NBER WORKING PAPER SERIES

THE (CHANGING) KNOWLEDGE PRODUCTION FUNCTION: EVIDENCE FROM THE MIT DEPARTMENT OF BIOLOGY FOR 1970-2000

Annamaria Conti
Christopher C. Liu

Working Paper 20037
http://www.nber.org/papers/w20037

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2014

We are indebted to Adam Jaffe, Ben Jones, Paula Stephan, Marie Thursby, Fabian Waldinger, and seminar participants at the NBER Changing Frontier Conferences (October 2012 and August 2013) for their valuable comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The (Changing) Knowledge Production Function: Evidence from the MIT Department of Biology for 1970-2000
Annamaria Conti and Christopher C. Liu
NBER Working Paper No. 20037
April 2014
JEL No. D2,H41,I2,I20,I28

ABSTRACT

Considerable attention has been focused, in recent years, on the role that graduate and postdoc students play in the production of academic knowledge. Using data from the MIT Department of Biology for the period 1970-2000, we analyze the evolution over time of four fundamental aspects of their productivity: i) training duration; ii) time to a first publication; iii) productivity over the training period; and iv) collaboration with other scientists. We identified four main trends that are common to graduate students and postdocs. First, training periods have increased for later cohorts of graduate and postdoc students. Second, later cohorts tend to publish their initial first-author article later than the earlier cohorts. Third, they produce fewer first-author publications. Finally, collaborations with other scientists, as measured by the number of coauthors on a paper, have increased. This increase is driven by collaborations with scientists external to a trainee's laboratory. We interpret these results in light of the following two paradigms: the increased burden of knowledge that later generations of scientists face and the limited availability of permanent academic positions.

Annamaria Conti
Georgia Institute of Technology
Scheller College of Business
800 West Peachtree Street NW
Atlanta, Georgia 30308
annamaria.conti@mgt.gatech.edu

Christopher C. Liu
Rotman School of Management
University of Toronto
105 St. George Street
Toronto, Ontario
Canada M5S 3E6
chris.liu@rotman.utoronto.ca

I. Introduction

Knowledge has been recognized as a major contributor to technological change and, more generally, to economic growth (Romer, 1990). In the knowledge production function, one of the most important inputs is knowledge created by university researchers. Indeed, a report by the National Science Board (2008) has revealed that university researchers are responsible for more than 70% of all scientific articles. Moreover, scholars have shown that academic knowledge is responsible for a large percentage of industrial innovations (Jaffe, 1989; Mansfield, 1995).

Academic knowledge has increasingly become a collective phenomenon. Seminal studies have documented the increase in the size of scientific collaborations, with special focus on the evolution of the geographic dispersion of team members (e.g. Adams *et al.*, 2005; Wuchty *et al.*, 2007). Even though university scientists collaborate more and more across research institutions, the scientific laboratory remains the major locus of knowledge production (Stephan, 2012b). These laboratories are largely populated by graduate students and postdocs, whose contributions to their laboratory's knowledge stock have been recognized in a number of studies (see, for instance, Stephan, 2012b; Conti *et al.*, forthcoming). These research trainees have coauthored an important percentage of their laboratory's papers and, moreover, have produced a considerable share of the articles published in highly ranked journals (Black and Stephan, 2010).

In this study, we use a unique database that allows us to examine the productivity, training duration, and the collaborative behavior of graduate students and postdocs as well as the extent to which these aspects have evolved over time. We interpret the patterns we find in light of two paradigms: the increased burden of knowledge that successive generations of scientists face (Jones, 2009 and 2010) and the limited availability of permanent academic positions (Stephan, 1996; Freeman *et al.*, 2001).

Our data encompass the complete set of laboratories in the MIT Department of Biology, observed from 1970 to 2000. This department has been a major locus of basic and applied discoveries in the life sciences for the latter half of the 20th century. Through the time frame of our dataset, the scientists working at the MIT Department of Biology made discoveries as varied as the molecular mechanisms underpinning recombinant DNA (e.g., the discovery of splicing and introns), cell death, aging, and the progression of cancer. This work has resulted in six Nobel Laureates and 43 members of the National Academy of Sciences between 1966 and 2000. MIT's Department of Biology has roughly doubled in size, from 27 laboratories in 1966 to 49 laboratories in the year 2000. Given this department's elite status, the findings in this paper may be difficult to extend beyond other elite North American laboratories. With this caveat in mind, we follow in the footsteps of other scholars and trade analytical depth with a focus on an elite setting (Azoulay, 2010; Zuckerman, 1977).

We collected a detailed set of information on the graduate students and postdocs who populated these laboratories, including their publication output. For the purposes of this study, we use this information to analyze the evolution over time of four fundamental aspects of their productivity: i) training duration, ii) time to a first publication, iii) productivity over the training period, and iv) collaboration with other scientists.

We identified four main trends that are common to graduate students and postdocs. First, training periods have increased for later cohorts of graduate students and postdocs. Second, recent cohorts tend to publish their first article later than the earlier cohorts. Third, they produce fewer first-author publications. Finally, collaborations with other scientists, as measured by the number of coauthors on a paper, have increased. This increase is driven by collaborations with scientists outside of a trainee's laboratory.

The remainder of this study is organized as follows. Section II describes the empirical setting. Section III presents the scientific productivity trends for graduate students and postdocs. Section IV concludes and discusses policy implications.

II. Empirical setting

For the period under study, the MIT Department of Biology generated an Annual Report, which serves as our core data source. The primary purpose of the Annual Report was to, internally, distribute information about the department's scientific activities. As a result, the report includes technical summaries of ongoing projects as well as a list of publications produced during the prior year. From 1966-1989, technical summaries were at the project level and individuals could contribute to multiple projects. The size of the Annual Report grew in accordance with the size of the department. After the Annual Report reached 629 pages in 1987, summaries were limited to two pages per laboratory, regardless of its size. Unfortunately, starting in 2001, even the summaries ceased to be published and subsequent data have been lost to posterity.

The Annual Report documents a roster of each laboratory's members. We know the names of every individual in each laboratory as well as the individual's personnel type (e.g., postdoc, graduate student, technician). As a result, we know the characteristics of the department, its laboratories, and its individual members over the course of 35 years. Figure 1 provides an example of the data available for any given laboratory-year. We know of no other data source that provides as detailed a view into the organization of scientific work as this one.

< Insert Figure 1 about here>

We supplemented this departmental personnel roster with a number of other data sources. To examine scientific outputs, we hand collected each principal investigator 's (PI) paper output from Medline. We then matched each publication's author list with our personnel roster to examine the extent to which individual laboratory members contributed to the scientific output. In instances where matching was ambiguous (e.g. Liu), we examined the article directly. It is exceedingly rare for laboratory members to publish scientific papers without their PI listed as an author. Hence we do not believe we are missing any publications.

Overall, our dataset comprises 1,494 laboratory-years and 20,324 laboratory member-years that span 1966-2000. Within this dataset, there are 120 professors and 6,938 laboratory members who collectively produced 7,553 journal publications.

We restrict our analysis to the 1970-2000 period as there was ambiguity in personnel categories prior to 1970. We begin with a description of the laboratories and their changes over time. We then turn our attention to examine the laboratory members with a particular emphasis on two major types, postdocs and graduate students, who comprise more than half of our personnel roster.

Within our dataset, the average laboratory has 10 members of which 5 are postdocs, 3 are graduate students, and 2 are technicians. Staff scientists are rare, but their prevalence has increased over time. As shown in Figure 2, laboratories have grown in size through the latter part of the 20th century, and this increase has been fostered by the number of postdoctoral scientists. There is no change in the number of graduate students or technicians over time, although the number of salaried staff (i.e., technicians and staff scientists) appears to have increased in the late 1990s[4].

< Insert Figure 2 about here>

Figure 3 presents trends in scientific output for our laboratories. As shown, the average number of articles has steadily increased over time, from an average of 4 articles per laboratory-year in the 1970s to 6 articles per laboratory-year in the 1990s. We observe a very similar trend in the number of impact factor-weighted publications.

< Insert Figure 3 about here>

[4] A likely reason why the number of graduate students remained steady over the years is that university departments in the US tend to set a limit to the number of students that can enroll in a PhD program.

We restrict our analysis of laboratory members to graduate students and postdocs for the following reasons. First, these individuals make large contributions to a PI's publication output. Their purpose is to directly produce scientific publications, rather than to play a supporting role (e.g., technicians). Second, these two types are the most prevalent personnel categories within the roster. Together they make up more than half of the laboratory. Third, these two personnel types have been the focus of recent interest in the literature because of their contributions to knowledge and technology production (e.g. Dasgupta and David, 1994; Waldinger, 2010). A quick look at faculty websites convinces one of the importance of these contributions, be it measured by publications, citations, or grants. Lastly, we note that graduate students and postdocs are easily and unambiguously identified, as opposed to less clear categories such as visiting scientists.

Our sample is composed of 991 graduate students and 2,427 postdocs. Figures 4a and 4b provide descriptive results of the distribution of graduate students and postdocs by their publication count. Interestingly, a significant proportion of them (about 35%) did not publish any articles during their training period. Conditioned upon having published, the mean number of papers is about three articles for both graduate students and postdocs.

< Insert Figure 4a about here>

< Insert Figure 4b about here>

III. Trends in scientific productivity of graduate students and postdocs

This section explores the trends in four major dimensions of graduate student and postdoc scientific productivity. First, we look at training duration. Second, we investigate the timing to a first publication. Third, we examine scientific output. Finally, we explore collaboration patterns.

In analyzing these trends, we should keep in mind that while both postdocs and graduate students are formally considered laboratory trainees, they fundamentally differ in a number of aspects. Postdocs are more experienced that graduate students and have accumulated a greater wealth of knowledge and skills. As a consequence, matching between postdocs and PIs is based upon prior ability and experience, rather than the future expectation of productivity as in the case of graduate students (Stephan, 2012b).

A. Training duration

We begin this section by presenting descriptive statistics for the average training duration of postdoc and graduate students over our sample period. We then investigate whether the length of training has changed over time. Figures 5a and 5b show the distribution of graduate students and postdocs by their training duration. The training period for graduate students is longer than postdoctoral training. Indeed, the majority of graduate students in our sample completed their training between five and seven years, while postdocs tended to spend between two and four years in a PI's laboratory[5].

< Insert Figure 5a about here>

< Insert Figure 5b about here>

Figure 6 shows the evolution of training periods for graduate students (in red) and postdocs (in blue) over the period 1970-1995. We exclude the years 1996 through 2000, since students who enrolled in these years might not have completed their training by the end of 2000. In line with previous studies[6], we find that training periods for recent cohorts of students tend to be about one year longer than those for the earliest cohorts. The training period increases from three to approximately four years for postdocs and from five to six years for graduate students over our dataset.

< Insert Figure 6 about here>

[5] It is possible for postdocs to have worked in more than one PI's laboratory before they are offered a faculty position. However, from discussions with MIT PIs as well as from an examination of a CV sample, it is evident that, at least for the period we examine, this is rarely the case for MIT postdocs.

[6] See, for instance, the findings by the National Research Council (1990), Tilghman (1998), Jones (2009), Jones and Weinberg (2011), and Freeman *et al.* (2001).

There are at least three reasons that can explain these trends. The first reason is that as knowledge accumulates, earlier trainee cohorts face a greater educational burden than do the older cohorts (Jones, 2009 and 2010). Second, it is also possible that the recent cohorts of postdocs and graduate students tend to stay longer in their positions because of the increased mismatch between the trainees' supply and the availability of permanent academic positions (Stephan, 1996; Freeman *et al.*, 2001). Finally, one cannot exclude the possibility that the increased pressure on PIs to publish and apply for grants has led them to impose longer training periods on their students (Freeman *et al.*, 2001).

To more formally assess the evolution of training periods over time, we estimate Poisson regression models, with robust standard errors, in which we relate the training duration of graduate students and postdocs to whether these trainees had enrolled during the following periods: *i*) 1970-1979; *ii*) 1980-1989; and *iii*) 1990-1995. The distribution of students across enrollment periods is reported in Table 1.

The equation we estimate is:

$$y_i=\exp(\beta_1 \text{ D1980-1989}+ \beta_2 \text{ D1990-1995}+ \nu_i+ \theta_i +\varepsilon_i)$$

where y_i is training duration, measured in number of years. Moreover, D1980-1989 is an indicator variable that equals one if trainee *i* enrolled during 1980-1989 and equals zero otherwise. D1990-1995 equals one if trainee *i* enrolled during 1990-1995 and, similarly, equals zero otherwise. We omit the 1970-1979 indicator variable and use it as a reference. Hence, the coefficients of β_1 and β_2 should be interpreted as the change in training duration relative to the duration of trainees enrolled in 1970-1979. When investigating training duration, it is important to consider the scientific field in which a laboratory operates (Galison, 1997). Different scientific fields use different tools and it is likely that trends in training durations vary across fields. To account for field effects, we include a series of indicator variables, ν_i, corresponding to the modal experimental organism used in each laboratory. Specifically, we generated indicators for protein biochemists, bacteriologists, unicellular systems (e.g., HeLa cells), genetic systems (e.g., yeast), rodents, and other (e.g., frog). Finally, we include a set of PI dummies, θ_i, to capture variations in duration trends across laboratory heads.

Table 2 presents the regression results for graduate student and postdoc training duration. For each trainee category, we first include biology field fixed effects (column I) and, subsequently, we add PI fixed effects (column II). We begin by describing the results for graduate students and then for postdocs.

As Table 2 shows, in the baseline model, the dummies D1980-1989 and D1990-1995 have a positive and statistically significant coefficient. These results confirm the descriptive evidence that later cohorts of students take longer to complete their PhD than earlier cohorts (cohorts who enrolled during the 1970-1979 period). In the second column, we add PI effects and the magnitude of the coefficients declines together with their statistical significance. This last result suggests that PI characteristics are a source of positive correlation between period dummies and training duration.

We find similar results for postdocs. The coefficients of the 1980-1989 and 1990-1995 period dummies are positive and statistically significant regardless of the model specification, although the magnitude and significance is reduced with PI fixed effects.

To summarize, the results in this section suggest that training periods have increased in recent years for both graduate students and postdocs. While we cannot precisely disentangle the mechanisms behind these trends, we believe that increasing challenges imposed on recent trainees, in terms of increased educational burden or reduced availability of permanent academic positions, may play an important role.

< Insert Table 1 about here>

< Insert Table 2 about here>

B. Time to a first publication

In this section, we focus on the time it takes trainees to publish their first article. We considered the time interval between a trainees' enrollment and their first publication as the time it takes them to acquire the knowledge to develop publishable findings. This interval becomes then a measure of trainee distance to the existing knowledge frontier. Figure 7 presents Kaplan-Meier estimates of the time to a first publication for postdocs and graduate students. As shown, the probability of publishing a paper in each training year appears to be higher for postdocs than for graduate students. This holds true even when we focus exclusively on first-author publications, which we take as a proxy for those projects to which trainees have given their greatest contribution[7].

< Insert Figure 7 about here>

Once more, we are interested in the evolution of time to a first publication over our sample period, for both graduate students and postdocs. If the knowledge burden for the more recent cohorts is larger than that for the oldest ones, then we should expect that the time it takes to publish a first article has increased for the most recent cohorts. There are other reasons to expect such a trend. One of these could be a lengthening of the review process at scientific journals. While this is a documented trend in the economic field (Ellison, 2002), there are grounds for believing that this phenomenon is not confined to economic journals. By way of an example, statistics available for the EMBO journal reveal an increase over time in the number of days from submission to final decision[8].

Figures 8 and 9 display Kaplan-Meier estimates of the time it takes to publish a first article, distinguishing between the following periods: i) 1970-1979; ii) 1980-1989; and iii) 1990-2000. They provide evidence that the probability of publishing a paper at any given period is higher for the oldest cohorts than for the more recent ones. These trends seem to be more accentuated for postdocs than for graduate students. Moreover,

[7] For the sake of brevity, we do not show the results for first-author publications, but they are available upon request.

[8] Statistics are available from http://www.nature.com/emboj/about/process.html

for graduate students, they are more evident in first-author publications than they are in other publications.

< Insert Figure 8 about here>

< Insert Figure 9 about here>

What we need to understand is whether these trends persist once we take into account field or PI characteristics, which are likely to be a source of correlation between enrollment periods and time to a first publication. For this purpose, we estimate a series of Cox proportional hazard models in which the hazard of publishing a first article is a function of our period indicators and controls.

Hence, we estimate the following equation:

$$h(t|x_i) = h_0(t)\exp(x_i\beta_x)$$

where $h(t|x_i)$ is the hazard of publishing a first article, $h_0(t)$ is the baseline hazard (i.e. the hazard when all covariates are equal to zero), and x_i is a matrix of covariates. As in our previous equation, x_i includes period indicator variables as well as field and PI dummies. This time we also include in the sample trainees who had enrolled after 1995. Hence, the last period indicator variable equals one for trainees who had enrolled during 1990-2000 and zero otherwise. The results for graduate students are presented in Table 3, while those for postdocs are in Table 4. Standard errors are clustered around PI.

We begin by presenting the results for graduate students, distinguishing between the time to a first publication and the time to an initial first-author publication. Estimates are presented in terms of their effect on the odds of publishing a first paper. Hence, a coefficient smaller (larger) than one reflects a negative (positive) effect. When we only include field fixed effects, the coefficients of the 1980-1989 and 1990-2000 period dummies are smaller than one, as expected, but not statistically significant. They decrease in magnitude and become significant once we introduce PI fixed effects. This result indicates that trends in the time to a first publication vary across PIs.

When we examine first-author publications, we find stronger evidence that the time to a first publication has increased for later cohorts of graduate students relative to earlier ones. Indeed, the coefficients of both period dummies are smaller than one and the coefficient for the 1990-2000 indicator is statistically significant. The coefficient magnitudes suggest that the hazard of publishing an initial first-author paper, for graduate students who enrolled in the 1980-1989 period, is 0.9 times the hazard of those who enrolled in the 1970-1979 period. It declines to 0.8 times for graduate students who enrolled during 1990-2000. As before, once we introduce PI fixed effects the significance of the coefficients improves and the magnitude declines.

< Insert Table 3 about here>

In the case of postdocs, both the time to a first publication and that to an initial first-author publication appear to have increased for later cohorts relative to earlier ones. Regardless of the regression specification, the hazard of publishing a first paper is lower for postdocs who started in the 1980-1989 period, than for postdocs who enrolled during 1970-1979, and it is lowest for those who started during 1990-2000. Moreover, the coefficients tend to be statistically significant with and without PI fixed effects[9].

< Insert Table 4 about here>

Overall, we provide evidence that the time to an initial first-author publication has increased for both graduate students and postdocs and this result is strongest for trainees in the most recent decade. Moreover, in the case of postdocs, results indicate that the time to a first publication has increased even for non-first author articles. In general, these

[9] In column three the coefficient for the 1980-1989 period dummy is not significant. However, a test of joint significance of period dummies rejects the null hypothesis that they are (jointly) equal to zero with a p-value of 0.00.

results seem to be consistent with our previous findings that training periods have increased over time. Taken together, these results may suggest that, at least in part, recent cohorts of trainees use their extra training time to achieve first publishable results.

C. Publication trends

In this section, we turn our attention to trends in the publication output of graduate students and postdocs. The question we want to explore is whether recent cohorts of graduate students and postdocs have become less productive than older ones. Indeed, if one posits that recent cohorts of scientists face a larger learning burden or that the reviewing process at scientific journals has increased over time, then we should observe a declining trend in the publication output of graduate students and postdocs.

To investigate this hypothesis, we estimate count regression models in which we relate publication outputs that graduate students and postdocs had produced during their training as a function of whether their enrollment year falls within the 1970-1979, 1980-1989, or 1990-1995 periods. We adopt a Poisson specification with robust standard errors. We measure publication output by counting the number of publications from the moment a trainee joins a PI laboratory until two years after the trainee was last observed in the laboratory. In this way, we account for the fact that there are lags between the moment a research project is completed and the moment its results are published. As for the analysis of training durations, we exclude the latest years because graduate students and postdocs who enrolled in these years might not have completed their training by the end of our sample period.

The equation we estimate is:

$$y_i=\exp(\beta_1\ D1980\text{-}1989+\ \beta_2\ D1990\text{-}1995+\ \beta_3\ \mathrm{Duration}_i+$$
$$+\nu_i+\ \theta_i\ +\varepsilon_i)$$

where y_i is either the total count of trainee *i*'s publications or the count of their first-author publications. D1980-1989 is an indicator variable that equals one if trainee *i* enrolled during 1980-1989 and equals zero otherwise. D1990-1995 equals one if trainee *i* enrolled during 1990-1995 and, similarly, equals zero otherwise. $Duration_i$ is defined as the number of years a trainee has spent in a laboratory. Finally, ν_i and θ_i are field and PI fixed effects, respectively.

The results for graduate students are displayed in Table 5, while those for postdocs are presented in Table 6. When we consider the total publication count (column I), we find that graduate students who enrolled in more recent periods are no less productive than their colleagues who enrolled during 1970-1979. In fact, none of the coefficients for the 1989-1990 and 1990-1995 period dummies are statistically significant. Once we include supervisor fixed effects, the coefficient of the dummy for student enrollment during 1990-1995 becomes statistically significant and has a negative sign. While this last result suggests that there are some supervisor characteristics that are correlated with productivity trends, we cannot conclude that there is a general declining tendency in the graduate student paper count. In support of this conjecture, descriptive evidence reported in Figure 10 does not reveal a decreasing trend for the annual publication count. In regressions, not reported here (but available upon request), we find very similar results when we use the impact-factor weighted publication count as the output measure.

We show different findings when analyzing first-author publications. In this case, both period dummies have a negative coefficient and that the coefficient for the 1990-1995 period variable is significant, regardless of whether we include PI fixed effects. One might wonder whether this effect is driven by the fact that fewer graduate students are publishing first-author papers in recent years. To investigate this possibility, we estimate a linear probability model in which the dependent variable is an indicator that takes a value of one if graduate students have published at least one article during their training. The results are displayed in the last column of Table 5. The coefficient for the 1990-1995 period dummy is negative and statistically significant, independent of the regression specification. These results suggest that at least part of the declining output trend is explained by a lower publishing probability for the most recent cohorts. Overall, we find that later graduate student cohorts produce fewer first-author articles than earlier ones and, this time, regression results seem to be supported by descriptive evidence reported in Figure 10.

< Insert Table 5 about here>

< Insert Figure 10 about here>

When we turn our attention to postdocs (Table 6), we find strong evidence that the postdoc cohorts enrolled during 1980-1989 and 1990-1995 produce less articles than cohorts enrolled during 1970-1979. This result holds true regardless of whether we look at total or first-author publication counts. Indeed, the coefficients of our period dummies are negative and statistically significant, with and without PI fixed effects. When we analyze the probability of publishing at least one first-author paper, we find that part of the declining trend for the first-author paper count is explained by a lower publishing probability for the most recent cohorts. Overall, these findings are consistent with the descriptive trends presented in Figure 11, which shows an over-time decline in publication outputs by postdoc students.

< Insert Table 6 about here>

< Insert Figure 11 about here>

In analyses not presented here, for the sake of brevity, we attempted to analyze whether the decline in the number of first-author graduate student publications was correlated with larger time intervals between papers, for subsequent publications. Thus we estimated hazard models for publishing a second first-author paper, conditioned on having published an initial one, and for publishing a third first-author paper, conditioned on having published a second. Because we have annual data, we cannot analyze the time interval between two papers published in the same year. With this caveat in mind, we find that the time intervals between first-author publications, subsequent to the first, are not larger for the most recent graduate student cohorts. This seems to suggest that the decline in the number of first-author papers for graduate students could be explained by the fact that trainees take longer to publish a first article or they publish fewer articles per year. Similar results were obtained when we estimated the hazard that postdoc students publish a paper or a first-author paper, conditioned on an initial publication.

To summarize, the results from this section lead us to infer that when we measure graduate student productivity by their first-author publication count, later cohorts appear to be less productive than earlier ones. As for postdocs, recent cohorts appear to be less productive in terms of both first-author and total paper counts.

D. Collaboration trends

We have analyzed the training period and productivity trends of postdoc and graduate students in light of the challenges that recent cohorts of scientists face relative to later ones. The question remaining to be answered is whether trainees have reacted to these challenges by working in larger teams, in a similar fashion to other researchers.

The benefits of teamwork have been extensively discussed in the economics literature and include output gains derived from labor specialization (Becker and Murphy, 1992) and from the circulation of new ideas among team members (Adams *et al.*, 2005). In the economics of science, scholars have found that scientists increasingly work in teams (Zuckerman and Merton, 1973; Wuchty *et al.*, 2007)[10] and that team size has expanded over time (Adams *et al.*, 2005), largely due to an intensification of multi-university collaborations (Jones *et al.*, 2008).

Figure 12 reports trends over time in the average number of coauthors per paper, distinguishing between postdocs and graduate students. In line with previous studies, we observe that for both trainee categories the average number of coauthors per paper has increased over time from approximately 1.5 at the beginning of the 1970s to approximately 3.5 by the second half of the 1990s. Interestingly enough, we also observe that the increased collaboration size was mainly driven by an increase in the number of outside laboratory coauthors.

[10] See also Agrawal and Goldfarb (2008) and Forman and Van Zeebroeck (2012).

< Insert Figure 12 about here>

Overall, this suggests that trainees, similar to other scientists across a broad range of disciplines, are increasingly working in teams and these teams tend to encompass authors from outside the trainees' laboratories.

IV. Conclusions and policy implications

A. Summary

While knowledge production is considered one of the main determinants of economic growth, there is no doubt that academic knowledge is one of the most decisive inputs in the knowledge production function representing by far the largest source of codified knowledge.

This study focuses on the contributions to academic knowledge by postdocs and graduate students. Using data from the MIT Department of Biology from 1970 to 2000, we look at the evolution of four fundamental aspects of their productivity: i) training duration, ii) time to a first publication, iii) productivity over the training period, and iv) collaboration with other scientists.

We identified four main trends that are common to graduate students and postdocs. First, training periods have increased for later cohorts of research trainees. Second, recent cohorts tend to publish their initial first-author article later than the earlier cohorts. Third, they produce fewer first-author publications. Finally, collaborations with other scientists, as measured by the number of coauthors on a paper, have increased. This increase is driven by collaborations with scientists outside of a trainee's laboratory.

B. Interpreting the results

What are the mechanisms that drive our results? Our findings are consistent with Jones' educational burden story (Jones, 2009, 2010), which states that, as knowledge accumulates, future generations of scientists require a greater effort to stand on a giant's shoulders. Hence, they can either make a greater effort or they can specialize in a narrower field and collaborate with other scientists. Our first three results –longer training periods, longer time to publish, lower productivity for later trainee cohorts– could be interpreted as an indication that the knowledge burden has increased. The final result regarding increased trainee collaboration provides an indication that these cohorts have become more specialized.

While the educational burden story is indeed a compelling explanation, we nevertheless think that other mechanisms might also be responsible for our results. One of these mechanisms is the mismatch between the supply of trainees and the availability of post-training academic positions that scholars have discussed in recent decades (Stephan, 2012a; Freeman *et al.*, 2001). Data from the NSF-NIH Survey of Graduate Students & Postdoctorates in Science and Engineering, shows that enrollment into PhD life science programs has increased by 80% between 1972 and 2005[11]. While we do not have information on the availability of post-training positions, it is plausible that selection into (desirable) postdoctoral positions has become harder over time. Lastly, we also should note that longer training periods certainly benefit and are encouraged by PIs. Specifically, many PIs are reluctant to allow their most productive laboratory members (i.e., high-tenure trainees) to depart. In fact, their compensation is, increasingly, assigned according to the rules of a tournament model in which trainee contributions have become key to making discoveries, first (Freeman *et al.*, 2001).

If market frictions were to be responsible for longer training periods, should we also expect them to explain the lower productivity of recent trainee cohorts and their increased propensity to work in collaboration with other scientists? Is it plausible to posit that market disequilibria last for decades? Why is the market not redirecting the excess supply of trainees to other fields?

To answer the first question, one might consider that the excess supply of scientists has led to an increase in academic journal submissions, without a corresponding

[11] Data is available from https://webcaspar.nsf.gov/.

increase in the number of publications. If there is an excess supply of submissions, then the direct consequence is that publishing becomes more difficult, which might explain the lower productivity of recent trainee cohorts. Moreover, specialization and collaboration become ways of dealing with market disequilibria and one wonders whether the reduction in recent cohort productivity could have been even more accentuated had recent trainees not worked with other scientists. This mechanism is not necessarily in contrast with the educational burden explanation, rather, it offers a complementary perspective. In fact, market imbalances might act as a stimulus for scientists to expand the knowledge frontier in order to publish, thus increasing the burden on future generations.

While the mechanisms we have highlighted seem to be plausible, one cannot exclude the possibility that the mismatch between the supply of trainees and the availability of academic positions has led the most brilliant students to shy away from careers in the life sciences. Thus, the increase in training periods and the reduced productivity of the most recent cohorts is a reflection of their lower quality skills.

To answer the second and third questions regarding the duration of market imbalances, we should refer to studies by Freeman *et al.* (2001) and Stephan (2012b) and mention that, increasingly, PhD programs in life science, among others, tend to be populated by foreign students. Indeed, while domestic students might be discouraged from continuing their studies in the life sciences PhD programs, these remain attractive to foreign students not only because of their prestige, but also because salary differentials between foreign countries and the US are typically large. Clearly, if the average salary of a PhD holder in Italy is about $2,000 per month, then Italian students will be attracted by a US graduate degree because by the conclusion of their studies, they will potentially earn more than they would have earned at home. To verify that the proportion of foreign graduate students in the MIT Department of Biology has increased over time, we examined our trainees' first and last names. We then codified those who had a Chinese last name as well as those with an Italian or French first and last name[12]. We found that the proportion of Asian, Italian, or French students has increased from 17% in 1970 to

[12] Given the authors' backgrounds, we found it easiest to codify these student ethnicities.

27% in 1995. While these figures are only suggestive, given that we cannot distinguish between foreign or native-born students, they seem to provide an indication that foreign trainees have recently become an important proportion of the graduate student population. There are important policy implications arising from the interpretations of our results. We will discuss them below.

C. Policy Implications

Ultimately, this paper has served to document the mechanisms underlying two important trends in the scientific community: the increasing duration of scientist trainees and an increasing propensity for collaborative activity (e.g., Agrawal et. al, this volume; Tilghman, 1998). Additionally, we have provided evidence of a decline in the scientific output of recent trainees. What implications do these trends have for the scientific community?

First, regardless of the reasons for the observed trends, it is important to note that the costs of science have increased (Jones, 2011). These costs are paid by the individual, who must endure longer training and uncertain future prospects, as well as by society at large, which does not recuperate the returns from its investment. As previous scholars have highlighted (Jones, 2011; Stephan 2012a), costs can be reduced by ensuring that graduate students and postdocs receive adequate pedagogical support during their training period. This, in turn, improves the efficiency of trainee learning and may serve to offset increases in learning burdens. Moreover, decision makers could cap the trainee teaching load, thereby ensuring that the majority of their time is dedicated to research.

It is also worth mentioning that, as the pre-PI career path for life scientists has become incredibly long, talented scientists may increasingly choose to opt out. Our data show that trainee duration creeps above 10 years and this evidence is not unique to the MIT Department of Biology and to elite institutions, in general (Stephan, 2012b). Longer training duration raises the opportunity costs of a scientific career and makes other occupations more attractive. After all, if employment in other fields entails shorter training periods, lower uncertainty and higher salaries, why would the most brilliant minds opt for a career in the life sciences?

The increase in the opportunity costs of a life science career is likely to affect women more severely than men, further exacerbating issues of female participation in the sciences (Ding *et al*., 2006). Women's participation in academia has been found to be very sensitive to considerations such as family constraints and career uncertainty (Kaminski and Geisler, 2006). Certainly, longer training durations do not help ease these concerns.

We conclude with a final important issue that has attracted the attention of recent scholars, namely the allocation of research credit in collaborations (Bikard *et al*., 2013). Working in teams entails a tradeoff. On the one hand, teamwork seems to produce more knowledge breakthroughs than solo work (Singh and Fleming, 2010). On the other, it involves costs, some of which are related to the assessment of the team members' contributions (Dasgupta and David, 1994). This tradeoff is especially relevant for trainees given that access to tenure-track positions requires that they be able to prove their ability to conduct impactful independent research.

References

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific Teams and Institutional Collaborations: Evidence from US universities, 1981–1999. *Research Policy*, 34(3), 259-285.

Agrawal, A., & Goldfarb, A. (2008). Restructuring Research: Communication Costs and the Democratization of University Innovation. *American Economic Review*, 98(4), 1578-1590.

Agrawal, A., McHale, J., & Oettl, A. (Forthcoming). Collaboration, Stars, and the Changing Organization of Science: Evidence from Evolutionary Biology. *NBER Changing Frontier Volume.*

Azoulay, P., Zivin, J. G., & Wang, J. (2010). Superstar Extinction. *Quarterly Journal of Economics,* 125(2), 549-589.

Becker, G. S., & Murphy, K. M. (1992). The Division of Labor, Coordination Costs, and Knowledge. *The Quarterly Journal of Economics*, 107(4), 1137-1160.

Bikard, M., Murray, F., & Gans, J. (2013). Exploring Tradeoffs in the Organization of Scientific Work: Collaboration and Scientific Reward. NBER Working Paper*,* No. 18958.

Black, G. C., & Stephan, P. E. (2010). The Economics of University Science and the Role of Foreign Graduate Students and Postdoctoral Scholars. In *American universities in a global market* (pp. 129-161). University of Chicago Press.

Conti, A., Denas, O., & Visentin , F. (Forthcoming). Knoweldge Specialization in PhD Students Groups. *IEEE Transactions on Engineering Management.*

Dasgupta, P., & David, P. A. (1994). Toward a New Economics of Science. *Research Policy*, 23, 487–521.

Ding, W.W., Murray, F. & Stuart, T.E. (2006). Gender Differences in Patenting in the Academic Life Sciences. *Science*, 313 (5787), 665-667.

Ellison, G. (2002). The Slowdown of the Economics Publishing Process. *Journal of Political Economy*, 110 (5), 947-993.

Forman, C., & Van Zeebroeck, N. (2012). From Wires to Partners: How the Internet has Fostered R&D Collaborations within Firms. *Management Science*, 58(8), 1549-1568.

Freeman, R., Weinstein, E., Marincola, E., Rosenbaum, J., Solomon, F. (2001). Competition and Careers in Biosciences. *Science*, 294, 2293-2294.

Galison, P. (1997). Image and Logic: A Material Culture of Microphysics. University of Chicago Press.

Jaffe, B. A. (1989). Real Effects of Academic Research. *American Economic Review*, 79(5), 957-970.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322(5905), 1259-1262.

Jones, B. F. (2009). The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder. *Review of Economics and Statistics*, 76(1), 283-317.

Jones, B. F. (2010). Age and Great Invention. *Review of Economics and Statistics*, 92(1), 1-14.

Jones, B. F., & Weinberg, B. A. (2011). Age Dynamics in Scientific Creativity. *Proceedings of the National Academy of Sciences*, 108(47), 1-5.

Jones, B. F. (2011). As Science Evolves, How Can Science Policy?. In *Innovation Policy and the Economy, Volume 11* (pp. 103-131). University of Chicago Press.

Kaminski, D, & Geisler C. (2006). Survival Analysis of Faculty Retention in Science and Engineering by Gender. *Science*, 335 (6070), 864-866.

Mansfield, E. (1995). Research Underlying Industrial Innovations: Sources, Characteristics, and Financing. *Review of Economics and Statistics*, 77(1), 55-65.

National Science Board (2008). Science and Engineering Indicators. Arlington, VA: NSB.

National Science Foundation (2005). Industrial Research and Development Information System, Table H-19, www.nsf.gov/sbe/srs/iris/start.cfm.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, 71-102.

Singh, J. & Fleming, L. (2010). Lone Inventors as Sources of Breakthroughs: Myth or Reality? *Management Science*, 56(1), pp.41–56.

Stephan, P. (1996). The Economics of Science. *Journal of Economic Literature*, 34(3), 1199-1235.

Stephan, P. (2012a). Perverse Incentives. *Nature*, 484, 29-31.

Stephan, P. (2012b). How Economics Shapes Science. Cambridge MA: Harvard University Press.

Tilghman, S. (1998). Trends in the Early Careers of Life Sciences. Report by the Committee on Dimensions, Causes, and Implications of Recent Trends in the Careers of Life Scientists, National Research Council. Washington DC: National Academy Press.

Waldinger, F. (2010). Quality Matters: The expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany. *Journal of Political Economy*, 118(4), 787-831.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827), 1036-1039.

Zuckerman, H. (1977). Scientific Elite: Nobel Laureates in the United States. New York, NY: Free Press

Figure 1: Personnel composition of Professor Baltimore's laboratory

Professor:	David Baltimore
Visiting Scientists:	Samuel Latt and Richard Van Etten
Postdoctoral Associates:	Brygida Berse, Mark Feinberg, Michael Lenardo, Jing-Po Li, Shiv Pillai, Louis Staudt and Xiao-Hong Sun
Postdoctoral Fellows:	Raul Andino, Patrick Baeuerle, Andre Bernards, Lynn Corcoran, Sunyoung Kim, Towia Libermann, Ricardo Martinez, Mark Muesing, Cornelis Murre, Jacqueline Pierce, Stephen Smale, Didier Trono, Anna Voronova and Astar Winoto
Technical Assistants:	Ann Gifford, Carolyn Gorka, Patrick McCaw, Michael Paskind and Gabrielle Rieckhof
Graduate Students:	George Daley, Peter Jackson, Marjorie Oettinger, David Schatz and Dan Silver
Undergraduate Student:	Anna Kuang

Figure 2: Number of laboratory's personnel by type

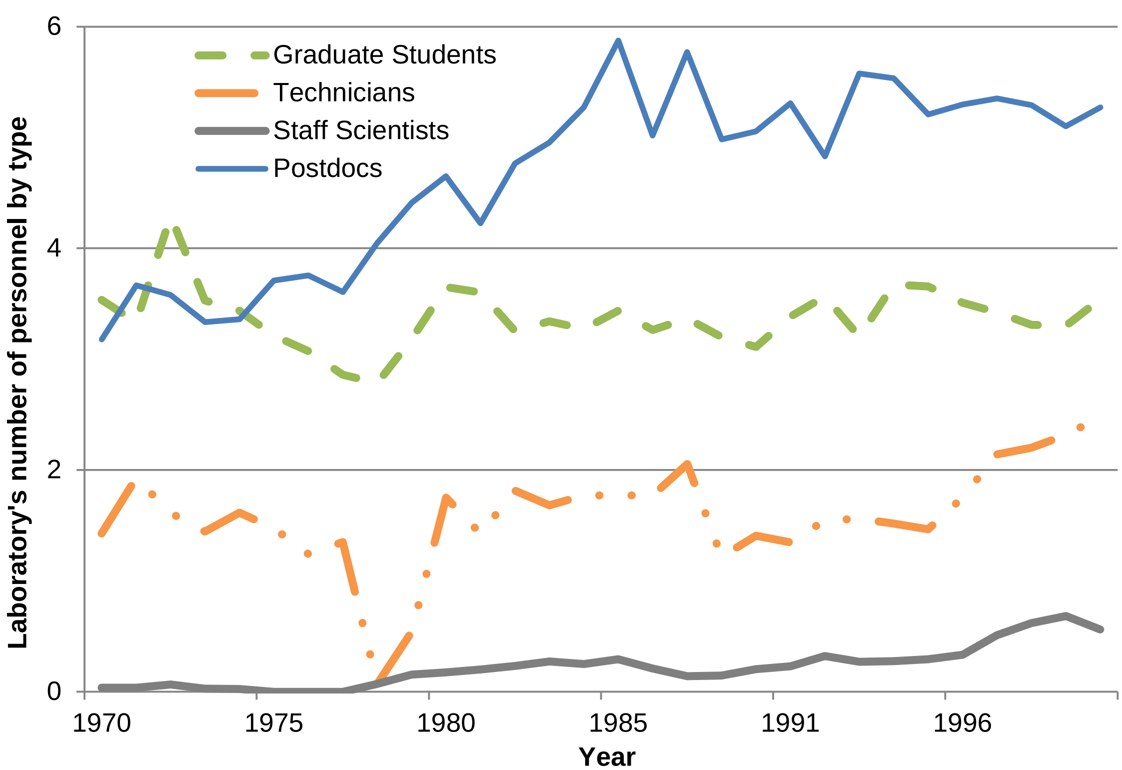


Figure 3: Number of laboratory's publications and impact factor-weighted publications

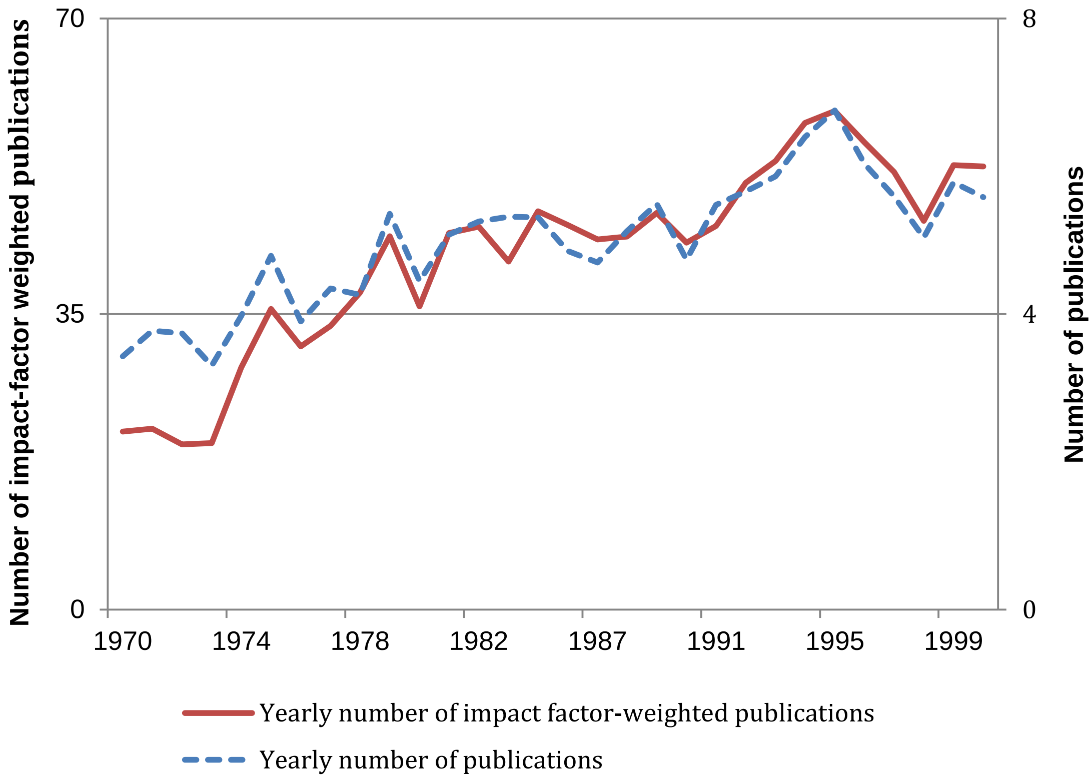


Figure 4a: Distribution of graduate students by their number of papers

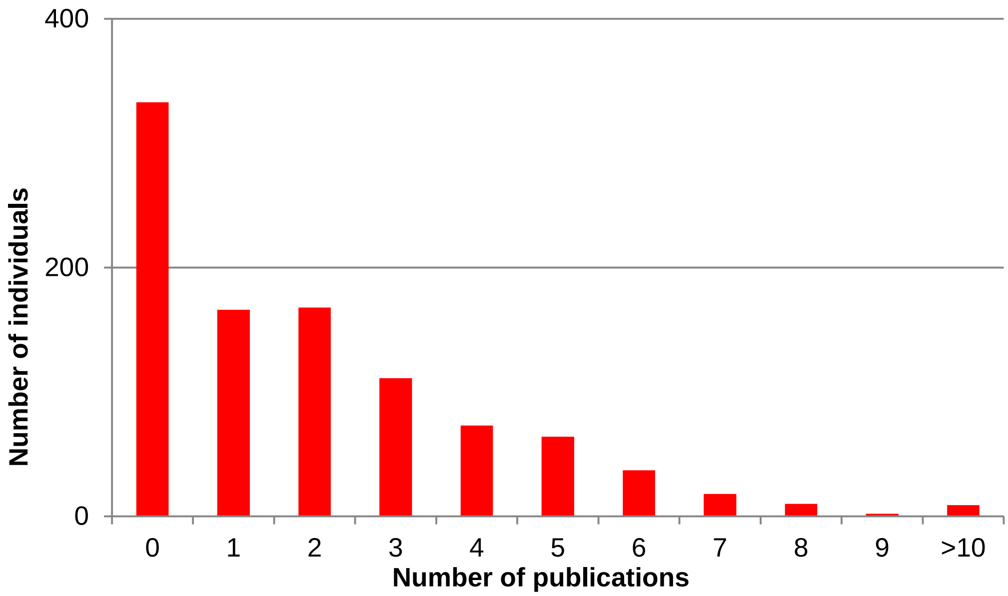


Figure 4b: Distribution of postdocs by their number of papers

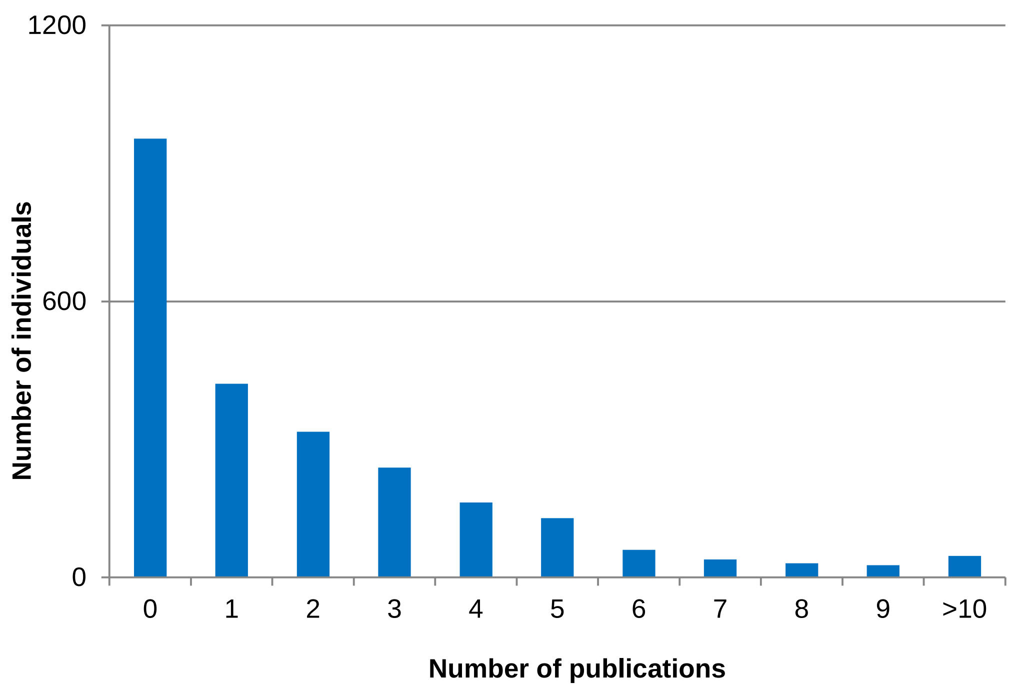


Figure 5a: Distribution of graduate students by their training duration

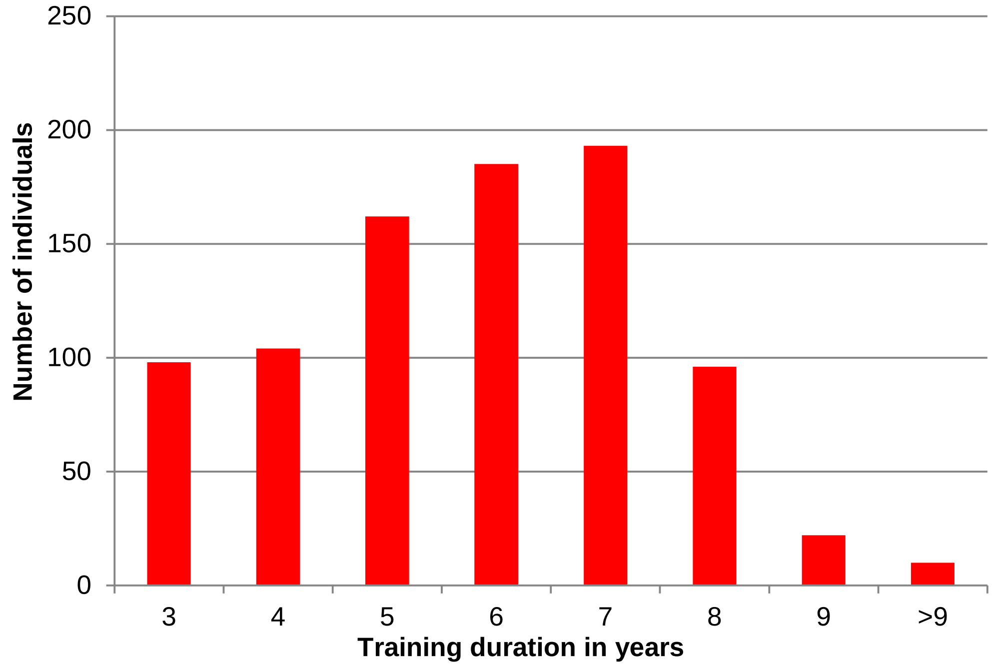


Figure 5b: Distribution of postdocs by their training duration

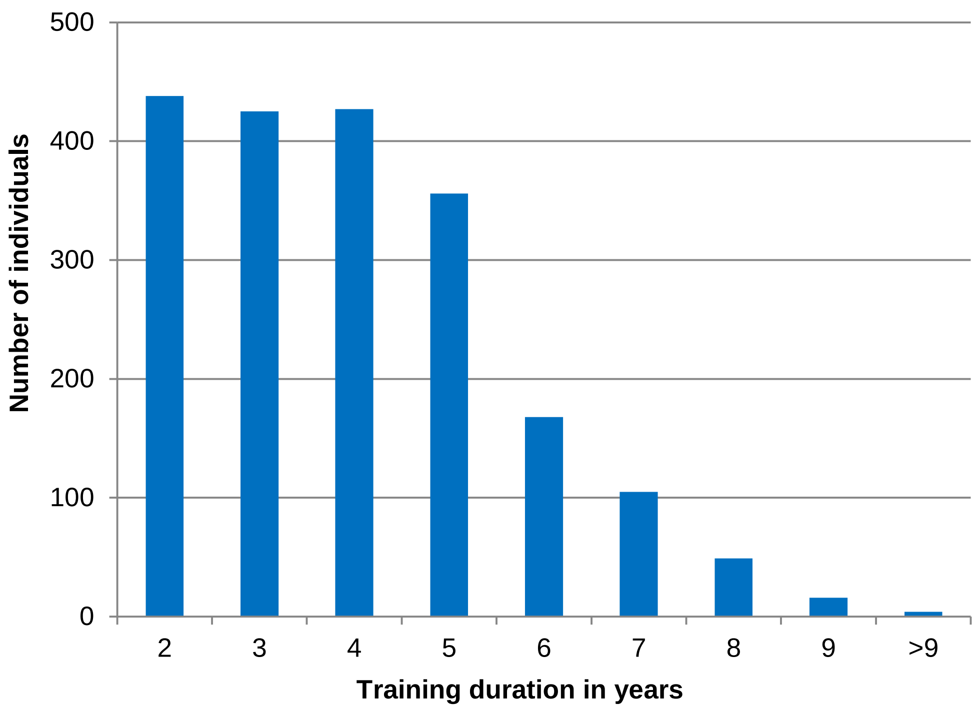


Figure 6: Training duration for graduate students and postdocs over time

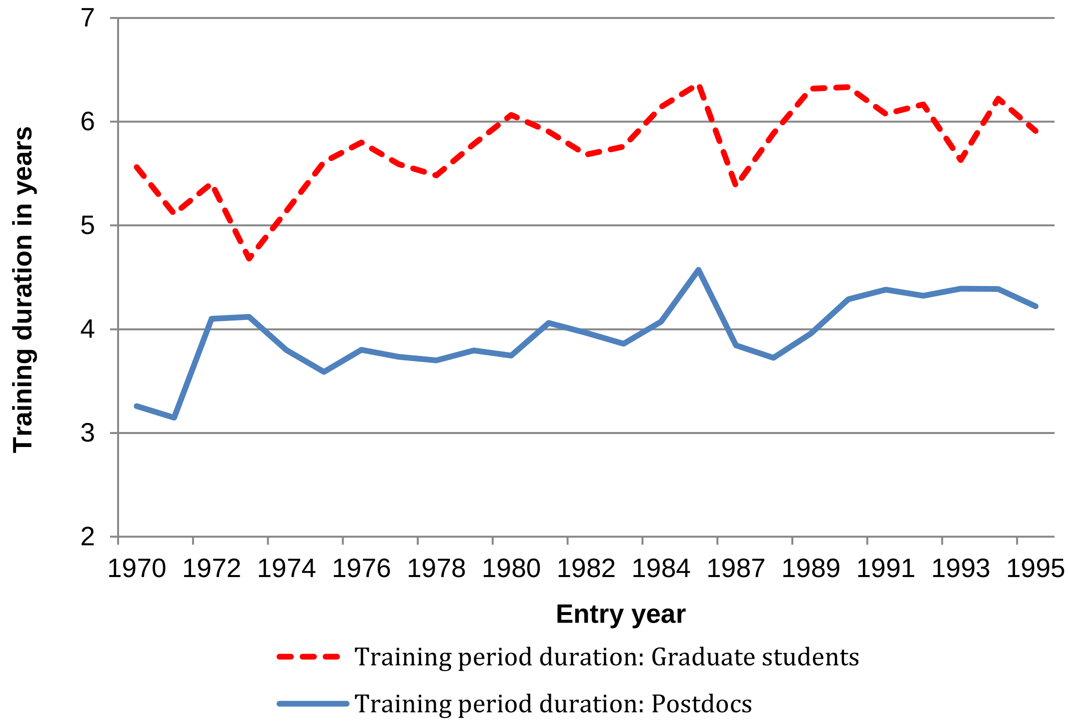


Table 1: Distribution of graduate students and postdocs by enrollment period

	Graduate students	Postdocs
1970-1979	289	560
1980-1989	334	868
1990-1995	247	565
1996-2000	121	434

Table 2: Regression results for graduate student and postdoc training duration

	Graduate students		Postdocs	
	Coeff.	*Coeff.*	*Coeff.*	*Coeff.*
D1980-1989	0.103***	0.075**	0.111***	0.065*
	(0.028)	(0.033)	(0.032)	(0.036)
D1990-1995	0.128***	0.055	0.209***	0.143***
	(0.029)	(0.039)	(0.034)	(0.041)
Field FE	YES	YES	YES	YES
PI FE		YES		YES
R^2	0.01	0.04	0.01	0.03
N	870		1993	

Note: We estimated Poisson models. Robust standard errors are in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.10. For these analyses we only consider trainees who had enrolled before 1996.

Figure 7: Kaplan-Meier estimates of the time to a first publication: graduate students and postdocs

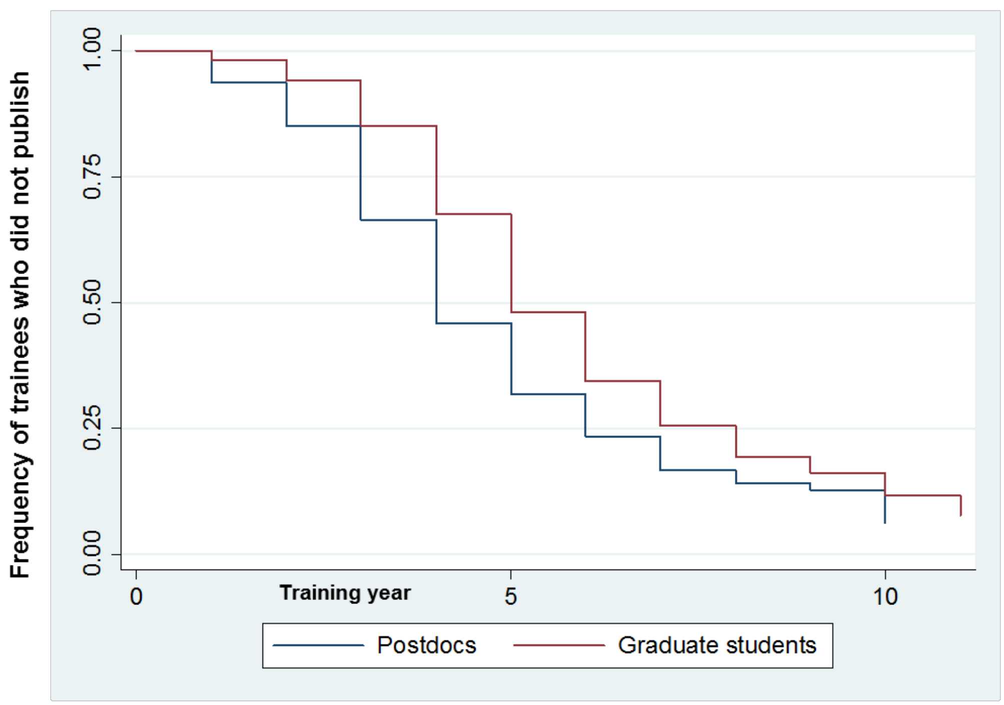


Figure 8: Kaplan-Meier estimates of the time to a first publication: graduate students over time

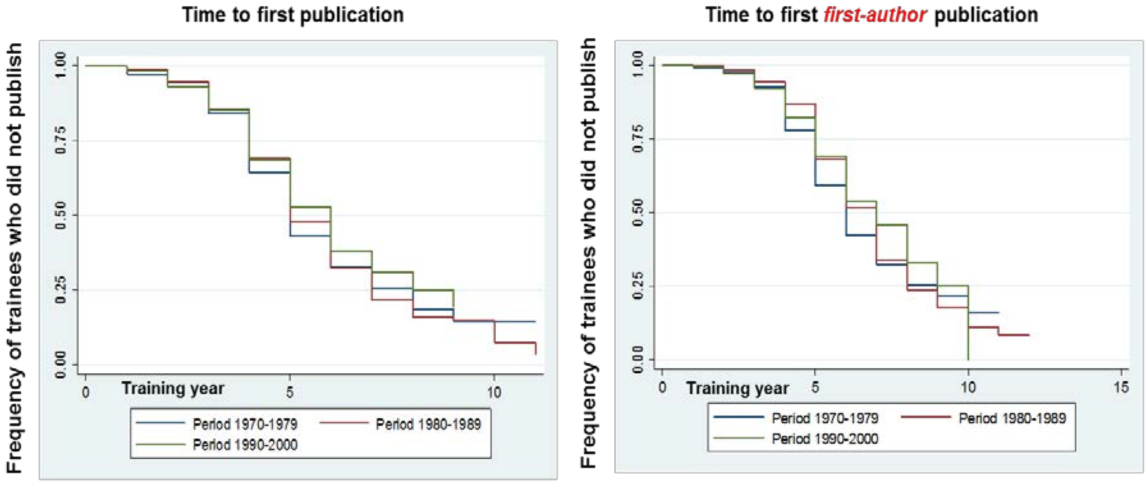


Figure 9: Kaplan-Meier estimates of the time to a first publication: postdocs over time

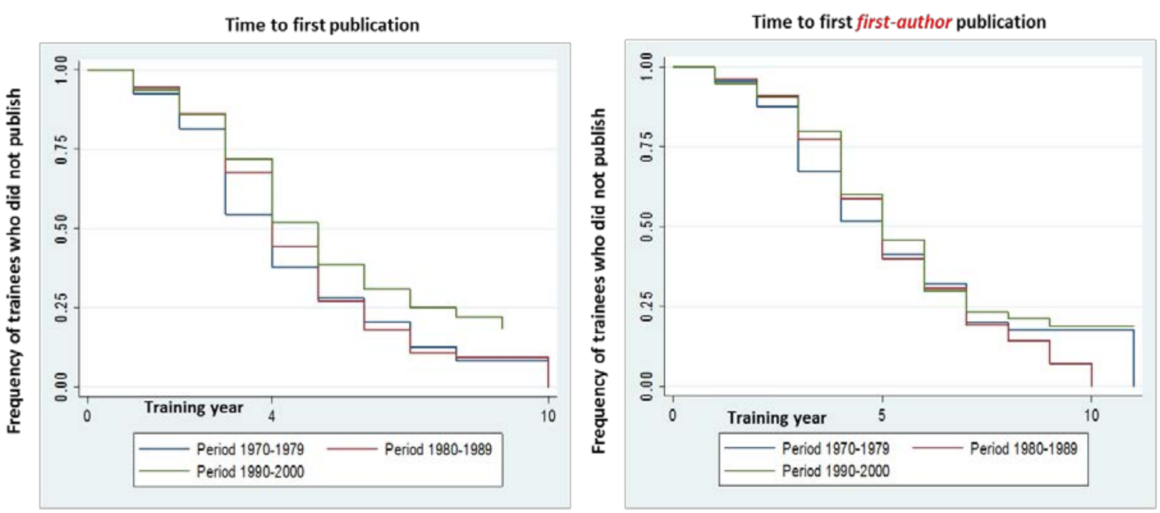


Table 3: Hazard models for the time to a first publication: graduate students over time

	Any publication		First-author publications	
	Hazard Ratios	*Hazard Ratios*	*Hazard Ratios*	*Hazard Ratios*
D1980-1989	0.969	0.768**	0.888	0.718***
	(0.121)	(0.095)	(0.110)	(0.091)
D1990-2000	0.837	0.650***	0.780*	0.613***
	(0.110)	(0.103)	(0.099)	(0.081)
Field FE	YES	YES	YES	YES
PI FE		YES		YES
Log likelihood	-4121	-4042	-3467	-3380
N	991			

Note: We estimate Cox proportional hazards models with standard errors clustered around PI. We report hazard ratios. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 4: Hazard models for the time to a first publication: postdocs over time

	Any publication		First-author publications	
	Hazard Ratios	*Hazard Ratios*	*Hazard Ratios*	*Hazard Ratios*
D1980-1989	0.850***	0.788***	0.862	0.795**
	(0.061)	(0.056)	(0.083)	(0.075)
D1990-2000	0.665***	0.615***	0.658***	0.602***
	(0.061)	(0.061)	(0.071)	(0.062)
Field FE	YES	YES	YES	YES
PI FE		YES		YES
Log likelihood	-10583	-10478	-8626	-8517
N	2427			

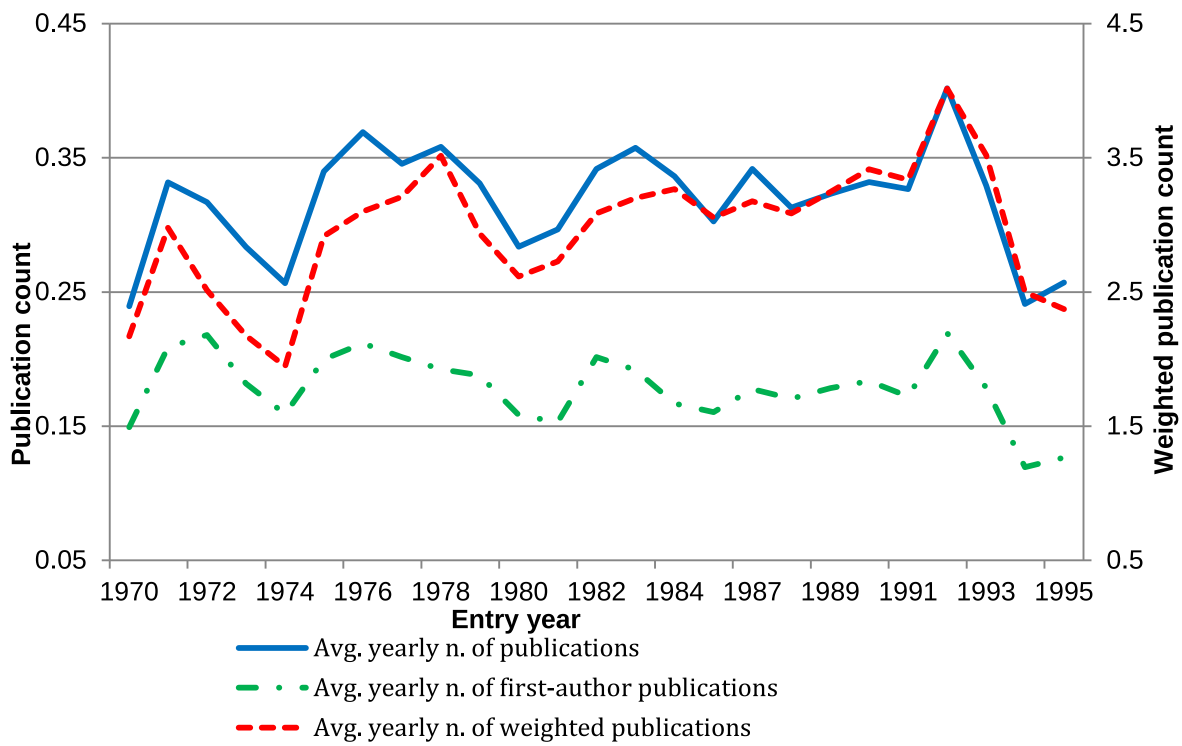
Note: We estimate Cox proportional hazards models with standard errors clustered around PI. We report hazard ratios. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 5: Regression results for graduate student publications

	# Publications		# First-author publications		Probability of publishing a first-author publication	
	Coeff.	*Coeff.*	*Coeff.*	*Coeff.*	*Coeff.*	*Coeff.*
D1980-1989	0.022	-0.103	-0.114	-0.241**	-0.009	0.071
	(0.083)	(0.103)	(0.087)	(0.104)	(0.038)	(0.048)
D1990-1995	-0.071	-0.257**	-0.221**	-0.499***	-0.084**	-0.203***
	(0.090)	(0.120)	(0.096)	(0.130)	(0.041)	(0.059)
Duration	YES	YES	YES	YES	YES	YES
Field FE	YES	YES	YES	YES	YES	YES
PI FE		YES		YES		YES
R^2	0.05	0.15	0.03	0.12	0.105	0.28
N	870					

Note: Standard errors are in parentheses. For the Poisson models we use robust standard errors, while for the linear probability model we cluster standard errors around PI. *** p < 0.01, ** p < 0.05, * p < 0.10. For these analyses we only consider trainees who had enrolled before 1996.

Figure 10: Publication output of graduate student cohorts



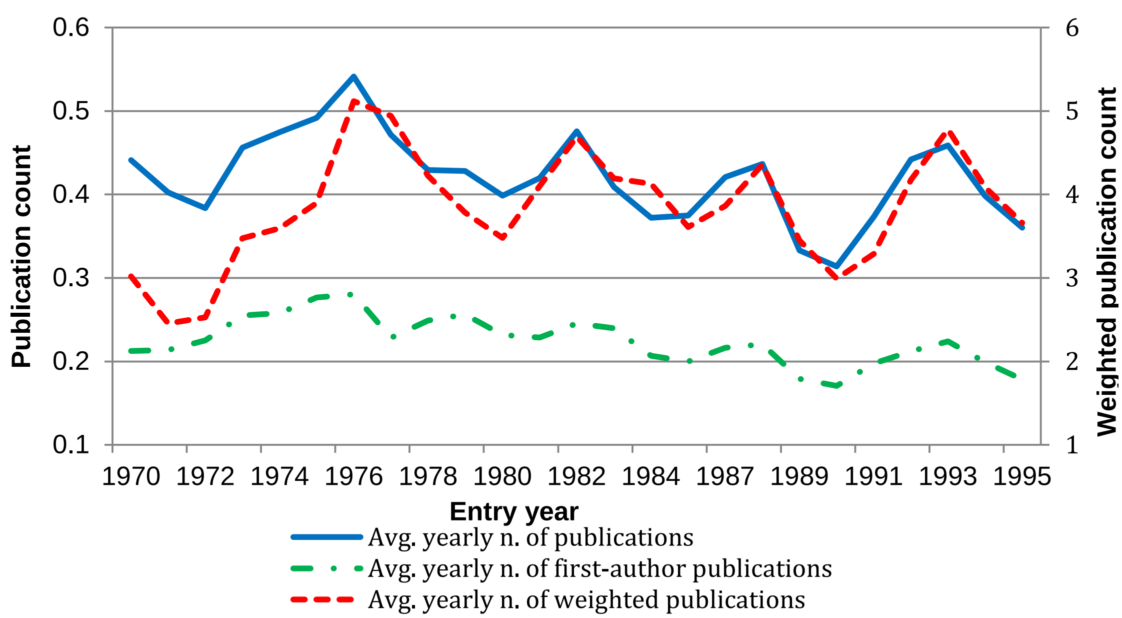
Note: Counts normalized by duration

Table 6: Regression results for postdoc publications

	# Publications		# First-author publications		Probability of publishing a first-author publication	
	Coeff.	*Coeff.*	*Coeff.*	*Coeff.*	*Coeff.*	*Coeff.*
D1980-1989	-0.160**	-0.250***	-0.174**	-0.255***	-0.018	0.019
	(0.067)	(0.071)	(0.068)	(0.074)	(0.026)	(0.029)
D1990-1995	-0.173**	-0.314***	-0.238***	-0.384***	-0.064**	-0.076**
	(0.075)	(0.086)	(0.076)	(0.089)	(0.028)	(0.035)
Duration	YES	YES	YES	YES	YES	YES
PI FE	YES	YES	YES	YES	YES	YES
Entry Year FE		YES		YES		YES
R^2	0.10	0.18	0.08	0.14	0.22	0.23
N	1993					

Note: Standard errors are in parentheses. For the Poisson models we use robust standard errors, while for the linear probability model we cluster standard errors around PI. *** p < 0.01, ** p < 0.05, * p < 0.10. For these analyses we only consider trainees who had enrolled before 1996.

Figure 11: Publication output of postdoc cohorts



Note: Counts normalized by duration

Figure 12: Average yearly number of coauthors per paper