Knowledge Specialization in PhD Student Groups *

Annamaria Conti †, Olgert Denas ‡and Fabiana Visentin §

September 19, 2013

Abstract

Researchers have argued that specialization within groups yields productivity gains. We evaluate this statement with a focus on groups of PhD students. Using an established technique in computer science called Latent Dirichlet Allocation, we construct a novel measure of the dispersion of PhD students' research interests based on their dissertation abstracts. We then relate this measure to PhD group publications. For our study, we use a rich dataset on groups of PhD students who studied at a major Swiss university, during the 1993-2008 period. We find robust evidence that within-group knowledge specialization is associated with a larger number of publications. However, when specialization increases beyond a critical level, it hinders the group's publication output. We interpret these results as an indication that gains, in the amount of research output, can be achieved if PhD students specialize according to their comparative advantages. However, beyond a certain level, knowledge specialization has a detrimental impact on research output, due to increasing communication costs and an increased likelihood of conflict insurgence.

Managerial relevance statement: Our study makes an important contribution to understanding the implications of within-group knowledge specialization on the productivity of knowledge-intensive groups. We argue that knowledge specialization generates benefits to a research group in terms of increased output because it allows the group members to specialize in those research areas for which they have a comparative advantage. However, knowledge specialization also involves communication costs and favors the insurgence of conflicts among group members. Hence, when the head of a research group sets the optimal level of knowledge specialization for her group, she has to solve a

*We are indebted to Paula Stephan, Peter Thompson, Marie Thursby, Jerry Thursby, and two anonymous reviewers for insightful comments and suggestions. We also thank Alberto Galasso, Ramesh Nallapati, Henry Sauermann, and seminar participants at the REER conference, Georgia Institute of Technology, Atlanta, November 2010, for their valuable comments. Mark Steyvers and Tom Griffiths kindly shared with us the Matlab Topic Modeling Toolbox for applying Latent Dirichlet Allocation. Tatiana Benavides and Eloisa De Santiago provided outstanding research assistance. Conti gratefully acknowledges support from the Hal and John Smith Chair in Entrepreneurship for support via a TI:GER Postdoctoral Fellowship and the Swiss National Science Foundation (Fellowship for Prospective Researchers n. 118734). Visentin gratefully acknowledges support from the University of Lugano and the Swiss National Science (Foundation Fellowship for Prospective Researchers n. 125573).

†Scheller College of Business, Georgia Institute of Technology, GA 30308, Atlanta, USA, annamaria.conti@scheller.gatech.edu

‡Dept. of Math & CS., Emory University, GA 30322, USA, odenas@emory.edu

§Chair in Economics and Management of Innovation, EPFL, CH-1015, Lausanne, Switzerland, fabiana.visentin@epfl.ch

trade-off between the gains and costs generated by knowledge specialization. Our empirical findings suggest that there is an inverted U-shaped relationship between the research output of a PhD student group and the level of within-group knowledge specialization. From a managerial perspective, this result has implications for the optimal design of firms' research-intensive groups. Indeed, the mechanisms that govern the functioning of these groups have similarities with those of PhD student groups, while professors increasingly resemble firms' division heads as they devote a large share of their time to organizational tasks.

Keywords: Knowledge, Specialization, PhD Students, Productivity, Research Output, Group Organization

1 Introduction

As knowledge accumulates, successive generations of innovators face a rising educational burden [28]. If standing alone "on the shoulders of a giant"[1] becomes increasingly cumbersome, innovators have the alternative of seeking narrower expertise and working in groups. This alternative has gained momentum in recent years, and the empirical evidence demonstrates that innovators increasingly work in groups [28], [52] and that collaborative work produces more breakthrough inventions than individual work [44].

The traditional argument is that the optimal size of research groups reflects a tradeoff between the gains from labor specialization and the costs of coordination [5]. In their seminal work, Becker and Murphy show that an increase in the size of a group positively affects the group's output by favoring task specialization and division of labor: in large groups, individuals can specialize in a narrow set of tasks and become experts at what they do. Despite these advantages, Becker and Murphy have pointed out that the costs of coordinating group members increase with the size of the group, and that group managers should account for these costs when determining the optimal size of their group. As shown by Holmstrom and Nalebuff [26], when groups become large, their members have an incentive to shirk because they receive a smaller fraction of the output they produce. Moreover, borrowing from the vast literature on the conflicts that arise in heterogeneous groups (see, for example, [24], [29], [51]), it is likely that as groups become large, they also become more heterogeneous along a number of characteristics,

[1] In 1676, Isaac Newton wrote to Robert Hooke: "If I have seen further it is by standing on ye sholders of Giants."

which gives rise to conflicts, thereby diminishing cohesion and negatively affecting the group's performance.

Recent empirical studies have shown that, over time, the size of research groups in research-intensive industries and academia has increased, and that these groups have become more geographically dispersed [1], [28]. Moreover, Adams *et al.* [1] have found that geographical dispersion has accelerated since the beginning of the 1990s. They argue that this trend might be due to the decreasing costs of collaboration. Building on Adam *et al.*'s study, Agrawal and Goldfard [2] and Forman and van Zeebroeck [17] have shown that the adoption of BITNET [2] and the diffusion of the Internet [17] have lowered communication costs, thereby fostering collaboration among universities and among geographically dispersed firm employees.

An important aspect that these empirical studies have neglected is the possibility that, *for a given research group size*, gains can be achieved by optimally allocating knowledge among its members [30], [31]. In light of Ricardo's theory of comparative advantage, The overall output of a research group can be enhanced, when members of a group are allowed to specialize in the research areas and tasks for which they have or could develop a comparative advantage. However, it is also true that knowledge specialization triggers communication costs and the insurgence of conflicts, which might be detrimental to research output. We extend the existing literature by empirically addressing the optimal degree of knowledge specialization within research-intensive groups in the context of PhD student groups.

Focusing on PhD student groups is particularly relevant, given the contributions that academic research makes to industrial innovation [27], [34], [37], [38] and the fact that PhD recipients are one of the most important sources of new talent in many professions [21]. As of today, few studies have analyzed the determinants of PhD student productivity. Stuen *et al.* [49] analyzed the productivity of foreign and domestic PhD students in the US, while Groen *et al.* [22] studied the determinants of receiving a PhD among US college graduates and Waldinger [50] assessed the role of faculty quality in PhD students' outcomes.

We use a dataset of 1,938 PhD student groups who studied at the Swiss Federal Institute of Technology (EPFL) in Lausanne, Switzerland, during 1993-2008. We derive detailed biograph-

ical information from each student's dissertation and then match this information with the publication record of the group and collect fine-grained data on the group's supervisor. Next, we extract the research areas in which the students had specialized during their PhD studies from the dissertation abstracts, applying an established technique in computer science known as Latent Dirichlet Allocation. We then measure the degree of knowledge specialization in the groups, based on the distribution of these topics among student abstracts.

We relate our measure of knowledge specialization to the research productivity of each group, which is defined by the number of scientific articles published by the group. We estimate a Poisson model for the number of publications and find that, holding constant the group members' ability, the group's research breadth and size, the supervisor's knowledge capital stock and other controls, an increase in the level of within-group knowledge specialization is associated with a larger number of publications. However, specialization becomes detrimental to the group's publication output at a certain level. We find similar results when we adopt a more restrictive definition of student publications and consider in the publication count only those articles that had received a certain number of citations. We interpret these results as an indication that gains in research output can be achieved if PhD students specialize according to their comparative advantages. However, beyond a certain level, knowledge specialization has a detrimental impact on research output. We argue that this is due to communication costs and to the likelihood of conflict insurgence, both of which increase with the level of within-group knowledge specialization.

2 State-of-the-art and the research hypothesis

Several studies have acknowledged the contribution of PhD students to a country's innovation capacity [27], [34], [37], [38]. Their *leitmotiv* is that the movement of PhD students from academia to industry or from one academic institution to another is one of the primary means of knowledge and technology transfer [14]. Recently, a few studies have attempted to uncover the black box of PhD student contributions to the production of scientific output. For instance,

Groen *et al.* [22] have documented the increased propensity of students to participate in PhD programs. Stuen *et al.* [49] have shown that both foreign and domestic PhD students are central inputs in knowledge creation in the US; however, at the margin, foreign PhD students contribute more than domestic students to the creation of knowledge output. This result is in line with the finding by Black and Stephan [7] and Libaers [33] that foreign researchers (including, but not limited to, PhD students) are consistently more productive than their domestic counterparts. Moreover, Waldinger [50] has analyzed the impact of faculty quality on PhD student outcomes. He finds a strong effect of faculty quality on the probability that a PhD student publishes her dissertation in a top journal, the probability that the PhD student becomes a full professor, and the number of citations.

None of the empirical studies above have explicitly considered that PhD students are members of a research group or that the organization of this group by the supervisor has important consequences on its productivity. Increasingly, research groups resemble "quasi-firms" [16], in which the supervisor acts as a team leader by setting the research direction of its members and providing mentorship to the group's trainees. Within a supervisor's research group, PhD students represent an essential component providing complementary work to that of their senior colleagues [36]. Mangematin and Robin suggest that PhD students constitute a sub-group within a supervisor's research group and that this sub-group is characterized by specific traits. These traits include temporary participation in the supervisor's research group and the objective of training through research [43].

The lifecycle of PhD students is typically characterized by an initial phase in which they acquire information on a research topic that they have agreed upon with their supervisor and a subsequent phase in which they contribute to expanding the state-of-the-art. In both phases, the supervisor plays a crucial mentoring role. This role often takes the form of i) directing the research efforts of her PhD students toward solving research puzzles that are considered to be relevant by the scientific community, ii) providing information on the state-of-the-art, or iii) assisting her students to ensure that they acquire the necessary skills to solve the research puzzle [35].

Having acknowledged that mentoring is essential for the intellectual development of PhD students, we should also mention that mentoring requires a considerable amount of time and effort on the part of the supervisor. Given the numerous tasks a supervisor must accomplish, the opportunity cost of mentoring a student is likely to be high [34]. Thus, it is natural that the supervisor implicitly expects that the mentored PhD student will exert effort to produce scientific papers, with the supervisor's name as a coauthor [47]. The supervisor expects returns not only because of the mentoring costs incurred, but also in light of the implicit contracts that she stipulates with her PhD students, which offer funding and interesting research careers in exchange for research output [46].

Given the returns a supervisor expects from the scientific work of her PhD students, a central question for the supervisor becomes how to optimally organize knowledge in a PhD student group to maximize the production of research output. Our focus is on the relationship between the degree of knowledge specialization in a PhD student group and the group's research output. Thus, we ask *what the optimal degree of knowledge specialization is within a PhD student group.*

The supervisor can achieve gains in terms of research output by having each PhD student specialize in the research topic for which she has a comparative advantage or for which she could develop a comparative advantage, if she does not initially have one. The importance of the gains from specialization has been stressed in recent studies [28] and [52], which have documented the increased propensity of scientists to work in teams and specialize in narrow research fields, due to the higher education burden they face relative to previous generations of scientists. In these works, team members are defined as coauthors on a paper, and the underlying assumption is that members' specialization in given research tasks increases overall research output. In our study, PhD students may or may not work on the same research project, and thus not all the PhD students will necessarily appear as coauthors on a paper. In this last case, the supervisor's gain from specialization derives from the diversification of her research portfolio, which reduces the risk that her research will not be published [39].

Knowledge specialization within a PhD group is not exempt from drawbacks. Indeed, it is likely to be positively correlated with communication costs, as well as with the insurgence

of conflicts, both of which are detrimental to a PhD group's research output. These costs are relevant not only when students work on a same project but also when they are assigned to different projects [24], [29],[51]. It is likely that when within-group specialization increases beyond a certain level, communication costs and the insurgence of conflicts prevail over the gains of specialization, as PhD students are no longer able to exchange ideas and provide advice to one another [19]. Based on the argumentation we have discussed above, we expect the following:

Hypothesis 1: There is an inverted U-shaped relationship between the research output of a PhD student group and the level of within-group knowledge specialization.

3 PhD student groups at EPFL

The empirical context for testing our hypothesis is groups of PhD students from EPFL, Lausanne, Switzerland. This university has a reputation for producing cutting-edge research. Publication data retrieved from Scopus show that during the 1993-2008 period, EPFL produced approximately 13,000 articles in science and engineering. Moreover, it holds a high ranking in the fields of engineering, technology, and computer science, according to a ranking of world universities published by Shanghai Jiao Tong University.

Our dataset consists of groups of PhD students who studied at EPFL during the 1993-2008 period. We define a PhD student group as a group of PhD students who are supervised by the same professor in year t and are driven by the ultimate goal of producing research output. Extensive interviews with supervisors at EPFL revealed that PhD students tend to meet once a week to discuss their research progress with one another and their supervisor. These meetings allow the supervisor to assess the progress of the PhD group, set research goals, and assign research tasks to the group members.

PhD students at EPFL are selected by the faculty through a formal interview process. Once they are hired, they are paid a salary from the Swiss Confederation, which in some cases is integrated from a supervisor's grant. The salary is guaranteed for the duration of their PhD.

Typically, students are required to complete their PhDs within four years. Extensions are possible, but they tend to be no longer than six months. Students who start their PhD work have already obtained a master's degree, and need to take only a few additional courses. Hence, their major task consists of doing research. Another important feature of the PhD program is that, once applicants are admitted to the program by a professor, they will work with that professor for the duration of the PhD program. Discussions with administrative personnel at EPFL confirmed that PhD students only rarely switch to another supervisor. Moreover, official statistics provided by EPFL reveal that, over the 1993-2008 period, the average drop out rate among PhD students was only 11%, and dropouts occurred mainly within the first year. This might be explained, at least partially, by the relatively high salary offered by the Swiss Confederation and by the fact that the salary is guaranteed for the entire duration of the PhD.

We construct these PhD student groups based on each student's graduation year. Although we only observe the students who eventually graduated, we feel confident that our group definition is sufficiently accurate. The low drop out rate and the rarity of PhD students switching supervisors ensure that, as a general rule, a student who completed her PhD in year t with professor x started her PhD in year $t-3$ with the same professor. Thus, we construct the groups as follows. Suppose that three students, A, B, and C, all completed their PhDs in year t with supervisor x and that another PhD student, D, completed her PhD in $t-1$ with the same supervisor. Then, in year $t-3$, the group of PhD students supervised by professor x consisted of students A, B, C, and D. The group's composition remained unchanged in years $t-2$ and $t-1$ and, in t, after D graduated, the group consisted of only students A, B, and C.

Our main sources of information are 2,302 PhD dissertations, which also include biographical information for the corresponding 2,302 PhD students. These students form 1,938 research groups supervised by 249 professors. For the purposes of our study, we exclude groups that only include a single PhD student. A professor, on average, supervised 7.80 PhD student groups during the 1993-2008 period, with a minimum of 4 and a maximum of 16 groups. Of these groups, 84% had unique PhD membership compositions. When classified by discipline,

15% of PhD student groups were in the computer science field, 46% in engineering, 5% in life science, and the remaining groups were in basic science. Within engineering, we distinguish between civil engineering (26% of all engineering groups), electrical engineering (24%), micro-engineering (19%), mechanical engineering (17%), and material science (14%). Within basic science, we distinguish between physics and mathematics (66% of all basic science groups) and chemistry (34%). The average size of a PhD group is 4.61 PhD students, and no group had more than 30 members. PhD students at EPFL come from a variety of countries; 93% of all PhD student groups have at least one member from a country other than Switzerland. This implies that there is a large variety in terms of the universities from which the students received their master's degrees. The average age of the group members is 29, with a minimum of 23 and a maximum of 47. On average, each group has 1.13 first-year students, 1.17 second-year students, 1.18 third-year students and 1.13 fourth-year students.

The information from the dissertations is complemented with publication data available from Scopus. The average number of publications is 3.12; 33% of the groups did not have any publications. Fifty percent of the groups with one or more publications have at least one paper that is coauthored by at least two group members. Moreover, 79% of the group's publications listed their supervisor as a coauthor. This is an indication that academic professors derive a return from the PhD student groups they supervise. The average number of citations is 65.25, and the median is 10. The difference between the mean and the median suggests that the distribution of citations is highly skewed.

Finally, our data on PhD student groups are matched with data on their supervisors. We collect information on the professors' ages, nationalities, pre-sample publication stock, and grants received during the 1993-2008 period. We define a professor's pre-sample stock of publications as the number of publications she authored in the five years preceding the first group of PhD students she supervised. The major source of grants in Switzerland is the Swiss National Science Foundation (SNSF), an institution whose primary goal is to promote scientific research. Of the supervisors, 49% are Swiss and the rest are primarily German, French, or Italian. The professors' average age was 49.7. On average, professors had 40.90 publications in

the pre-sample period, none having more than 370.

4 Knowledge specialization in PhD student groups

To assess the specialization of knowledge within a PhD group, we have to identify the research topics that the members' abstracts describe. These topics define the students' research areas of expertise. We then build a measure of group knowledge specialization that captures the degree of disparity between the group members' dissertation topics. This measure compares the group members' dissertation abstracts and assesses the diversity of their research topics.

We begin by eliminating words that are not instrumental in characterizing the content of a PhD dissertation. These words are known as "stop words" in the data mining community and include prepositions, conjunctions, and general adjectives. Examples are "and," "if," "or," and "agree." By inspecting a random sample of 500 abstracts, we manually identify 1,550 instances of such words, which we eliminate from every abstract.

Successively, we apply a stemming procedure that strips each word to its bare root. In this procedure, the words "agree," "agrees," and "agreement," for example, are stripped to the root "agree." For this purpose, we use an algorithm that was initially implemented in computer science by Porter [41]; it is one of the most widely used algorithms for stemming in computer science.

We infer the topics of each PhD dissertation by using a method called Latent Dirichlet Allocation (LDA), which is a technique that computer scientists commonly use to infer the topics in texts [20], [42], [45], [48][2]. The idea of LDA is that documents are mixtures of topics, and topics are probability distributions over words [48]. For example, if a document contains frequent occurrences of the words "water," "flood," and "hydrolog," we might reasonably infer that one of the topics of the document is hydraulics. In using LDA, we have to pre-define the number of topics for which we want to search in the entire corpus of dissertation abstracts, with the idea that each abstract is composed of a fraction of these topics. In order to choose

[2]For details of the algorithm, refer to Blei et al. [8].

the number of topics, we must consider the interpretability of the results. There is a trade-off between choosing a small number of topics, which produces topics that are too broad, and choosing a large number of topics, which leads to "uninterpretable topics that pick out idiosyncratic word combinations."[48] To select the number of topics, we rely on advice from several computer scientists and select 120 topics. In Table 1, we present 6 of the 120 topics obtained. The table shows the ten words that are most likely to occur under each topic, as derived from the entire sample of abstracts. As a robustness check, we also present the analysis with 100 topics, closely following the work of Griffiths and Steyvers [20].

Once we have inferred the research topics from an abstract, we apply the second step of LDA and redefine the abstract as a probability distribution over topics. The resulting output from this step is a matrix of counts with dimensions $D \times T$, where D is the number of abstracts, and T is the number of topics. Each entry contains the number of times a topic t is assigned to some word token in abstract d. We then assess the organization of knowledge in a group as follows. First, within each abstract, d, we consider only those topics with relative frequencies that are greater than 5%. By applying this cutoff, we avoid including topics in our comparisons with low levels of relative importance within an abstract.[3] The average number of topics studied by a group of PhD students is 8.19, with a minimum of 1 and a maximum of 25.

Second, we build a measure of group knowledge specialization that captures the degree of disparity between the group members' dissertation topics. We use a dispersion index defined on the set of abstracts from students of the same group, $\{D_1, ..., D_k\}$. To this scope, we define the count of any topic t as $c_t = |\{D_i | t \in D_i, 1 \leq i \leq k\}|$, i.e., the number of group abstracts with topic t, and its relative frequency as $f_t = c_t/F$, where F is the sum of c_t over all t.

The measure is then:

$$Within\text{-}Group\ Specialization = 1 - \sum_t \left(\frac{c_t}{F}\right)^2$$

The term $\sum_t \left(\frac{c_t}{F}\right)^2$ is equivalent to a Herfindahl Index (HI) applied to topic counts. HI varies between 0 and $1/F$, with lower values indicating greater dispersion in the group's research.

[3]The results are robust to applying different cutoffs to the topics' relative frequencies.

Hence, we normalize $\sum_t \left(\frac{c_t}{F}\right)^2$ to ensure that it varies between 0 and 1.

Hence, we compute the normalized HI index as follows:

$$HIN = \frac{\sum_t \left(\frac{c_t}{F}\right)^2 - \frac{1}{F}}{1 - \frac{1}{F}}$$

The new dispersion measure is:

$$Within\text{-}Group\ Specialization\ (N) = 100 \times [1 - HIN]$$

This measure varies between 0 and 100, with larger values indicating a higher degree of knowledge specialization within the PhD student group. As an example, a group with a value of *Within-Group Specialization* (N) equal to 0 has all of its members conducting research on Parkinson's disease. Conversely, a group with a value of *Within-Group Specialization* (N) equal to 100 has a member working on dynamic optimization to reduce production costs and the other investigating problems of disturbance rejection applied to the fedbatch fermentation of *Saccharomyces cerevisiae.* For the groups in our sample, the mean of *Within-Group Specialization* (N) is 89, with a minimum of 0 and a maximum of 100. The high value of the mean suggests that the PhD students in our groups, on average, tend to be highly specialized and conduct research in diverse areas within their larger fields.

⟨ Insert Table 1 about here ⟩

5 Econometric Methodology

To test our hypothesis, we estimate the impact of our measure of within-group knowledge specialization, *Within-Group Specialization (N)*, on the publications produced by the PhD student groups. Consistent with the literature on the research productivity of academic researchers [12], [4], [32], which allows for lags between the time, *t,* of the beginning of a research project and the resulting output, we measure a group's research output using the group's scientific articles published in *t+1.* As a robustness check, we also use scientific articles published in *t+2.*

We also adopt a more restrictive definition of a PhD group's publication count in order to

pin down the impact of within-group knowledge specialization on quality research output[4]. For each student in a PhD group, we count only those publications having received a number of citations that is higher than a certain cutoff. The cutoff we use is the median number of citations received by student articles published in the same year and in the same field as the observed student's articles. We then aggregate these publications at the level of the PhD group and denote them as *highly cited publications*. In Figure 1, we show that by applying this criterion, the percentage of PhD groups who had not published any article increases from 33.23 to 47.57. As robustness checks, we use alternative cutoffs for the number of citations. We prefer these specifications over a simple count of a PhD student group's number of citations, because we want to mitigate the standard problems inherent in measuring research quality using citation counts. For instance, previous studies have presented evidence that citation counts significantly vary across subject fields, and they are correlated with factors such as article length or article styles (whether they are reviews or not), which are not necessarily informative of the quality of research output [3]. Other studies have shown that self-citations are a non-trivial share of the total citations received by a scientific article [18].

⟨ Insert Figure 1 about here ⟩

We estimate count regression models that account for the fact that our dependent variables can only take positive integer values. We adopt a Poisson specification and use robust standard errors, clustered by the supervisors of the PhD student groups. This specification ensures consistent estimates of the parameters in our model. We prefer this specification to a negative binomial maximum likelihood estimation because the latter is not consistent if the variance specification is incorrect [13].[5]

The conditional expectation of group i's number of publications can then be expressed as follows:

[4]Other studies (see for example [11]) have used journal impact factors to measure the quality of researchers' articles. We refrain from choosing this option because a large proportion of our PhD groups are in the computer science field, and, in this field, conference proceedings (which in many cases have at least the same importance as journal articles) rarely have an associated impact factor.

[5]In robustness analyses not reported here (but available upon request), we estimate negative binomial models and obtain very similar results to those using a Poisson specification.

$$\mu_i = \exp(\beta_1 Within\text{-}Group\ Specialization\ (N)_i + \beta_2 Within\text{-}Group\ Specialization\ (N)_i^2 + X_i'\gamma + Z_i'\rho + W_i'\nu + V_i'\vartheta)$$

To test our hypothesis, we include the measure of the within-group knowledge specialization we have built and its squared term. As knowledge specialization increases, communication costs and the likelihood of conflicts also increase; thus, we should expect an inverted-U shaped relationship between knowledge specialization and the group's output. Finally, we also include four types of controls. The variables in X_i capture the attributes of the PhD student group, those in Z_i capture the characteristics of the supervisor, those in W_i capture the characteristics of the department, and the variable in V_i captures the characteristics of EPFL.

Characteristics of the PhD student group

We measure the size of a group by the number of group members. We denote this variable as *Group size*. Including this control is important because our measure of within-group knowledge specialization is likely to be correlated with group size. Hence, we cannot assess the effect of knowledge specialization on group productivity unless we net out the impact of group size. We expect the relationship between the size of the group and its research output to have an inverted U-shape. In fact, after a certain level, the positive impact of having a large number of contributors to the group's research output is likely to be offset by an increase in coordination costs. Additionally, given the scarcity of positions available once the students have completed their PhDs [47], large PhD groups might foster situations of competition and conflict among students, which discourage collaboration and have a detrimental impact on the group's productivity. For these reasons, we include a squared term of *Group size*. Moreover, we control for the breadth of a group's research with a count of the number of topics that the group investigated, which we label as *Group research breadth*. The topics are obtained using the LDA technique described above. As with the size of the group, we include a squared term of knowledge breadth. Additionally, we control for the research tenure of the group (*Mean group tenure*) by computing the mean number of years the PhD students have spent in the

PhD program. We include a squared term to account for the possibility that the relationship between research tenure and output is concave. We also include a measure of the quality of the group members (*N students with research awards*). This is a count of group members who were awarded a prize for the best dissertation at the end of their PhD studies. These awards are the result of student quality as well as other factors, such as the supervisor's knowledge stock, the funds available for research, and the quality of other group members. However, to the extent that we control for these factors, any residual impact of our variable on the group's research productivity should be attributable to student quality. We also include the variable *Mean group age* to measure the average age of the group members. We follow a large body of literature that examined the impact of demographic heterogeneity on group productivity (see [25], [30], [31], [40], [53]) and control for the dispersion of members' ages and for the number of different universities at which the group members obtained their master's degrees. This last variable is a proxy for the diversity of the students' university backgrounds. We denote these variables as *StdDev group age* and *Master background diversity*, respectively. We also include a squared term of *Master background diversity*. Furthermore, we include a dummy to control for group members who had two supervisors. These students may have been assigned an additional supervisor because one of the supervisors was not a full professor. However, the assignment of an additional supervisor may also occur because a student conducts interdisciplinary research, which requires more than one supervisor.

Characteristics of the group's supervisor

We include a variable that measures the supervisor's age, *Professor age*, and five dummies that control for the professor's nationality. These dummies control whether a professor is Swiss, German, Italian, French, or some other nationality. Additionally, we use the stock of pre-sample publications to measure a professor's knowledge capital stock. This variable is defined as the number of articles that a professor published in the five years prior to supervising her first PhD student group in the sample. We include a squared term, given that highly productive professors might have little time to supervise their PhD students. We also have a measure of the research budget available to the professor. This measure, which we denote as *SNSF grants*,

is defined as the amount of grant funds (in thousands of real Swiss Francs) that are awarded by the SNSF in the five years prior to the creation of the group. This variable controls for the size of a supervisor's laboratory, including postdoc students, as SNSF grants are mainly used to pay the salaries of laboratory personnel.

Characteristics of the department

To account for the characteristics of the department with which a group is affiliated, we include department fixed effects. Specifically, we have dummy variables for the following departments: Chemistry, Physics and Mathematics, Life Science, Civil Engineering, Electrical Engineering, Mechanical Engineering, Micro Engineering, Material Science, and Computer Science. Moreover, we control for the budget of each department with a variable, called *Department funds*, which measures the amount of SNSF grants (in millions of real Swiss Francs) that were assigned to a department in year t. This variable controls for the impact of the time-varying size of the department on a group's research output.

Characteristics of EPFL

Finally, we include a linear time trend to control for time effects. In the case of EPFL, it is very important to control for time effects because the university has undergone an important evolution over the last two decades. Initially a technical university with no aspirations to academic excellence, EPFL has progressively become a reputable university in the fields of engineering and computer science. Summary statistics are reported in Table 2.

⟨ Insert Table 2 about here ⟩

An important issue when analyzing the impact of group characteristics on productivity is selection into groups [25]. A professor may select students into a group based on unobserved characteristics that are correlated with our measure of within-group knowledge specialization and the group's research output. Among these characteristics, those of the group's supervisor are the most relevant. We refrain from estimating a fixed-effects model, given that the fixed-effects Poisson estimator leads to a substantial loss of data. This is because observations for PhD student groups with an overtime sum of publications equal to zero do not contribute to

the estimation [23]. However, to account for the potential confounding role of professor-level unobserved factors, we adopt an approach similar to that developed by Blundell *et al.* [9] and Blundell *et al.* [10] and applied by Belenzon and Schankerman [6] to the context of university licensing. Blundell *et al.* [9] and Blundell *et al.* [10] show that, under certain assumptions, the pre-sample mean of the dependent variable is a consistent estimator of unobserved, fixed, heterogeneity, and therefore can be used as a control for such heterogeneity. Similar to Belenzon and Schankerman [6], we do not have pre-sample information on the dependent variable, so we follow their approach and use pre-sample information on a variable that is correlated with our dependent variables. In our case, we use a professor's pre-sample stock of publications, noting that 79% of the PhD students' publications are coauthored by their supervisors. To account for the possibility that some of the professors' characteristics might be time variant, we include the one-year lagged value of a supervisor's publication count as a robustness check. Another important issue in our analysis is reverse causality: professors choose a level of knowledge specialization for their PhD group to increase expected productivity and, at the same time, expected productivity induces supervisors to select a given level of within-group knowledge specialization. While we do not have suitable instruments to estimate an instrumental variable regression model, we follow Blundell *et al.* [9] and estimate a dynamic feedback model where we introduce the one-year lagged value of a professor's PhD group productivity.[6] The underlying assumption of such a model is that professors form expectations about the productivity of their future PhD student group, based on the productivity of their current PhD group. This last variable is measured either by the count of publications or by the count of highly cited publications, depending on the definition we use for the regressand. In constructing our lagged variables, we note that the composition of the PhD group at time t might be different from that of the PhD group at time *t-1*. This occurs because, at the end of time *t-1*, some students might have left the group, while other students might have joined. Finally, we note that the number of observations decreases to 1,689, in estimating such a model, given that we lose the

[6]Without loss of insight, we measure the research output of the group in *t-1* using a one-year lag between the group organization and its research output. Results remain the same if we use a two-year lag.

first period observation for each of the 249 professors in our sample[7].

6 Results

6.1 Baseline Results

Table 3 presents the regression results for the number of PhD student groups' publications. In column I, within-group specialization is related to the number of publications in *t+1*; in column II, we introduce the one-year lagged value of a supervisor's publication count; in column III, we add the one-year lagged value of a supervisor's PhD group's publication count. In columns IV to VI, we estimate the same models as in columns I to III, respectively, but we use the number of publications in *t+2* as the dependent variable.

The results have several notable features. First, all else being equal, greater within-group knowledge specialization is associated with a larger number of group publications. The coefficient of *Within-Group Specialization* (N) is positive and statistically significant, regardless of whether we include the lagged count of a professor's publications. As expected, the magnitude for the coefficient of the linear term is slightly lower when we include the lagged term of a supervisor's publications. The results hold when we include the publication count of a supervisor's PhD group at time *t-1*, despite the reduced sample size. Introducing this variable induces a decrease in the magnitude of the coefficient of *Within-Group Specialization* (N) by about 6%, relative to the coefficient we would obtain without the lagged publication count of a supervisor's PhD group and using the reduced sample size of 1,689 PhD groups. In all cases, the relationship between knowledge specialization and research output has an inverted U-shaped form, as indicated by the negative and statistically significant coefficient of the squared term of *Within-Group Specialization* (N). These results confirm our hypothesis that allowing group members to specialize in the research areas at which they are relatively good has a positive impact on a group's productivity. However, beyond a certain level, specialization becomes

[7]We could not retrieve earlier information of a supervisor's PhD group because it is not available from EPFL records.

detrimental to the group's output due to increasing communication costs and the likelihood of conflicts.

In what follows, we describe the results for the other controls of interest. The relationship between group size and group research output, and that between group research tenure and research output, have inverted-U shapes. This last result occurs because last-year PhD students need to spend a significant portion of their time writing their dissertations; thus, they may not be as productive as they were during their second or third years. The coefficients of *Group research breadth* and its squared term are positive and negative, respectively, in all regressions. While the coefficients are not statistically significant, we reject the null hypothesis that they are jointly equal to zero at the 10% confidence level, regardless of the regression specification. All else being equal, groups with more members who were awarded prizes for their PhD dissertations are also more productive. Having controlled for a supervisor's pre-sample stock of publications and the funds she was awarded, we are confident that the residual impact of *N students with research awards* on group research output captures the intrinsic quality of the group members. The mean age of a PhD student group is negatively related to the number of publications produced by the group. This may be because older PhD students need to allocate their time among a larger number of tasks than younger students, including, for instance, taking care of their children. The coefficient for the number of different universities from which the students received master's degrees is positive, and its squared term is negative. In all regression specifications, these coefficients are jointly and significantly different from zero. If the number of different universities from which the students received master's degrees is positively correlated with the diversity of the group members' backgrounds, our results suggest an inverted U-shaped relationship between background diversity and the group's number of publications. Surprisingly, having two supervisors is negatively related with the number of a group's publications. Discussions with PhD candidates at EPFL revealed that it is often the case that PhD students with two supervisors risk being neglected by both supervisors. It could also be that having two supervisors is an indicator that these students are problematic. For instance, those students who are not able to integrate and work with a given research group,

might be assigned to an additional group. Their problematic behavior ultimately could result in poor scientific productivity.

Regarding the supervisor's controls, a larger stock of capital knowledge is associated with greater productivity for her PhD student group, although the relationship is non-linear. When we include the one-year lagged value of a supervisor's publications, the coefficient is positive and highly significant. The amount of SNSF grant money received by the supervisor in the five years prior to the creation of the group has a positive and statistically significant impact on the number of publications. To the extent that these grants control for the size of a supervisor's laboratory, this is an indication that laboratory size is positively associated with PhD student group productivity. Finally, supervisor age is negatively correlated with PhD group productivity. This result is consistent with the finding of Levin and Stephan [32] that a scientist's research productivity declines over her lifecycle.

The coefficient of *Department funds* is negative and statistically significant in the regression for the number of publications in *t+1*. To the extent that this is a proxy for the department's size, the result suggests that larger departments are associated with fewer publications. Finally, department dummies tend to be highly significant, suggesting that there are important differences across departments with respect to the PhD student groups' research productivity.

Table 4 presents the regression results for the number of highly cited publications. In column I, the organization of the group is related to the number of highly cited publications, published in *t+1*; in column II, we introduce the one-year lagged value of a supervisor's publication count; in column III, we add the one-year lagged value of a supervisor's PhD group's (highly cited) publication count. In columns IV and VI, we estimate the same models as in columns I to III, respectively, but we use as a dependent variable the number of highly cited publications, published in *t+2*. The results are similar to those presented in Table 3. The relationship between within-group knowledge specialization and the number of highly cited publications has an inverted U-shaped form. The coefficients of *Within-Group Specialization* (N) and its squared term are positive and negative, respectively, and are highly statistically significant. In regression results not presented here (but available upon request), we conduct robustness

checks and use the 75th and 90th percentiles as cutoffs for the number of citations. In the regressions for the number of highly cited publications published in *t+1*, we continue to find that the coefficients of *Within-Group Specialization* (N) and its squared term are, respectively, positive and negative, and are both statistically significant. In the regressions for the number of highly cited publications published in *t+2,* the coefficients of *Within-Group Specialization* (N) and its squared term maintain the expected signs, but are no longer significant.

⟨ Insert Tables 3 and 4 about here ⟩

6.2 Robustness Analysis

To address the concern that our results may be sensitive to the number of topics we have pre-selected in constructing our measure of within-group knowledge specialization, we estimate the same equations as those in Table 3, this time following Griffiths and Steyvers [20] and pre-selecting 100 topics. The results are reported in Table 5. The new measure of specialization with 100 topics is positively correlated with the previous one, and the correlation coefficient is 0.55. The mean of this measure is 92.9 and ranges between 0 and 100. As we have pre-selected a smaller number of topics, the standard deviation of the measure drops by almost half, as shown in Table 2. The results in Table 5 tend to confirm the results presented in Table 3 for the simple count of publications. Regarding the number of highly cited publications (Table 6), the coefficients of the within-group specialization measure and its squared term have the expected signs, but are no longer significant. This is not unexpected. Given the smaller number of pre-selected topics, the new measure captures within-group knowledge specialization less precisely, which translates to higher standard errors for the measure's coefficients.

⟨ Insert Tables 5 and 6 about here ⟩

7 Discussion and Conclusions

Our study makes an important contribution to understanding the implications of within-group knowledge specialization on the productivity of knowledge-intensive groups, such as those composed of PhD students. Knowledge specialization can generate benefits to a research group in terms of larger output, by allowing the group members to specialize in those research areas for which they have a comparative advantage. At the same time, it involves communication costs and might favor the insurgence of conflicts among group members, both of which increase with knowledge specialization. Thus, when the head of a research group sets the optimal level of knowledge specialization for her group, she has to solve a trade-off between the gains and the costs that knowledge specialization can generate.

We empirically test the impact of knowledge specialization on the research productivity of PhD groups, using a novel dataset on PhD student groups at EPFL, Lausanne, Switzerland. This dataset contains a large amount of fine-grained information on students' and supervisors' demographics, as well as student research quality. It also includes the research areas students focused on in their PhD research, as described in their dissertation abstracts. Using this last piece of information, we build a novel measure of within-PhD student group knowledge specialization, which applies an established technique in computer science known as Latent Dirichlet Allocation. We then relate this measure to the research output of PhD groups. We estimate panel data models and find that, holding constant the size of a PhD student group and the breadth of its knowledge, the relationship between within-group knowledge specialization and group publication productivity has an inverted U-shape. This relationship holds, under certain regression specifications, when we adopt a more restrictive definition of student publications and consider in the publication count only those articles that had received a certain number of citations.We interpret this result as follows. On the one hand, within-group knowledge specialization has a positive impact on research output because it takes advantage of the group members' comparative advantages. On the other hand, knowledge specialization induces communication costs and the insurgence of conflicts, which, after a certain point, prevail

over the gains from specialization.

Our results have implications for the literature on the organization of research-intensive groups. Adams *et al.* [1] and Jones [28] have shown that the size of research groups has increased over time. Agrawal and Goldfard [2], Ding *et al.* [15], and Forman and Zeebroeck [17] have shown how the advent of information technology has affected the organization of R&D collaborations. The underlying assumption in these studies is that larger research groups can benefit from the division of labor and task specialization, especially when communication costs are low. Our work shows that, *holding constant the size of these groups and the breadth of their knowledge*, benefits can be achieved by ensuring that the group members specialize in the research areas for which they have a comparative advantage. However, to the extent that communication costs and the likelihood of conflicts among group members increase with knowledge specialization, complete specialization may not be achieved. By investigating the dispersion of research topics within the dissertation abstracts of PhD students, our study allows us to examine the distribution of knowledge within groups, *independent* of whether these groups have published their research. We believe that this is an important contribution, as it allows us to assess the impact of within-group knowledge specialization on the research output of the group, including the realistic possibility that there is no output. Finally, we have applied Latent Dirichlet Allocation to the analysis of PhD students' abstracts. We believe this technique can be easily used for the analysis of other research documents, such as publication abstracts, to derive areas of interest in which scientists specialize.

From a managerial perspective, our results have implications for the optimal design of firms' research-intensive groups. The mechanisms that govern the functioning of these groups and their goals have certain similarities with PhD student groups. As an example, professors increasingly resemble firms' division heads in that they devote a large share of their time to organizational tasks [16]. Our results suggest that firms seeking to maximize the output of their research groups can do so by allowing the group members to specialize in those research areas for which they have a comparative advantage or by hiring members who are already specialized. However, the optimal level of knowledge specialization should take communication costs into

account.

From a policy perspective, our results have important implications because they shed light on the functioning of PhD student groups, whose contributions to a country's innovation capacity have been widely recognized. Although these results are based on a single academic institution with a PhD program that has particular features, they can be extended to other programs. For instance, PhD students at EPFL are similar to their US colleagues during the final years of their PhD programs. Similar to these students, EPFL PhD students focus on conducting research rather than taking classes because they fulfilled their course requirements while earning their master's degree.

A few caveats are in order. First, despite the fact that our results on the impact of knowledge specialization on the research productivity of PhD groups hold even after having estimated a dynamic feedback model, perhaps a better approach would be to estimate instrumental variable regression models to fully address the endogeneity of a supervisor's choice. Unfortunately, we did not find plausible instruments that were uncorrelated with the error term. Additionally, although the level of detail in our data allows us to elucidate important aspects of the organization of research groups, they are limited to a single academic institution. Further work might seek to extend these results to other academic institutions, possibly using cross-country data. Finally, although PhD student groups are important contributors to a country's innovative capacity, it would be interesting to extend these types of analyses to the overall organization of a supervisor's laboratory. In the present study, we only control for the characteristics of the non-PhD components of a supervisor's laboratory using the grant money available to pay staff other than PhD students. In the future, we plan to closely examine the entire organization of the professor's laboratory.

References

[1] J. D. Adams, G. C. Black, J. R. Clemmons, and P. E. Stephan, "Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999," *Res. Policy*, vol. 34, no. 3, pp. 259–285, 2005.

[2] A. Agrawal and A. Goldfarb, "Restructuring research: Communication costs and the democratization of university innovation," *Am. Ec. Rev.*, vol. 98, no. 4, pp. 1578–1590, 2008.

[3] M. Amin and M. Mabe, "Impact factors: Use and abuse," *Perspectives in Publishing*, vol. 1, no. 2, pp. 1–6, 2000.

[4] A. Arora and A. Gambardella, "The impact of NSF support on basic research in economics," *Ann. Econ. Statist.*, vol. Special Issue in Honor of Zvi Griliches, pp. 79–80, 2005.

[5] G. S. Becker and K. M. Murphy, "The division of labor, coordination costs, and knowledge," *Q. J. Econ.*, vol. 107, no. 4, pp. 1137–1160, 1992.

[6] S. Belenzon and M. Schankerman, "University knowledge transfer: Private ownership, incentives and local development objectives," *J. Law Econ.*, vol. 52, pp. 111–144, 2009.

[7] G. Black and P. E. Stephan, "The economics of university science and the role of foreign graduate students and postdoctoral scholars," in *American Universities in a Global Market*, C. T. Clotfelter, Ed. University of Chicago Press, 2010.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[9] R. Blundell, G. Rachel, and J. V. Reenen, "Market shares, market value and innovation in a panel of British manufacturing firms," *Rev. of Econ. Stud.*, vol. 66, pp. 529–554, 1999.

[10] R. Blundell, G. Rachel, and F. Windmeijer, "Individual effects and dynamics in count data models," *J. Econometrics*, vol. 108, 2002.

[11] M. Calderini, C. Franzoni, and A. Vezzulli, "If star scientists do not patent. The effect of productivity, basicness and impact on the decision to patent in the academic world," *Res. Policy*, vol. 36, no. 3, pp. 303–319, 2007.

[12] ——, "The unequal benefits of academic patenting for science and engeneering research," *IEEE Trans. Eng. Manag.*, vol. 56, no. 1, pp. 16–30, 2009.

[13] A. C. Cameron and P. K. Trivedi, *Microeconometrics: Methods and applications.* Cambridge University Press, 2005.

[14] P. Dasgupta and P. A. David, "Toward a new economics of science," *Res. Policy*, vol. 23, pp. 487–521, 1994.

[15] W. Ding, S. Levin, P. E. Stephan, and A. Winkler, "The impact of information technology on academic scientists productivity and collaboration patterns," *Manage. Sci.*, vol. 56, no. 9, pp. 1439–1461, 2010.

[16] H. Etzkowitz, "Research groups as quasi-firms: The invention of the entrepreneurial university," *Res. Policy*, vol. 32, pp. 109–121, 2003.

[17] C. Forman and N. van Zeebroeck, "From wires to partners: How the Internet has fostered R&D collaborations within firms," *Manage. Sci.*, vol. 58, no. 8, pp. 1549–1568, 2012.

[18] E. Garfield, "Is citation analysis a legitimate evaluation tool?" *Scientometrics*, vol. 1, no. 4, pp. 359–375, 1979.

[19] P. Gosling and B. Noordam, "Mastering your Ph.D: Mentors, leadership, and community," *Science Career Magazine*, August 2007.

[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, pp. 5228–5235, 2004.

[21] J. Groen and M. Rizzo, "The changing composition of U.S. citizen PhDs," in *Science and the University*, R. Ehrenberg and P. Stephan, Eds. University of Wisconsin Press, 2007, pp. 177–196.

[22] J. A. Groen, M. P. Nagowski, and R. G. Ehrenberg, "PhD attainment of graduates of selective private academic institutions," *Educ. Financ. Policy*, vol. 2, no. 1, pp. 100–110, 2007.

[23] S. Gurmu, G. Black, and P. E. Stephan, "The knowledge production function for university patenting," *Econ. Inqu.*, vol. 48, pp. 192–213, 2010.

[24] D. C. Hambrick, T. S. Cho, and M. Chen, "The influence of top management team heterogeneity on firms' competitive moves," *Admin. Sci. Quart.*, vol. 41, pp. 659–684, 1996.

[25] B. Hamilton, J. A. Nickerson, and H. Owan, "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation," *J. Polit. Econ.*, vol. 111, no. 3, pp. 465–497, 2003.

[26] B. Holmstrom and B. Nalebuff, "To the raider goes the surplus: A re-examination of the free-rider problem," *J. Econ. Manage. Strat.*, vol. 1, no. 1, pp. 37–62, 1992.

[27] A. Jaffe, "Real effects of academic research," *Am. Ec. Rev.*, vol. 79, pp. 957–970, 1989.

[28] B. F. Jones, "The burden of knowledge and the 'death of the Renaissance man': Is innovation getting harder?" *Rev. of Econ. Stud.*, vol. 76, no. 1, pp. 283–317, 2009.

[29] D. C. Knight, L. Pearee, K. G. Smith, J. D. Olian, H. P. Sims, K. A. Smith, and P. Flood, "Top management team diversity, group process, and strategic consensus," *Strategic Manage. J.*, vol. 20, pp. 445–465, 1999.

[30] E. P. Lazear, *Personnel economics for managers*. New York: Wiley, 1998.

[31] ——, "Globalisation and the market for team-mates," *Econ. J.*, vol. 109, no. 454, pp. 15–40, 1999.

[32] S. G. Levin and P. E. Stephan, "Research productivity over the life cycle: Evidence for academic scientists," *Am. Ec. Rev.*, vol. 81, no. 1, pp. 114–132, 1991.

[33] D. P. Libaers, "Role and contribution of foreign-born scientists and engineers to the public U.S. nanoscience and technology research enterprise," *IEEE Trans. Eng. Manag.*, vol. 54, pp. 423–432, 2007.

[34] ——, "Time allocation decisions of academic scientists and their impact on technology commercialization," *IEEE Trans. Eng. Manag.*, vol. 59, no. 4, pp. 705–716, 2012.

[35] W. Ling, "The ideal PhD mentor–A student's perspective," *Science Career Magazine*, Dec. 2002.

[36] V. Mangematin and S. Robin, "The two faces of PhD students: Management of early careers of French PhDs in life sciences," *Sci. Public Policy*, vol. 30, pp. 405–414, 2003.

[37] E. Mansfield, "Academic research and industrial innovation," *Res. Policy*, vol. 20, pp. 1–12, 1991.

[38] G. McMillan and R. Hamilton, "The impact of publicly funded basic research: An integrative extension of Martin and Salter," *IEEE Trans. Eng. Manag.*, vol. 50, no. 2, pp. 184–191, 2003.

[39] R. R. Nelson, "The simple economics of basic scientific research," *J. Polit. Econ.*, vol. 67, no. 3, pp. 297–306, 1959.

[40] J. Pleffer, "Organizational demography," in *Research In Organizational Behavior*, L. Cummings and B. M. Staw, Eds. JAI Press. Greenwhich, CT, 1983.

[41] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14(3), pp. 130–137, 1980.

[42] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proc. of Fourth Int. AAAI Conf. on Weblogs and Social Media*, 2010.

[43] T. Shinn, "Hirarchies des chercheurs et formes de recherche," *Actes Recher. Sci. Soc.*, vol. 74, pp. 2–22, 1988.

[44] J. Singh and L. Fleming, "Lone inventors as sources of breakthroughts: Myth or reality," *Manage. Sci.*, vol. 56, no. 1, pp. 41–56, 2010.

[45] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, "Topic and keyword re-ranking for lda-based topic modeling," in *Proc. of the 18th ACM Conf. on Information and Knowledge Management*. ACM, 2009, pp. 1757–1760.

[46] P. Stephan and S. Levin, "The importance of implicit contracts in collaborative scientific research," in *Science Bought and Sold: Essays in the Economics of Science*, P. Mirowski and E.-M. Sent, Eds. University of Chicago Press, 2002.

[47] P. E. Stephan, *How economics shapes science*. Harvard University Press, 2012.

[48] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent semantic analysis: A road to meaning*, T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, Eds. Laurence Erlbaum, 2007.

[49] E. T. Stuen, A. M. Mobarak, and K. E. Maskus, "Foreign graduate students and knowledge creation at U.S. universities: Evidence from enrollment fluctuations," 2007, working paper. University of Colorado.

[50] F. Waldinger, "Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany," *J. Polit. Econ.*, vol. 118, no. 4, pp. 787–831, 2010.

[51] K. Y. Williams and C. O'Reilly, "Forty years of diversity research: A review," in *Research In Organizational Behavior*, B. M. Staw and L. L. Cummings, Eds. Greenwhich, CT: JAI Press., 1998.

[52] S. Wuchty, J. F. Benjamin, and B. Uzzi, "The increasing dominance of teams in the production of knowledge," *Science*, vol. 316, pp. 1036–1039, 2007.

[53] T. R. Zenger and B. S. Lawrence, "Organizational demography: The differential effects of age and tenure distributions on technical communication," *Acade. Manage. J.*, vol. 32(2), pp. 353–376, 1989.

Figure 1: Density distribution of group publications and highly cited group publications

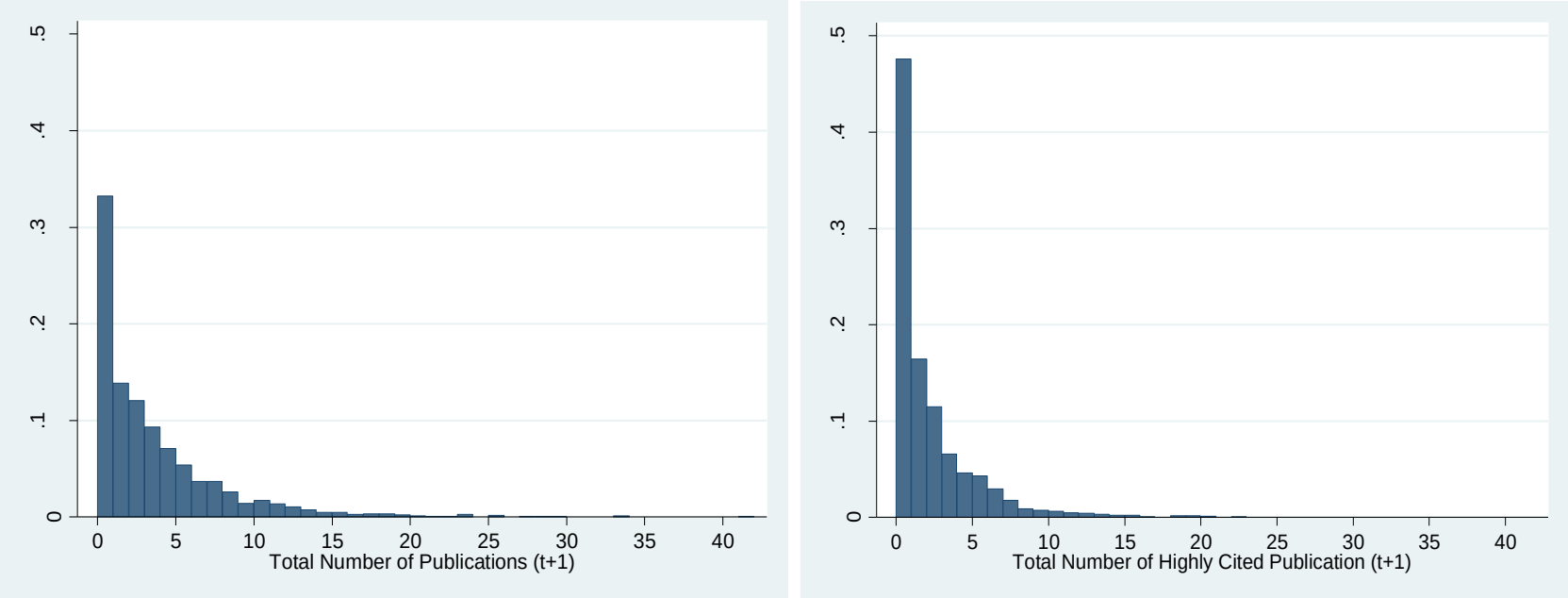

Table 1: An illustration of six (out of 120) topics extracted from the abstracts' corpus

Topic 16 *terms*	**Topic 21** *terms*	**Topic 25** *terms*	**Topic 73** *terms*	**Topic 99** *terms*	**Topic 111** *terms*
WATER	PHYSIC	POLYM	ION	FREQUENC	NUMER
FLOW	DECAI	POLYMER	COMPOUND	PERIOD	EQUAT
SLOPE	LEVEL	CHAIN	EXCHANG	RESON	SOLUT
RIVER	MEASUR	STABIL	METAL	OSCIL	METHOD
HYDRAUL	CP	CONCENTR	SOLUT	AMPLITUD	FINIT
SEDIMENT	STANDARD	MONOM	STABIL	EXPERIMENT	NONLINEAR
FLOOD	EXPERI	VISCOS	PH	MEASUR	SPACE
HYDROLOG	BS	DISPERS	IONIC	NONLINEAR	ELEMENT
NATUR	DETECTOR	COPOLYM	RELAX	HARMON	APPROXIM
CATCHMENT	THEORI	ENCAPSUL	PROTON	RING	LOCAL

Table 2: Summary statistics

Variable	Description	Mean	Std. Dev.	Min.	Max.
# of publications (t+1)	# of PhD student group publications at *t+1*	3.12	4.23	0.00	42.00
# of highly cited publications (t+1)	# of PhD student group highly cited publications at *t+1*	1.78	2.82	0.00	23.00
# of publications (t+2)	# of PhD student group publications at time *t+2*	3.39	4.56	0.00	34.00
# of highly cited publications (t+2)	# of PhD student group highly cited publications at *t+2*	1.86	2.94	0.00	23.00
Within-group specialization (N) (120 topics)	Normalized HI applied to topic counts (120 topics)	89.01	10.70	0.00	100.00
Within-group specialization (N) (100 topics)	Normalized HI applied to topic counts (100 topics)	92.98	6.11	0.00	100.00
Group research breadth (120 topics)	# of topics investigated by the PhD student group (120 topics)	8.19	4.08	1.00	25.00
Group research breadth (100 topics)	# of topics investigated by the PhD student group (100 topics)	7.93	3.51	1.00	22.00
Characteristics of the PhD student group					
Group size	# of PhD student group members	4.61	2.88	2.00	30.00
Mean group tenure	Mean tenure of the PhD student group members	2.50	0.78	1.00	4.00
# students with research award	# of students awarded a prize for the best dissertation	0.13	0.48	0.00	4.00
Mean group age	Average age of PhD student group members	29.21	2.09	23.67	47.00
StdDev group age	Dispersion of PhD student group members' age	2.55	1.99	0.00	31.11
Master background diversity	# of different masters' universities	3.25	1.86	1.00	19.00
Thesis is co-supervised	Dummy = 1 if at least one thesis is co-supervised	0.10	0.30	0.00	1.00
Professor PhD group publications (t-1)	# of Supervisor PhD group publications (t-1)	3.23	4.41	0.00	42.00
Professor PhD group highly cited publications (t-1)	# of Supervisor PhD group highly cited publications (t-1)	3.57	4.79	0.00	42.00
Characteristics of the group's supervisor					
Professor Age	Supervisor's age	49.72	8.37	21.00	71.00
Professor is Swiss	Dummy =1 if the supervisor is Swiss	0.49	0.50	0.00	1.00
Professor is German	Dummy =1 if the supervisor is German	0.12	0.32	0.00	1.00
Professor is Italian	Dummy =1 if the supervisor is Italian	0.03	0.17	0.00	1.00
Professor is French	Dummy =1 if the supervisor is French	0.10	0.30	0.00	1.00
Professor knowledge capital stock	Stock of pre-sample publications	40.90	46.97	0.00	370.00
Professor publications (t-1)	# of supervisor's publications at *t-1*	6.30	6.20	0.00	45.00
SNSF grants (in real Swiss Francs)	SNSF money received by the supervisor in the five years prior to the creation of the group	188,000	279,000	0.00	2,250,000
Characteristics of the department					
Department funds (in real Swiss Francs)	Budget assigned to a department in *t*	8,203.17	5,773.82	0.00	22,174.46

Table 3: Regressions for the number of publications

	N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+2)		N. Pubs (t+2)		N. Pubs (t+2)	
Within-group specialization (N)	0.0319	***	0.0312	***	0.0302	***	0.0270	**	0.0263	**	0.0316	***
	(0.0107)		(0.0103)		(0.0099)		(0.0118)		(0.0116)		(0.0119)	
Within-group specialization (N)^2	-0.0003	***	-0.0003	***	-0.0002	**	-0.0002	**	-0.0002	**	-0.0002	**
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristics of the PhD student group												
Group size	0.1930	***	0.1767	***	0.163	***	0.1749	***	0.1596	***	0.1264	***
	(0.0535)		(0.0498)		(0.0411)		(0.0517)		(0.0494)		(0.0424)	
Group size^2	-0.0040	**	-0.0035	**	-0.0031	***	-0.0030	**	-0.0025	*	-0.0017	
	(0.0016)		(0.0015)		(0.0012)		(0.0014)		(0.0014)		(0.0012)	
Group research breadth	0.0511		0.048		0.0200		0.0632		0.0626		0.0601	
	(0.0438)		(0.0399)		(0.0385)		(0.0480)		(0.0448)		(0.0417)	
Group research breadth^2	-0.0022	*	-0.0021	*	-0.0014		-0.0026	*	-0.0026	**	-0.0025	**
	(0.0013)		(0.0012)		(0.0011)		(0.0014)		(0.0013)		(0.0013)	
Mean group tenure	0.8918	***	0.9399	***	0.7373	**	0.5699	**	0.6050	***	0.4379	*
	(0.2091)		(0.2110)		(0.2864)		(0.2267)		(0.2111)		(0.2510)	
Mean group tenure^2	-0.1816	***	-0.1982	***	-0.1666	***	-0.1533	***	-0.1669	***	-0.1325	***
	(0.0397)		(0.0408)		(0.0528)		(0.0429)		(0.0397)		(0.0455)	
N students with research award	0.1562	***	0.1421	***	0.0922	***	0.1706	***	0.1600	***	0.1113	***
	(0.0347)		(0.0305)		(0.0318)		(0.0364)		(0.0337)		(0.0359)	
Mean group age	-0.0610	**	-0.0591	**	-0.0543	***	-0.0543	**	-0.0522	**	-0.0376	**
	(0.0237)		(0.0231)		(0.0210)		(0.0225)		(0.0220)		(0.0184)	
StdDev group age	0.0231		0.0252		0.0276		0.0166		0.0180		0.0164	
	(0.0227)		(0.0212)		(0.0191)		(0.0230)		(0.0220)		(0.0192)	
Master background diversity	0.0777	**	0.0785	**	0.1117	***	0.0983	**	0.0992	**	0.1121	**
	(0.0385)		(0.0374)		(0.0388)		(0.0412)		(0.0404)		(0.0441)	
Master background diversity^2	-0.0028		-0.0032	*	-0.0068	***	-0.0041	**	-0.0044	**	-0.007	***
	(0.0017)		(0.0017)		(0.0021)		(0.0019)		(0.0019)		(0.0023)	
Thesis is co-supervised	-0.2792	***	-0.2646	***	-0.1925	***	-0.3107	***	-0.3003	***	-0.2593	***
	(0.0766)		(0.0737)		(0.0659)		(0.0811)		(0.0800)		(0.0737)	
Professor PhD group publications *(t-1)*					0.0399	***					0.0081	
					(0.0066)						(0.0051)	
Characteristics of the group's supervisor												
Professor age	-0.0156	***	-0.0133	***	-0.0134	***	-0.0145	***	-0.0124	**	-0.0115	***
	(0.0047)		(0.0045)		(0.0037)		(0.0051)		(0.0050)		(0.0042)	
Professor is Swiss	-0.0532		-0.0439		-0.0059		-0.0673		-0.0588		-0.0315	
	(0.0892)		(0.0834)		(0.0690)		(0.0908)		(0.0867)		(0.0691)	
Professor is German	-0.0959		-0.0771		-0.0239		-0.1392		-0.1242		-0.0587	
	(0.1131)		(0.1023)		(0.0874)		(0.1145)		(0.1052)		(0.0902)	
Professor is Italian	0.1010		0.1624		0.1754		0.1945		0.2467		0.2503	*
	(0.1986)		(0.1775)		(0.1364)		(0.2027)		(0.1809)		(0.1474)	
Professor is French	-0.1124		-0.1032		-0.0756		-0.1181		-0.1109		-0.0659	
	(0.1104)		(0.1006)		(0.0939)		(0.1125)		(0.1051)		(0.0950)	
Professor knowledge capital stock	0.0052	***	0.0044	***	0.0039	***	0.0053	***	0.0046	***	0.0042	***
	(0.0017)		(0.0017)		(0.0014)		(0.0016)		(0.0016)		(0.0014)	
Professor knowledge capital stock^2	-0.0001	**	-0.0001	**	-0.0001	**	-0.0001	***	-0.0001	***	-0.0001	***
	(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Professor publications *(t-1)*			0.0188	***	0.0105	**			0.0167	***	0.0081	
			(0.0051)		(0.0045)				(0.0058)		(0.0051)	
SNSF grants	0.0003	**	0.0003	**	0.0002	*	0.0003	**	0.0002	**	0.0002	**
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristics of the department												
Department funds	-0.0144	**	-0.0144	**	-0.0118	*	-0.0114		-0.0113		-0.0069	
	(0.0073)		(0.0072)		(0.0066)		(0.0078)		(0.0078)		(0.0071)	
Department fixed effect	YES	***	YES	***	YES	***	YES	***	YES	***	YES	***
Characteristcs of EPFL												
Time trend	0.1831	***	0.1754	***	0.0105	**	0.1764	***	0.1700	***	0.1312	***
	(0.0130)		(0.0131)		(0.0045)		(0.0132)		(0.0138)		(0.0131)	
Constant	-1.8369	**	-1.9115	**	-1.5893	**	-1.1276		-1.1991		-1.3826	*
	(0.8171)		(0.7811)		(0.7847)		(0.8946)		(0.8398)		(0.7391)	
Observations	1938		1938		1689		1938		1938		1689	
Log-likelihood	-4076		-4042		-3638		-4361		-4334		-3845	

Note: Standard errors are in parentheses, clustered around supervisors. * 0.10% ** 0.05% ***0.01%.

Table 4: Regressions for the number of highly cited publications

	N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+2)		N. Pubs (t+2)		N. Pubs (t+2)	
Within-group specialization (N)	0.0329	***	0.0325	***	0.0308	***	0.0282	***	0.0276	***	0.0337	***
	(0.0103)		(0.0101)		(0.0088)		(0.0103)		(0.0100)		(0.0089)	
Within-group specialization $(N)^2$	-0.0003	**	-0.0002	**	-0.0002	**	-0.0003	**	-0.0003	**	-0.0003	***
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristics of the PhD student group												
Group size	0.2644	***	0.2461	***	0.2314	***	0.2397	***	0.2209	***	0.1777	***
	(0.0584)		(0.0547)		(0.0461)		(0.0600)		(0.0568)		(0.0510)	
Group $size^2$	-0.0069	***	-0.0063	***	-0.0057	***	-0.0068	***	-0.0061	***	-0.0045	***
	(0.0021)		(0.0020)		(0.0017)		(0.0019)		(0.0018)		(0.0017)	
Group research breadth	0.0186		0.0173		-0.0122		0.0406		0.0417		0.0539	
	(0.0475)		(0.0438)		(0.0417)		(0.0511)		(0.0471)		(0.0437)	
Group research $breadth^2$	-0.0015		-0.0015		-0.0006		-0.0015		-0.0015		-0.0020	
	(0.0015)		(0.0014)		(0.0013)		(0.0015)		(0.0014)		(0.0014)	
Mean group tenure	0.8537	***	0.9095	***	0.5955	**	0.5622	**	0.6097	**	0.3905	
	(0.2513)		(0.2505)		(0.2844)		(0.2683)		(0.2577)		(0.3257)	
Mean group $tenure^2$	-0.1906	***	-0.2096	***	-0.1534	***	-0.1557	***	-0.173	***	-0.1263	**
	(0.0468)		(0.0467)		(0.0514)		(0.0526)		(0.0508)		(0.0603)	
N students with research award	0.1662	***	0.1526	***	0.1021	***	0.1728	***	0.1622	***	0.1165	***
	(0.0342)		(0.0312)		(0.0290)		(0.0388)		(0.0362)		(0.0420)	
Mean group age	-0.0212		-0.0183		-0.0253		-0.0269		-0.0241		-0.0270	
	(0.0305)		(0.0299)		(0.0246)		(0.0276)		(0.0271)		(0.0223)	
StdDev group age	0.0082		0.0104		0.0143		0.0098		0.0115		0.0169	
	(0.0312)		(0.0294)		(0.0263)		(0.0300)		(0.0283)		(0.0250)	
Master background diversity	0.0649		0.0626		0.0961	**	0.0778		0.0767		0.0824	*
	(0.0477)		(0.0460)		(0.0421)		(0.0477)		(0.0471)		(0.0478)	
Master background $diversity^2$	-0.0012		-0.0015		-0.0056	**	-0.0014		-0.0017		-0.0039	
	(0.0024)		(0.0024)		(0.0023)		(0.0024)		(0.0024)		(0.0025)	
Thesis is co-supervised	-0.2735	***	-0.2588	***	-0.1986	**	-0.2896	***	-0.2782	***	-0.2622	***
	(0.0985)		(0.0961)		(0.0802)		(0.0948)		(0.0942)		(0.0806)	
Professor PhD group publications (t-1)					0.0625	***					0.0489	***
					(0.0089)						(0.0090)	
Characteristics of the group's supervisor												
Professor age	-0.0342	***	-0.0318	***	-0.0275	***	-0.0339	***	-0.0317	***	-0.027	***
	(0.0059)		(0.0058)		(0.0047)		(0.0060)		(0.0059)		(0.0052)	
Professor is Swiss	0.0403		0.0467		0.0957		0.0250		0.0316		0.0401	
	(0.1122)		(0.1073)		(0.0891)		(0.1131)		(0.1092)		(0.0923)	
Professor is German	0.0342		0.0571		0.0774		0.0316		0.0501		0.0432	
	(0.1519)		(0.1390)		(0.1037)		(0.1412)		(0.1305)		(0.1043)	
Professor is Italian	0.2920		0.3505	*	0.3829	**	0.4128	**	0.4668	***	0.4339	***
	(0.2126)		(0.1977)		(0.1566)		(0.1761)		(0.1590)		(0.1260)	
Professor is French	-0.0127		-0.0086		0.0346		-0.1035		-0.1003		-0.0477	
	(0.1437)		(0.1387)		(0.1212)		(0.1512)		(0.1461)		(0.1338)	
Professor knowledge capital stock	0.0075	***	0.0067	***	0.0052	***	0.0077	***	0.0069	***	0.0057	***
	(0.0020)		(0.0021)		(0.0016)		(0.0019)		(0.0019)		(0.0016)	
Professor knowledge capital $stock^2$	-0.0001	**	-0.0001	**	-0.0001	**	-0.0001	***	-0.0001	***	-0.0001	***
	(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Professor publications *(t-1)*			0.0202	***	0.0132	***			0.0189	***	0.0124	**
			(0.0053)		(0.0047)				(0.0058)		(0.0056)	
SNSF grants	0.0004	***	0.0003	***	0.0002	**	0.0003	**	0.0002	**	0.0002	**
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristics of the department												
Department funds	-0.0181	*	-0.0184	*	-0.015	*	-0.0136		-0.0139		-0.0118	
	(0.0099)		(0.0097)		(0.0087)		(0.0089)		(0.0087)		(0.0077)	
Department fixed effect	YES	***	YES	***	YES	***	YES	***	YES	***		
Characteristcs of EPFL												
Time trend	0.1911	***	0.1831	***	0.1526	***	0.1725	***	0.1654	***	0.136	***
	(0.0144)		(0.0144)		(0.0141)		(0.0136)		(0.0138)		(0.0136)	
Constant	-2.9785	***	-3.0696	***	-2.3102	***	-1.5452	*	-1.6171	*	-1.3200	
	(0.8956)		(0.8875)		(0.8182)		(0.9068)		(0.8613)		(0.8127)	
Observations	1938		1938		1689		1938		1938		1689	
Log-likelihood	-3141		-3120		-2790		-3331		-3312		-2974	

Note: Standard errors are in parentheses, clustered around supervisors. * 0.10% ** 0.05% ***0.01%.

Table 5: Regressions for the number of publications (with 100 topics)

	N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+2)		N. Pubs (t+2)		N. Pubs (t+2)	
Within-group specialization (N)	0.1648	**	0.1486	**	0.1074	*	0.1906	**	0.1771	**	0.1352	**
	(0.0720)		(0.0667)		(0.0588)		(0.0751)		(0.0713)		(0.0639)	
Within-group specialization (N)^2	-0.0010	**	-0.0009	**	-0.0006		-0.0011	**	-0.0010	**	-0.0007	*
	(0.0005)		(0.0004)		(0.0004)		(0.0005)		(0.0005)		(0.0004)	
Characteristics of the PhD student group												
Group size	0.2429	***	0.2213	***	0.2024	***	0.2439	***	0.2247	***	0.1978	***
	(0.0515)		(0.0496)		(0.0459)		(0.0473)		(0.0456)		(0.0431)	
Group size^2	-0.0061	***	-0.0054	***	-0.0048	***	-0.0056	***	-0.0049	***	-0.0043	***
	(0.0016)		(0.0015)		(0.0013)		(0.0014)		(0.0013)		(0.0012)	
Group research breadth	-0.0261		-0.0248		-0.0338		-0.0358		-0.0336		-0.0355	
	(0.0315)		(0.0310)		(0.0300)		(0.0316)		(0.0309)		(0.0301)	
Group research breadth^2	0.0004		0.0004		0.0005		0.0006		0.0006		0.0005	
	(0.0006)		(0.0006)		(0.0006)		(0.0006)		(0.0006)		(0.0006)	
Mean group tenure	0.8980	***	0.9433	***	0.7141	**	0.6127	***	0.6435	***	0.4399	*
	(0.2121)		(0.2141)		(0.2914)		(0.2316)		(0.2181)		(0.2526)	
Mean group tenure^2	-0.1827	***	-0.1985	***	-0.1627	***	-0.1610	***	-0.1735	***	-0.1332	***
	(0.0402)		(0.0413)		(0.0536)		(0.0438)		(0.0411)		(0.0459)	
N students with research award	0.1444	***	0.1316	***	0.0849	***	0.1568	***	0.1473	***	0.1017	***
	(0.0357)		(0.0315)		(0.0329)		(0.0381)		(0.0353)		(0.0379)	
Mean group age	-0.0608	***	-0.0586	**	-0.0539	***	-0.0561	**	-0.0538	**	-0.0388	**
	(0.0233)		(0.0228)		(0.0207)		(0.0224)		(0.0220)		(0.0185)	
StdDev group age	0.0272		0.0283		0.0290		0.0230		0.0236		0.0202	
	(0.0210)		(0.0198)		(0.0180)		(0.0213)		(0.0206)		(0.0180)	
Master background diversity	0.0768	**	0.0797	**	0.1114	***	0.0994	**	0.1019	**	0.1143	**
	(0.0386)		(0.0379)		(0.0395)		(0.0415)		(0.0410)		(0.0451)	
Master background diversity^2	-0.0030	*	-0.0034	**	-0.0069	***	-0.0044	**	-0.0048	**	-0.0072	***
	(0.0017)		(0.0017)		(0.0021)		(0.0019)		(0.0019)		(0.0024)	
Thesis is co-supervised	-0.2742	***	-0.2562	***	-0.1849	***	-0.3075	***	-0.2945	***	-0.2574	***
	(0.0744)		(0.0719)		(0.0641)		(0.0801)		(0.0791)		(0.0725)	
Professor PhD group publications (t-1)					0.0395	***					0.0359	***
					(0.0064)						(0.0058)	
Characteristics of the group's supervisor												
Professor age	-0.0157	***	-0.0134	***	-0.0133	***	-0.0147	***	-0.0127	**	-0.0118	***
	(0.0047)		(0.0046)		(0.0037)		(0.0051)		(0.0050)		(0.0042)	
Professor is Swiss	-0.0716		-0.0651		-0.0298		-0.0753		-0.0691		-0.0436	
	(0.0883)		(0.0832)		(0.0693)		(0.0896)		(0.0862)		(0.0696)	
Professor is German	-0.0819		-0.0653		-0.0130		-0.1280		-0.1144		-0.0519	
	(0.1114)		(0.1017)		(0.0873)		(0.1136)		(0.1052)		(0.0911)	
Professor is Italian	0.0796		0.1368		0.1567		0.1810		0.2277		0.2416	*
	(0.1872)		(0.1675)		(0.1304)		(0.1872)		(0.1667)		(0.1353)	
Professor is French	-0.1284		-0.1211		-0.0912		-0.1328		-0.1274		-0.0828	
	(0.1059)		(0.0976)		(0.0894)		(0.1093)		(0.1033)		(0.0924)	
Professor knowledge capital stock	0.0050	***	0.0042	**	0.0036	***	0.0050	***	0.0043	***	0.004	***
	(0.0017)		(0.0017)		(0.0013)		(0.0016)		(0.0016)		(0.0013)	
Professor knowledge capital stock^2	-0.0001	**	-0.0001	**	-0.0001	**	-0.0001	***	-0.0001	***	-0.0001	***
	(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Professor publications (t-1)			0.018	***	0.0096	**			0.0157	***	0.0073	
			(0.0051)		(0.0046)				(0.0057)		(0.0050)	
SNSF grants	0.0003	**	0.0003	**	0.0002	*	0.0003	**	0.0002	**	0.0002	**
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristcs of the department												
Department funds	-0.0097		-0.0099		-0.0071		-0.0069		-0.0070		-0.0023	
	(0.0075)		(0.0073)		(0.0068)		(0.0078)		(0.0077)		(0.0070)	
Department fixed effect	YES	***	YES	***	YES	***	YES	***	YES	***		
Characteristcs of EPFL												
Time trend	0.1808	***	0.1733	***	0.1419	***	0.1744	***	0.1683	***	0.1295	***
	(0.0130)		(0.0130)		(0.0133)		(0.0132)		(0.0137)		(0.0133)	
Constant	-7.7611	***	-7.1265	***	-5.1349	**	-8.4990	***	-7.9644	***	-6.1794	**
	(2.9618)		(2.7422)		(2.4315)		(2.9835)		(2.8318)		(2.4722)	
Observations	1938		1938		1689		1938		1938		1689	
Log-likelihood	-4075		-4045		-3641		-4346		-4322		-3842	

Note: Standard errors are in parentheses, clustered around supervisors. * 0.10% ** 0.05% ***0.01%.

Table 6: Regressions for the number of highly cited publications (with 100 topics)

	N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+1)		N. Pubs (t+2)		N. Pubs (t+2)		N. Pubs (t+2)	
Within-group specialization (N)	0.0671		0.0551		0.0372		0.1169		0.1061		0.0898	
	(0.0771)		(0.0733)		(0.0610)		(0.0746)		(0.0709)		(0.0646)	
Within-group specialization (N)^2	-0.0004		-0.0003		-0.0002		-0.0007		-0.0007		-0.0006	
	(0.0005)		(0.0005)		(0.0004)		(0.0005)		(0.0005)		(0.0004)	
Characteristics of the PhD student group												
Group size	0.3033	***	0.2786	***	0.2602	***	0.3009	***	0.2765	***	0.2327	***
	(0.0613)		(0.0592)		(0.0545)		(0.0572)		(0.0553)		(0.0527)	
Group size^2	-0.0085	***	-0.0077	***	-0.0069	***	-0.0088	***	-0.008	***	-0.0064	***
	(0.0023)		(0.0022)		(0.0019)		(0.0020)		(0.0019)		(0.0018)	
Group research breadth	-0.0225		-0.0200		-0.0355		-0.0214		-0.0176		-0.0121	
	(0.0361)		(0.0355)		(0.0347)		(0.0351)		(0.0342)		(0.0327)	
Group research breadth^2	0.0001		0.0000		0.0003		0.0002		0.0001		0.0000	
	(0.0008)		(0.0007)		(0.0007)		(0.0008)		(0.0007)		(0.0007)	
Mean group tenure	0.8274	***	0.8823	***	0.5527	*	0.5647	**	0.6077	**	0.3690	
	(0.2515)		(0.2522)		(0.2883)		(0.2724)		(0.2629)		(0.3293)	
Mean group tenure^2	-0.1862	***	-0.2052	***	-0.1464	***	-0.1565	***	-0.1729	***	-0.1231	**
	(0.0469)		(0.0471)		(0.0521)		(0.0537)		(0.0521)		(0.0610)	
N students with research award	0.1578	***	0.1456	***	0.0963	***	0.169	***	0.1589	***	0.1122	**
	(0.0337)		(0.0301)		(0.0299)		(0.0408)		(0.0381)		(0.0449)	
Mean group age	-0.0210		-0.0177		-0.0250		-0.0298		-0.0268		-0.0278	
	(0.0301)		(0.0297)		(0.0245)		(0.0274)		(0.0271)		(0.0223)	
StdDev group age	0.0108		0.0118		0.0140		0.0145		0.0154		0.0188	
	(0.0296)		(0.0282)		(0.0253)		(0.0283)		(0.0269)		(0.0237)	
Master background diversity	0.0617		0.0616		0.0937	**	0.0734		0.0741		0.0813	*
	(0.0480)		(0.0461)		(0.0430)		(0.0484)		(0.0479)		(0.0489)	
Master background diversity^2	-0.0011		-0.0015		-0.0055	**	-0.0013		-0.0016		-0.0039	
	(0.0024)		(0.0024)		(0.0023)		(0.0024)		(0.0024)		(0.0025)	
Thesis is co-supervised	-0.2648	***	-0.2456	***	-0.1838	**	-0.2961	***	-0.2812	***	-0.267	***
	(0.0956)		(0.0935)		(0.0792)		(0.0958)		(0.0950)		(0.0833)	
Professor PhD group publications (t-1)					0.0614	***					0.0471	***
					(0.0088)						(0.0091)	
Characteristics of the group's supervisor												
Professor age	-0.0337	***	-0.0313	***	-0.0267	***	-0.0337	***	-0.0314	***	-0.0269	***
	(0.0060)		(0.0059)		(0.0047)		(0.0061)		(0.0060)		(0.0053)	
Professor is Swiss	0.0163		0.0200		0.0631		0.0100		0.0154		0.0210	
	(0.1133)		(0.1087)		(0.0918)		(0.1140)		(0.1106)		(0.0943)	
Professor is German	0.0362		0.0569		0.0762		0.0161		0.0333		0.0279	
	(0.1486)		(0.1366)		(0.1033)		(0.1375)		(0.1277)		(0.1037)	
Professor is Italian	0.2641		0.3221	*	0.3565	**	0.3797	**	0.4331	***	0.4107	***
	(0.2082)		(0.1944)		(0.1574)		(0.1739)		(0.1573)		(0.1271)	
Professor is French	-0.0278		-0.0257		0.0139		-0.1199		-0.1180		-0.0725	
	(0.1392)		(0.1362)		(0.1147)		(0.1482)		(0.1443)		(0.1289)	
Professor knowledge capital stock	0.0071	***	0.0063	***	0.0048	***	0.0075	***	0.0067	***	0.0055	***
	(0.0020)		(0.0020)		(0.0015)		(0.0018)		(0.0019)		(0.0016)	
Professor knowledge capital stock^2	-0.0001	**	-0.0001	**	-0.0001	**	-0.0001	***	-0.0001	***	-0.0001	***
	(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Professor publications (t-1)			0.0198	***	0.0128	***			0.0185	***	0.0124	**
			(0.0053)		(0.0046)				(0.0058)		(0.0057)	
SNSF grants	0.0004	***	0.0003	***	0.0002	**	0.0003	**	0.0002	**	0.0002	**
	(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)		(0.0001)	
Characteristcs of the department												
Department funds	-0.0141		-0.0146		-0.0113		-0.0104		-0.0109		-0.0075	
	(0.0098)		(0.0095)		(0.0089)		(0.0088)		(0.0086)		(0.0079)	
Department fixed effect	YES	***	YES	***	YES	***	YES	***	YES	***	YES	***
Characteristcs of EPFL												
Time trend	0.1896	***	0.1816	***	0.1511	***	0.1725	***	0.1655	***	0.0124	**
	(0.0143)		(0.0143)		(0.0144)		(0.0137)		(0.0139)		(0.0057)	
Constant	-4.6466		-4.1766		-2.8435		-5.4311	*	-4.9931	*	-3.7714	
	(3.2497)		(3.0949)		(2.5116)		(3.0155)		(2.8806)		(2.5575)	
Observations	1938		1938		1689		1938		1938		1689	
Log-likelihood	-3149		-3128		-2800		-3323		-3305		-2973	

Note: Standard errors are in parentheses, clustered around supervisors. * 0.10% ** 0.05% ***0.01%.