Performance Computing & Comm	20,941
Graduate Involvement	20,147
Biotechnology	17,495
Advanced Materials & Processing Program	13,441

NSF = National Science Foundation.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-6

Top 10 program element codes for NSF awards: 1959–2020

(Number)

Program element code	Number
Molecular Biophysics	5,890
Undergrad Instrument & Lab Improve	5,543
EPSCoR Co-Funding	5,238
Major Research Instrumentation	5,014
Algebra, Number Theory, and Com	4,893
Applied Mathematics	4,757
Economics	4,609
Cross-Directorate Programs	4,109
Geophysics	4,087
S-STEM-Schlr Sci Tech Eng & Math	4,047

EPSCoR = Established Program to Stimulate Competitive Research; NSF = National Science Foundation; STEM = science, technology, engineering, and mathematics.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-7

Primary program and funding program mapping: 1959–2020

(Number and percent)

Funding program classification	Primary program	Number	Percent
Research	040100 NSF RESEARCH & RELATED ACTIVIT	149,770	45.8
	040101 RRA RECOVERY ACT	5,035	1.5
	490100 NSF RESEARCH & RELATED ACTIVIT	124,816	38.2
Training	040106 NSF Education & Human Resource	12,983	4.0
	040107 EHR RECOVERY ACT	97	0.0
	045176 H-1B FUND, EHR, NSF	3,570	1.1
	490106 NSF, EDUCATION & HUMAN RESOURC	13,658	4.2
	490106 NSF, Education & Human Resource	1,493	0.5
Facilities and equipment	040551 NSF MAJOR RESEARCH EQUIPMENT	65	0.0
	040552 MREFC RECOVERY ACT	3	0.0
	490150 NSF ACADEMIC RESEARCH FACILITI	820	0.3
Other	040180 NSF Agency Oper & Award Mgmt	14	0.0
	040301 OIG RECOVERY ACT	2	0.0
	048960 NSF TRUST FUND	694	0.2
	491014 DVLP FUND FOR AFRICA, A.I.D.	1	0.0
	491021 AGENCY FOR INTERNATIONAL DEVEL	3	0.0
	49T566 US INDIA FUND (RUPEES) (P&I)	74	0.0
NA	NA	13,617	4.2
	Total	326,715	

NA = not available.

AID = Agency for International Development; EHR = Education and Human Resources; MREFC = Major Research Equipment and Facilities Construction; NSF = National Science Foundation; OIG = Office of Inspector General; P&I = pensions and investments; RRA = research and related activities.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

TABLE A-8

NSF awards, by type of grantee: 1959–2020

(Number and percent)

Type of grantee	Number	Percent
College or university	2,618	13.29
Federally funded research and development center (FFRDC)	20	0.10
Hospital/ medical center	64	0.32
Individual	4,705	23.89
K–12 school	2,279	11.57
Private sector	9,379	47.62
Public sector (federal)	311	1.58
Public sector (local)	45	0.23
Public sector (state)	166	0.84
Unknown	107	0.54
Total	19,694	100.00

NSF = National Science Foundation.

Source(s):
National Center for Science and Engineering Statistics, processed NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp), accessed 15 May 2020.

Appendix B: Supplemental Data of NSF Support in the SEH Doctorate Population

Table	Title
B-1	Estimates and standard errors for SEH doctorate holders receiving an NSF award, by doctorate year and major field of doctorate
B-2	Estimates and standard errors for SEH doctorate holders receiving an NSF award, by employment sector
B-3	Estimates and standard errors for SEH doctorate holders receiving an NSF award, by employment sector, sex, race, and ethnicity

TABLE B-1

Estimates and standard errors for SEH doctorate holders receiving an NSF award, by doctorate year and major field of doctorate

(Number and percent)

	Estimates			Standard errors		
Doctorate year and major field of doctorate	NSF award	No award	Share (%)	NSF award	No award	Share (%)
All doctorates	86,850	961,050	8.30	1,325	1,700	0.15
Biological, agricultural, and environmental life sciences	18,000	236,600	7.10	600	825	0.25
Computer and information sciences	5,200	25,600	16.90	325	350	1.05
Mathematics and statistics	7,550	44,350	14.60	325	425	0.65
Physical sciences	22,500	154,550	12.70	625	675	0.35
Psychology	3,950	135,600	2.80	250	500	0.20
Social sciences	9,800	132,700	6.90	450	675	0.35
Engineering	19,150	183,750	9.40	700	800	0.35
Health	650	47,850	1.30	150	275	0.30
Doctorates since 2005	17,450	281,800	5.80	575	775	0.20
Biological, agricultural, and environmental life sciences	2,650	73,550	3.50	225	725	0.30
Computer and information sciences	1,950	11,950	14.10	225	325	1.50
Mathematics and statistics	1,400	13,000	9.60	150	375	1.10
Physical sciences	4,100	38,950	9.60	250	550	0.60
Psychology	800	29,350	2.70	125	475	0.40
Social sciences	1,850	35,550	5.00	175	525	0.45
Engineering	4,500	61,500	6.80	325	725	0.50
Health	200	18,000	1.10	75	375	0.35

NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

TABLE B-2

Estimates and standard errors for SEH doctorate holders receiving an NSF award, by employment sector

(Number and percent)

Employment sector	Estimates			Standard errors		
	NSF award	No award	Share (%)	NSF award	No award	Share (%)
All employment sectors	86,850	961,050	8.30	1,325	1,700	0.15
4-year educational institution	73,150	326,650	18.30	1,225	2,100	0.30
Other educational institution	1,000	28,900	3.30	175	750	0.55
Private, for-profit	2,900	288,100	1.00	300	2,125	0.10
Self-employed	500	47,400	1.10	125	1,050	0.25
Private, nonprofit	2,450	54,250	4.30	200	1,075	0.35
Federal government	1,250	49,500	2.50	150	1,000	0.30
State or local government	200	16,000	1.30	75	625	0.35
Non-U.S. government	150	13,950	1.00	50	600	0.35
Not applicable	5,150	136,250	3.70	400	1,475	0.30

NSF = National Science Foundation; SEH = science, engineering, and health.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).

TABLE B-3

Estimates and standard errors for SEH doctorate holders receiving an NSF award, by employment sector, sex, race, and ethnicity

(Number and percent)

	Estimates			Standard errors		
Employment sector, sex, race, and ethnicity	NSF award	No award	Share (%)	NSF award	No award	Share (%)
All doctorate holders						
Both sexes	86,850	961,050	8.30	1,325	1,700	0.15
Asian	13,600	245,200	5.30	625	1,000	0.25
Black	2,150	31,800	6.40	175	350	0.50
Hispanic	3,650	44,350	7.60	225	425	0.45
White	66,100	627,050	9.50	1,150	1,600	0.20
Other	1,300	12,700	9.40	175	400	1.05
Female	21,050	320,350	6.20	575	800	0.20
Asian	2,550	69,450	3.60	225	800	0.30
Black	700	14,450	4.70	100	300	0.60
Hispanic	1,100	16,400	6.30	100	275	0.55
White	16,250	214,650	7.00	525	1,100	0.25
Other	450	5,350	7.40	75	250	1.15
Male	65,750	640,750	9.30	1,175	1,500	0.20
Asian	11,050	175,750	5.90	575	1,050	0.30
Black	1,450	17,300	7.80	175	400	0.80
Hispanic	2,550	27,950	8.40	200	425	0.60
White	49,850	412,350	10.80	1,025	1,775	0.25
Other	900	7,350	10.90	150	375	1.60
Employed by 4-year educational institutions						
Both sexes	73,150	326,650	18.30	1,225	2,100	0.30
Asian	12,000	85,400	12.30	575	1,275	0.60
Black	1,900	12,350	13.40	175	325	1.10
Hispanic	3,150	19,250	14.10	175	425	0.75
White	54,950	205,200	21.10	1,075	1,725	0.40
Other	1,100	4,450	19.80	150	275	2.40
Female	18,050	119,500	13.10	525	1,100	0.40
Asian	2,250	25,900	7.90	200	600	0.70
Black	600	5,600	9.90	100	225	1.40
Hispanic	950	7,100	11.80	100	250	1.05
White	13,900	78,900	15.00	475	975	0.45
Other	350	2,000	15.10	75	150	2.45
Male	55,150	207,150	21.00	1,100	1,925	0.40
Asian	9,800	59,500	14.10	550	1,175	0.75
Black	1,300	6,800	16.10	150	300	1.65
Hispanic	2,200	12,150	15.40	175	350	1.05
White	41,100	126,300	24.50	950	1,575	0.55
Other	750	2,450	23.30	150	225	3.60

NSF = National Science Foundation; SEH = science, engineering, and health.

Note(s):
Other includes American Indian or Alaska Native, Native Hawaiian or Pacific Islander, and multiple race. Hispanic may be any race; Asian, Black, and White exclude Hispanic origin.

Source(s):
National Center for Science and Engineering Statistics, Survey of Doctorate Recipients, 2015, linked to the public NSF Awards Database (https://www.nsf.gov/awardsearch/download.jsp).