

UNIVERSITY OF MISSOURI ST. LOUIS AND GEORGIA STATE UNIVERSITY

Linking Investigator-Initiated Federal Research Grants with the Production and Scientific Development of Doctoral Scientists and Engineers

Prepared for Science of Science Measurement
Workshop, December 2-3, 2010, Washington, D.C.

Sharon G. Levin, Professor of Economics, University of Missouri-St. Louis,

St. Louis MO 63121-4400 (slevin@umsl.edu) and

Paula E. Stephan, Professor of Economics, Georgia State University,

Box 3992, Atlanta, GA 30302-3992 (pstephan@gsu.edu) and NBER

Introduction

Each year more than 25,000 students receive a PhD in a STEM field in the United States.¹ These individuals are vital to U.S. science. As students they play an important role in producing research; as graduates they play an even greater role in transmitting the knowledge they learned in graduate school as well as establishing research programs of their own. Moreover, PhD students are primarily supported in graduate school on research assistantships, traineeship awards or fellowships, most of the funding for which comes from the federal government. Thus they represent a major investment on the part of the federal government. To be a bit more specific, in 2008, the latest year for which data are available, for 75% of the PhD graduates in the life sciences the primary source of support was either a research assistantship/traineeship or a fellowship/grant. In physical sciences, 68% were supported by such means; the percent is even higher in engineering where almost 82% were supported either as a graduate research assistant or on a training grant or fellowship.² The large role that the federal government plays is not by accident. Both NIH and NSF by design provide funds for the support of graduate students. According to Rita Colwell, the Director of NSF from 1998-2004, “In the 1980s, NSF asked investigators to put graduate students on their research budgets, saying it preferred to fund graduate students rather than technicians.”³

The important role that graduate students play in research can readily be seen by studying a sample of articles published in the journal *Science*. For example, for a six-month period in 2007, 20% of all U.S. authors were graduate students; in almost one-out-of three instances the graduate student was the first-author—the author who generally does the heavy lifting.⁴

¹ In 2008, the latest year for which data are available, 11,088 received degrees in the life sciences, 8,129 in the physical sciences and 7,862 in engineering. See Table 5, National Science Foundation 10-309, *Doctorate Recipients from U.S. Universities: Summary Report 2007-2008*.

² Op. cit. Table 22. According to the *Survey of Graduate Students and Postdoctorates in Science and Engineering: 2008*, approximately 54,000 full time graduate students were supported by NSF, NIH and DOD in 2008. Although this figure includes support for students in the social sciences and psychology, the vast majority are in STEM fields. See, http://www.nsf.gov/statistics/srvygradpostdoc/pub_data.cfm.

³ See, “The Biocomplex World of Rita Colwell.” *Science*, 281, 25 September 1998, 1944-1947.

⁴ The sample is limited to papers having a last author from a U.S. academic institution. See Grant Black and Paula Stephan, “The Economics of University Lab Science and the Role of Foreign Graduate Students

Recent graduates play an even greater role in transmitting the knowledge they learned in graduate school as well as establishing research programs of their own. Face-to-face interaction is an important means of transmission since part of the knowledge acquired in graduate school is of a tacit nature which cannot be codified. Creating transgenic mice, for example, was not something that one could pick up by reading an article—one needed to train in the lab of someone who had the expertise. Likewise, the new technology of microfluidics requires hands on training. More generally, and to quote the physicist J. Robert Oppenheimer, “the best way to send information is to wrap it up in a person.”⁵ The placement of newly trained PhDs in industry, government, and academic research settings is an extremely important mechanism by which new knowledge is diffused.

Recently trained graduate students also set out on research careers. The direction of their research and their attitudes towards doing research are shaped largely in graduate school where they are socialized to research and acquire the skills and knowledge to become researchers after they leave graduate school. The mentoring they receive from the director of the lab plays a key role in determining their attitudes and tolerance for risk when it comes to research.⁶ It also plays a role in shaping their attitudes and perspective on doing interdisciplinary research as well as frontier research. Recent graduates are especially key to the research health of the nation given the evidence that while one does not have to be extremely young to do great research, age matters. Virtually no Nobel laureates, for example, received the prize for work that they started after the age of 50.⁷

The important role that young scientists play in the health of the research enterprise and the overall health of the U.S. economy was stressed in “Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future” which was issued in 2007

and Postdoctoral Scholars,” in *American Universities in a Global Market*,” edited by Charles Clotfelter, University of Chicago Press, 2010, pp. 129-162.

⁵ J. Robert Oppenheimer, as quoted in Anon, “The eternal apprentice,” *Time* magazine, vol. 52(8 November 1948): 70-81, on p. 81.

⁶ Michael T. Nettles and Catherine M. Millett, 2006. *Three Magic Letters: Getting to Ph.D.* Baltimore: Johns Hopkins University Press.

⁷ Paula Stephan and Sharon Levin, “Age and the Nobel Prize Revisited,” *Scientometrics*, 28(3): 387-99, 1993.

and re-issued in 2010.⁸ Most recently, the National Academies has convened a blue-ribbon committee to recommend ways to keep research universities healthy. Part of the mandate will be on the training role of universities as well as the ways in which universities nurture early career scientists.

Despite the critical role that doctoral recipients play in the growth and development of scientific knowledge, and by extension the economy, we have little up-to-date information about their production, in terms of research focus. This is because the Survey of Earned Doctorates (SED), the nation's key database for collecting information on doctoral recipients, is released one to two years after the degree is conferred and does not permit the matching of individual-level information with external sources. Field of training is collected at a relatively aggregate level. For example, currently a PhD recipient in science or engineering must choose among approximately 150 fields. Thus the survey provides at best a retrospective view at a relatively aggregate level of changing trends in the training of the doctoral workforce.

More importantly, the SED does not permit an analysis of how new ideas and techniques are embodied in the newly-trained scientific workforce. Perhaps most important of all, the survey provides no information regarding the impact of thesis advisors on the intellectual capital and creativity of their newly-trained mentees. Yet we know from the work of others that the training and mentoring-relationship at the dissertation stage as well as at the postdoctoral stage is instrumental to the future career of the doctoral student.⁹ In some instances funding for the student or postdoctoral fellow comes from fellowships or training grants but more generally funding comes from principal investigator research grants.

Here we suggest a somewhat novel approach that will permit a ready count of newly-created PhDs by narrowly defined research topic. The approach can also be used to analyze the degree to which new ideas and techniques are embodied in the recently trained. And, by

⁸See, *Rising Above the Gathering Storm Revisited: Rapidly Approaching Category 5*. Washington, DC: National Academies Press, 2010.

⁹Harriet Zuckerman (*Scientific Elite: Nobel Laureates in the United States*. New York: The Free Press, 1977) finds that over half of American Nobel laureate scientists studied or worked under other Nobel laureates. See, also, Robert Kanigel (1986), *Apprentice to Genius: The Making of a Scientific Dynasty*. New York: MacMillan and Azoulay, P., Liu, C. and Stuart, T. "Social Influence Given (Partially) Deliberate Matching: Career Imprints in the Creation of Academic Entrepreneurs." See, (http://pazoulay.scripts.mit.edu/Working_Papers.html).

linking the newly-created database proposed in this paper to data on federal funding, the approach can provide insights regarding the degree to which investigator-initiated Federal research grants affect intergenerational knowledge transfer.¹⁰

Proposed approach

Our approach is based on using the *ProQuest Dissertations & Theses Database*¹¹ to create a continuously updated database of *Newly-Minted Scientific Talent* (NMST) in the United States at the doctoral-level. ProQuest contains abstracts and full texts of PhD dissertations including the name of the student, the title of the dissertation, the subject area and keywords, the date of degree, the department as well as the university granting the degree, and the name of the chair of the dissertation committee. Furthermore, this database can be searched annually or over specific time periods.

The NMST database would permit a count of newly-minted PhDs by fine-field of research interest. It could also be used to provide information regarding the degree to which the focus of the newly trained is changing and the degree to which their focus changes in relationship to funding initiatives and special research foci of agencies. Furthermore, when linked to information collected by the federal and university partnership Star Metrics (Science and Technology in America's Reinvestment Measuring the Effect of Research and Innovation, Competitiveness and Science)¹² concerning faculty recipients of federal funding, the database could provide a powerful tool for analyzing how federal funding affects the future scientific workforce. By way of example, one could compare the PhD dissertation research of students who work with faculty supported on R01s to the dissertation work of those who work with Pioneer Award winners or Eureka (Exceptional, Unconventional Research Enabling Knowledge

¹⁰Azoulay, P., Liu, C. and Stuart, T., *op. cit.*

¹¹Recent work by P. Gaule and M. Piacentini

(http://siteresources.worldbank.org/INTINTERNATIONAL/Resources/1572846-1253029981787/6437326-1253030199852/Piacentini_Gaule_ppt.pdf) and M. MacGarvie ("Using

Published Dissertations to Identify Graduates' Countries of Origin," presented at the NBER Conference on Career Patterns of Foreign born and Engineers, Scientists November 7, 2007) have extracted data from ProQuest.

¹²See (http://nrc59.nas.edu/star_info2.cfm).

Acceleration) recipients.¹³ Do programs such as the latter that target investigators who test novel, often unconventional hypotheses or tackle major methodological challenges, engender PhD dissertations that differ in terms of novelty and creativity from those engendered from “bread and butter grants” such as the R01?¹⁴ While endogeneity may, in part, be responsible for observed differences, a first step in addressing such an issue is the creation of a strong database.

The NMST database could also potentially be used to analyze the future career outcomes of doctoral recipients by matching their ProQuest identifiers with data sources that track publications and grant awards.¹⁵ Thus one could examine the long-term research productivity of doctoral students who worked with R01 recipients and compare it to the long-term research productivity of students who worked with Eureka recipients or Pioneer recipients. One could also analyze the degree to which the student’s research evolves over the career vs. the degree to which it stays closely linked to the dissertation topic.¹⁶ In what follows, we give some practical examples of how using ProQuest can assist in answering some of the important questions we have posed concerning development of the doctoral labor force.

¹³“The NIH Director’s Pioneer Award Program is designed to “support individual scientists of exceptional creativity who propose pioneering – and possibly transforming approaches – to major challenges in biomedical and behavioral research ... To be considered pioneering, the proposed research must reflect ideas substantially different from those already being pursued in the investigator’s laboratory or elsewhere,” see (<http://nihroadmap.nih.gov/pioneer/index.aspx>). Eureka awards are conceptually similar to Pioneer Awards, see (<http://www.nigms.nih.gov/Research/Mechanisms/EUREKA.htm>). This is not to say that Pioneer or Eureka recipients may not have also received R01 grants.

¹⁴Recent work by Azoulay, P., Manso, G. and Zivin, J. (“Incentives and Creativity: Evidence from the Academic Life Sciences.” *NBER Working Paper #15466*) derives measures of novelty by comparing key words in the scientist’s work to key words in the field as well as key words in previous research undertaken by the scientist.

¹⁵We recognize that because of homonyms this will not be possible in all cases. However, newer bibliometric databases, such as Scopus and Google Scholar which provide information on first name and not just initials (as ISI has done) enhance the possibility of tracking people over time.

¹⁶Software such as Crawdad, for example, which allows one to compute the degree of relatedness of two documents could be used to analyze the degree to which the student’s research relates to the faculty member’s research and the degree to which the student’s dissertation relates to future work. See, Corman, S. and Dooley, K. (2006), *Crawdad Text Analysis System*, Chandler, Arizona: Crawdad Technologies, LLC. For an example of an alternative methodology see Conti, A., Denas, O. and Visentin, F. “Organization of PhD Teams and Research Productivity: Mickey and Goofy or Huey, Dewey and Louie?” Unpublished paper, Georgia Institute of Technology, September 2010.

Example 1. Determining the count of PhDs by fine-field of research interest.

As indicated earlier, the SED identifies field of study fairly broadly. For example, despite the fact that there are 34 degree fields listed in the biological sciences, one could not determine the number of newly-minted PhDs in the specific field of epigenetics using the SED codes. But by searching ProQuest using terms related to epigenetics,¹⁷ we could determine that over the past 25 years, 244 doctoral dissertations on this subject had been written in the United States; 206 of these were completed in just the last five years.¹⁸

Example 2. Determining how the research focus of doctoral students changes in relationship to major funding initiatives at the national or agency level.

To provide an example of how such a database could be used to answer this question we study the National Nanotechnology Initiative (NNI) brought forward at the level of a federal initiative in the Clinton administration's budget of 2001¹⁹ and the 10-year NIH led Protein Structure Initiative (PSI), which commenced in September 2000.²⁰

To gain some understanding as to how the NNI may have affected the production of doctoral recipients trained in nanoscience and nanotechnology, we searched ProQuest for the number of doctoral dissertations produced in the United States containing the term nano.²¹ The results shown in Figure 1 clearly indicate an increased rate of growth in nano-related dissertations subsequent to the initiative's start in 2001. Moreover, Figure 2 suggests that the NNI has also affected the research focus of students in S&E. Indeed, nano-related dissertations

¹⁷Epigenetics is the study of the factors -- anything other than the DNA sequence -- that control gene activity during the development of a complex organism. Today the field often focuses on the heritable traits (over rounds of cell division and perhaps over generations) that do not involve changes to the underlying DNA.

¹⁸The methodology also lends itself to measuring the degree to which students are being trained in multi-disciplinary areas such as systems biology.

¹⁹See <http://www.nano.gov/html/about/history.html>.

²⁰See <http://www.nigms.nih.gov/Initiatives/PSI/>.

²¹This means that the abstract contained a word (or words) starting with the term "nano." We then excluded cases where nano was used solely as an index of measurement such as "nanograms." See Mugoutov, A. and Kahane, B. "Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracing" *Research Policy*, 36, pp. 893-903, 2007. Clearly this methodology undercounts nano-related dissertations since some nano-related research does not contain the word "nano."

have grown from slightly more than 2% to more than 6% of the dissertations produced by students earning doctorates in S&E in the U.S since 2001.²²

Investigating the effects of the Protein Structure Initiative on S&T workforce development at the doctoral level proved to be more difficult. First, it is questionable how many dissertations would actually be associated with the initiative during its pilot phase (September 2002 – June 2005) since the focus initially was on building infrastructure and developing the technologies that would establish an automated pipeline for protein production and structure determination. In addition, because of the wide range of subjects and keywords associated with protein structure analyses, ProQuest might not reliably identify the doctoral dissertations directly related to the PSI. Instead, since all the protein structures solved by the Centers supported by the PSI had to be deposited in the Protein Data Bank (PDB)²³ so that they would be available to other researchers, we indirectly measured the effects of PSI on the production of doctoral students by tracking the number of doctoral dissertations utilizing the PDB, distinguishing between the pre and post PDB pilot phase.²⁴

The results of the ProQuest search for doctoral dissertations using the PDB are shown in Figure 3. These data are consistent with the expectation that the number of doctoral dissertations using the PDB especially after the completion of the pilot phase of PSI, would have increased sharply.

²²We exclude psychology and the social sciences and use summary information from the SED to determine the number of doctoral recipients in S&E. See <http://nsf.gov/statistics/doctorates/> for various years.

²³The PDB an NSF- and NIH-supported public repository of experimentally-determined structures of proteins, nucleic acids, and complex assemblies. These structures can be visualized, downloaded, and analyzed by users who range from students to specialized scientists. See <http://www.rcsb.org/pdb/>.

²⁴During its pilot phase, the initial 9 Structural Genomic Centers of the PSI solved more than 1,100 structures, while the PDB grew by about 18,000 structures; thus the PSI contributed about 6.1% of the total at that time. Today, (as of September 14, 2010), there are 10 PSI Centers and PSI has contributed about 5,100 of the 68,000 contained in the PDB (7.5%). Thus, in addition to signaling the growing national (and global) interest in protein structure determination that the Protein Structure Initiative heralded, PSI has increased the number of protein structures that researchers have access to in the databank.

Example 3. Determining the degree to which investigator-initiated Federal research grants affect intergenerational knowledge transfers.

The Pioneer Award Program began in 2004 with 9 recipients. Using ProQuest, we have identified the doctoral dissertations that were most-likely supported²⁵ by the Pioneer award won by one of the initial recipients, Sunny Xie of Harvard University.²⁶ Professor Xie was committee chair for 9 dissertations during the award period. Our goal here is simply to obtain some insight into the transmission of knowledge from PI to doctoral researcher.

A longer-term goal in subsequent research is to compare the research paths of those working with Pioneer-award recipient to those working with PIs supported by more bread-and-butter grants such as R01s. For example, one could compare, using a measure of novelty such as that used by Azoulay, Manso and Zivin,²⁷ the novelty of the work done by PhD students working under Pioneer and Eureka recipients to the novelty of the work produced by PhDs supported by RO1 grants. Because we have only a short-window of time through which to examine the career paths of these researchers, this goal cannot be met at this time.²⁸ Instead, although we have a limited time period for analysis, we examined the direction in which the publications produced by Xie's doctoral students went subsequent to their degree completion. We traced their publication records using PubMed.²⁹ The results are shown in Table 1.

As Table 1 indicates, in the short period of time that has elapsed since earning their doctorate, only 4 of the 9 doctoral recipients had started to publish with someone other than their mentor Xie, and of these 4, only 2 appear to be embarking on a research path somewhat

²⁵We used PubMed to examine the subject matter and timing of the published articles. We also examined a list of (selected) journal articles that the winner indicated had been supported by the Pioneer grant. See, (<http://nihroadmap.nih.gov/pioneer/Recipients04.aspx>).

²⁶ We also examined the eight dissertations that Larry Abbott of Columbia University chaired during the award period. In the interest of brevity, these results are not reported here.

²⁷ Azoulay, P., Manso, G. and Zivin, J., *op. cit.*

²⁸ This is because their doctorates were only completed within the past five years and often, before striking out on his or her own, a newly-minted PhD serves as a postdoctoral assistant for two or three years.

²⁹ See, (<http://www.ncbi.nlm.nih.gov/pubmed>).

different from the that of Xie.³⁰ Thus, we see that early in the career, the PhD mentor Xie is still playing an important role in career development.

Conclusion

The examples that we have presented in this preliminary paper indicate the usefulness of using ProQuest as an excellent source of up-to-date data on the development of the doctoral workforce as envisioned in the newly-created NMST database. Combined with other sources such as PubMed and the new STAR Metrics database, we are convinced that, with the right expertise and software, investigator-initiated Federal research grants can be linked with the production and scientific development of doctoral scientists and engineers and provide a “rigorous, quantitative basis from which policy makers and researchers can assess the impacts of the Nation’s scientific and engineering enterprise, improve their understanding of its dynamics, and assess the likely outcomes.”³¹

Much can be learned by the development of NMST in the short run; still more in the long run. To be more precise, the following research agenda can be pursued in the short term:

1. Extract data from ProQuest to develop the NMST database.
2. Use the NMST database to track the production of new scientific talent by fine field of expertise.
3. Use the NMST database to track the development of new scientific talent in response to specific federal funding initiatives.

In the medium/long term, it will be necessary to link the NMST database with STAR Metrics as well as sources of individual-level productivity such as publications (PUBMED or SCOPUS) and/or patents (Patent Database). Then, it will be possible to investigate such questions as:

³⁰With more years of data, a more refined analysis of the “scientific distance” between the work of the mentor and his/her mentee (in the biological sciences) can be conducted using the open-source software developed by Azoulay, P. Zivin, J. and Stellman, A ., (<http://www.steelman-greene.com/Scientific Distance/>).

³¹See, *The Science of Science Policy: A Federal Research Roadmap*, November 2008. (<http://www.scienceofsciencepolicy.net/blogs/sosp/pages/sosproadmap.aspx>)

1. The degree to which a newly-minted doctoral student's research evolves over the career vs. the degree to which it stays closely linked to the dissertation topic.
2. How investigator-initiated Federal research grants affect intergenerational knowledge transfers.
3. Whether mentorship by exceptionally innovative or creative scientists increases the likelihood that their mentees will be exceptionally innovative or creative when compared with their peers mentored by less distinguished scientists.

Figure 1. Nanoscience and nanotechnology dissertations, 1990-2009

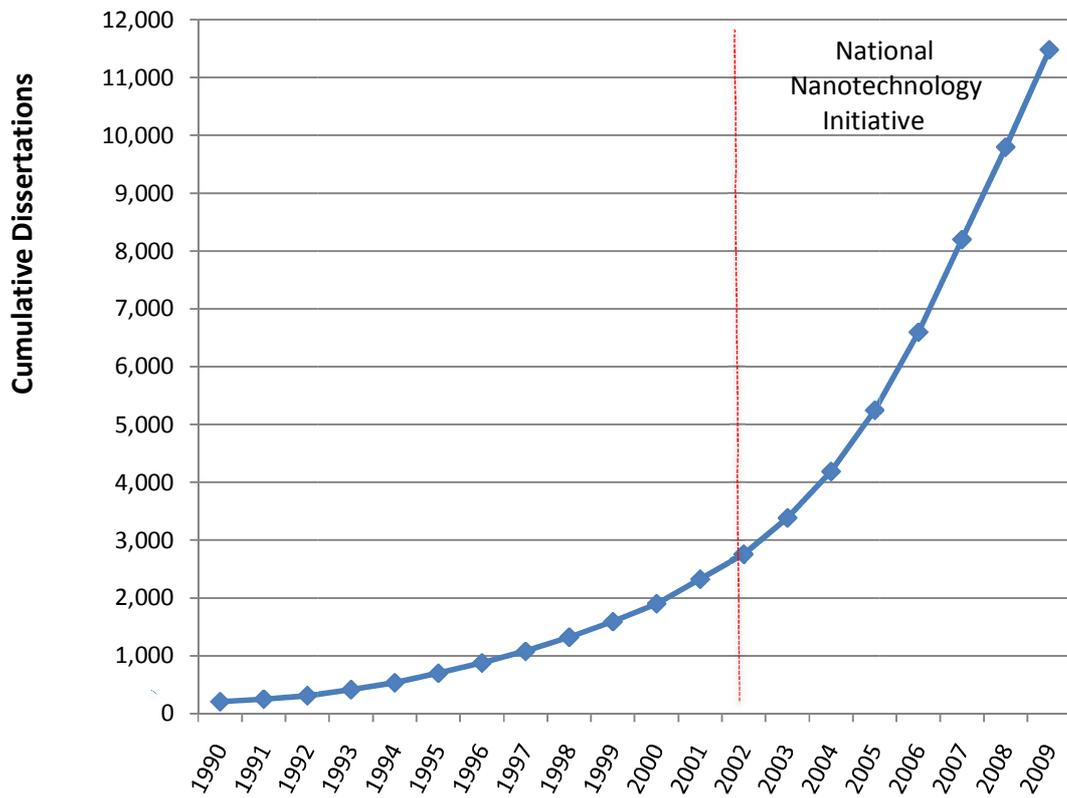


Figure 2. Percentage of Dissertations in S&E in Nanoscience and Nanotechnology, 1990-2008

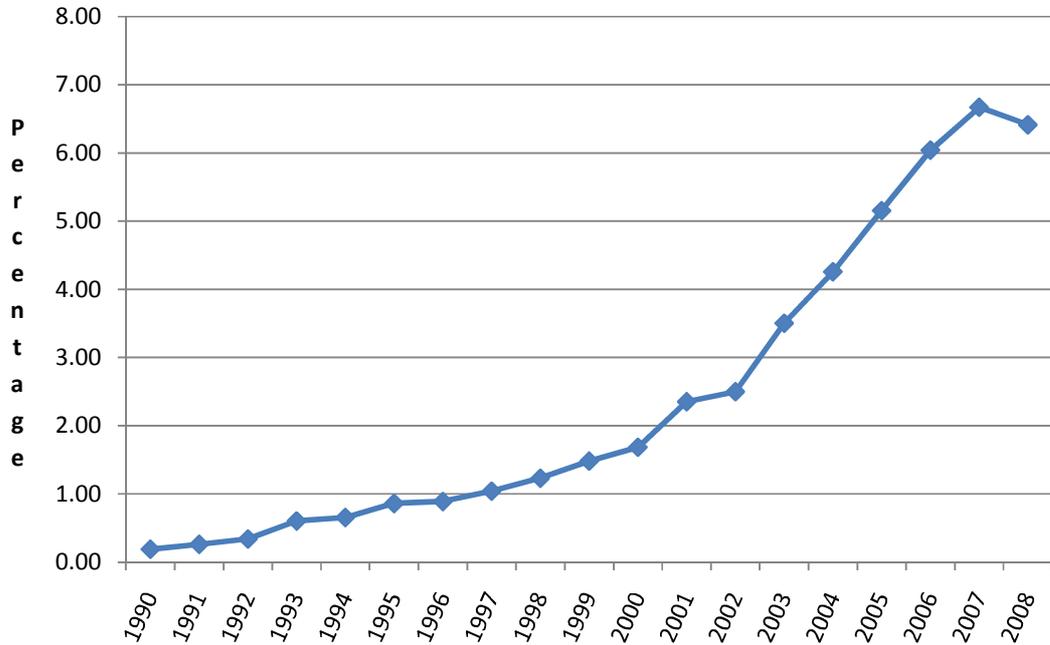


Figure 3. Protein Data Bank (PDR) Dissertations, 2000-2009

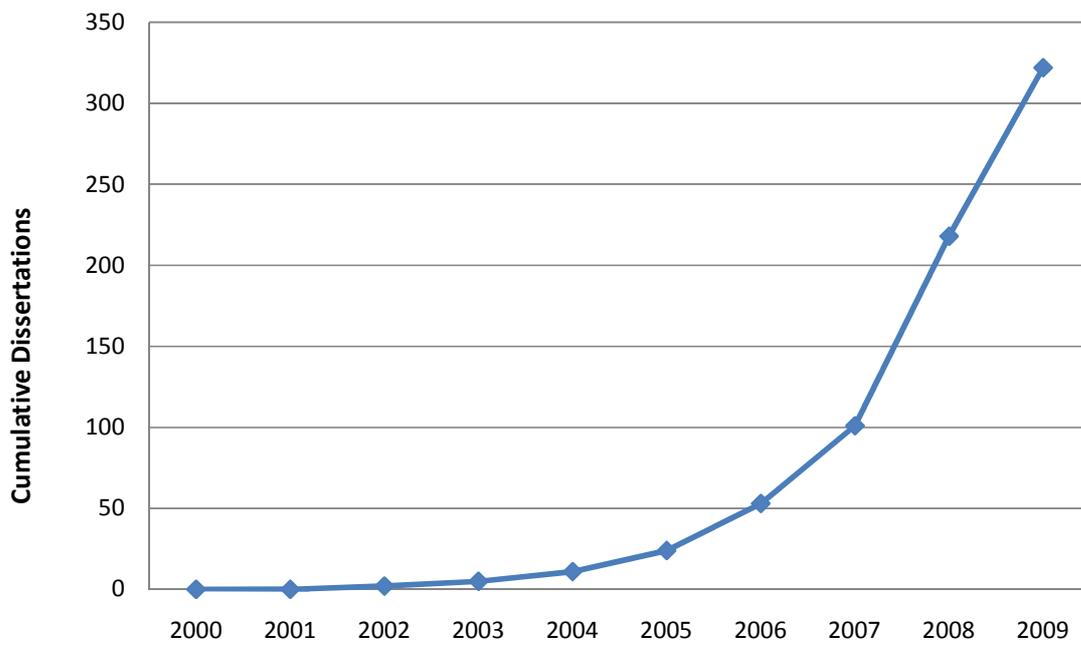


Table 1. Dissertations supervised by Sunny Xie, Harvard University's Department of Chemistry and Chemical Biology that were supported by his 2004 Pioneer Award

Scientist	Doctorate year	Years with Xie	Pubs	Coauthor Xie	New Direction	Present Location
G. Luo	2006	2000-2006	8	8		Indian Institute of Science
C. Long	2006	2001-2006	5	4	Yes	Cal Tech
P. Blainey	2007	2001-2007	6	5	No	Stanford
X. Nan	2007	2001-2007	6	6		Lawrence Berkeley National Lab
C. Evans	2007	2002-2007	13	8	Yes	Photomedicine Center, Harvard
W. Min	2008	2001-2007	14	14		Columbia
P. Choi	2009	2005-2010	5	5		Harvard Med
W. Li	2010	2005-2010	4	3	No	UCSF, Postdoc
P. Sims	2010	2005-2009	2	2		Postdoc with Xie

Source: PubMed.