New Science and Engineering Research Enterprise Data on U.S.-trained PhDs – Survey, Administrative, Transaction, and Bibliometrics

Jodi Basner (Clarivate), Wan-Ying Chang* (NCSES), Maryah Garner (Coleridge Initiative), Jason Owen-Smith (Institute for Research on Innovation and Science)

Value of Science Conference, June 23, 2021

National Center for Science and Engineering Statistics
Social, Behavioral and Economic Sciences
National Science Foundation
Disclaimer

This presentation reports findings of research undertaken by staff at the National Center for Science and Engineering Statistics (NCSES) at the National Science Foundation (NSF).

The views expressed are attributable to the authors and do not necessarily reflect those of the NCSES or NSF.
NCSES is the nation's leading provider of statistical data on the science and engineering enterprise
NCSES conducts more than a dozen surveys

Surveys on U.S.-trained PhDs: timeline and scope

Enter doctoral program

Census of Research PhDs from all fields

Sample of Science Engineering Health PhDs

~ 5 years

SED

2 years

SDR

2 years

SDR

2 years

SDR

... SDR until age 76
Research agenda

- How does funding of graduate students differ by research field, race, and sex?
- What are the federal source and type of graduate funding?
- What is the impact of federal research funding on subsequent research productivity?
- What is the impact of NSF and NIH funding for graduate education on job placement, career progression, are research output?
- How do the scientific careers of women and minorities differ from men and majority?

Other individual factors:
- Level of graduate debts
- Research experiences

Indicators of labor force mobility: mobility between employers, sectors, occupation
- Scientific international mobility
- Why do foreign nationals with US PhDs return home?
- Effect of diversity on research network, interdisciplinarity
New research data on U.S.-trained PhDs: survey + administrative + transaction + bibliometrics
Data linkage methods

SED-UMETRICS

- An "exact match" approach, matching names and birth dates of individuals from the same institution in 9 sequential steps
- For privacy protection, the personally identifiable information is hashed for the data sets
- Conducted in the Administrative Data Research Facility – FedRAMP certified secure platform

SDR-WoS

- Two stages of machine learning approaches implemented sequentially to address i) survey-to-publication linking and ii) publication-to-publication linking
- Required building resource-intensive Gold Standard datasets
- Conducted in the NCSES Secure Data Access Facility
New insights: Funding level and source for doctoral students varies by field of study

Match rate and funding source by doctorate field of study

<table>
<thead>
<tr>
<th>Field</th>
<th>Federal</th>
<th>Both</th>
<th>Non-federal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical sciences</td>
<td>9%</td>
<td>27%</td>
<td>64%</td>
</tr>
<tr>
<td>Engineering</td>
<td>6%</td>
<td>32%</td>
<td>60%</td>
</tr>
<tr>
<td>Computer sciences</td>
<td>13%</td>
<td>36%</td>
<td>51%</td>
</tr>
<tr>
<td>Psychology</td>
<td>5%</td>
<td>22%</td>
<td>29%</td>
</tr>
<tr>
<td>Non-SEH</td>
<td>17%</td>
<td>34%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Source: SED 2014-2017 cohort matched to UMETRICS 2020 data

Funding duration by primary source of support

- Research Assistant
- Fellowship Grant
- Teaching Assistant
- Other
- Own fund

- 0-1 year
- 1-2 year
- 2-3 year
- 3-4 year
- 4-5 year
New insights: Pre-PhD publishing paves the road to academic R&D

Pre-doctorate publishing rates varied by field of study and type of financial support during graduate school.

Those with pre-PhD publications show higher rate of persistent post-PhD publishing and holding academic R&D position in 2015.

Source: 2015 SDR-WoS linked data, subset for graduate cohort FY 1994 – 2007; alluvial plot shows only the U.S.-residing subset.
Future research

- Expanding data network by linking to PatentsView
- Matching methodology research
- Measuring linkage errors and developing data quality standards
- Developing meaningful measures and user-friendly analytical data files
- Privacy & confidentiality protection
- Building user community