NBER WORKING PAPER SERIES

A TALE OF TWO FIELDS? STEM CAREER OUTCOMES

Xuan Jiang
Joseph Staudt
Bruce A. Weinberg

Working Paper 31835
http://www.nber.org/papers/w31835

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

Thanks to Cheryl Grim, Elisabeth Perlman, participants at the NBER Investments in Early Career Scientists Meeting, the 2022 International Conference on the Science of Science & Innovation (ICSSI), and the 2023 Society of Labor Economists. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7507193: CBDRB-FY22-CES007-003, CBDRB-FY22-CES007-012). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Tale of Two Fields? STEM Career Outcomes
Xuan Jiang, Joseph Staudt, and Bruce A. Weinberg
NBER Working Paper No. 31835
November 2023
JEL No. I23,J24,O3

ABSTRACT

Is the labor market for US researchers experiencing the best or worst of times? This paper analyzes the market for recently minted Ph.D. recipients using supply-and-demand logic and data linking graduate students to their dissertations and W2 tax records. We also construct a new dissertation-industry "relevance" measure, comparing dissertation and patent text and linking patents to assignee firms and industries. We find large disparities across research fields in placement (faculty, postdoc, and industry positions), earnings, and the use of specialized human capital. Thus, it appears to simultaneously be a good time for some fields and a bad time for others.

Xuan Jiang
Department of Economics
Jinan University
School of Economics
Guangzhou 510000
China
jiangxuan@jnu.edu.cn

Joseph Staudt
Center for Economic Studies
U.S. Census Bureau
joseph.staudt@census.gov

Bruce A. Weinberg
The Ohio State University
Department of Economics
410 Arps Hall
1945 North High Street
Columbus, OH 43210
and NBER
weinberg.27@osu.edu

"I know we all want the U.S. to continue to be the world's center for innovation. But our position is at risk. There are many reasons for this but two stand out. First, U.S. companies face a severe shortfall of scientists and engineers with the expertise to develop the next generation of breakthroughs..."—Bill Gates, Testimony to Congress, 2008.[1]

"I think in all good conscience, PhD programmes now know that they're taking on more people than have a high probability of getting research positions...we just can't keep on growing these PhD programmes without good outlets... I think if you train a lot of people who don't end up in research positions, the PhD, as I've understood it and as we've always discussed it, is to train people in research, so that's how I think it should be assessed."—Paula Stephan, *Nature*, 2019.[2]

1 Introduction

Is the labor market for researchers in the United States experiencing the best or worst of times? Many policymakers, institutions, and practitioners see the STEM labor market as experiencing the worst of times. They (rightly) note that there has been a large expansion in the share of Ph.D. recipients placed in increasingly long and poorly-compensated postdoc positions. At the same time, a growing number of Ph.D. recipients appear to be pushed by excess supply into a wide range of jobs that do not seem to utilize the specialized human capital they accumulated during graduate school (Cyranoski et al., 2011; Stephan, 2012b; Alberts et al., 2015; Gould, 2015). These trends are especially acute in fields such as Biology, where the fraction of Ph.D. recipients taking postdoc positions has increased particularly rapidly (Powell, 2015; Heggeness et al., 2017).

Others view the STEM labor market as experiencing the best of times. They emphasize (again, rightly) that Ph.D. recipients have low unemployment rates after receiving their degrees. For this camp, long postdocs do not indicate supply outstripping demand, but steep training requirements arising from the ever-growing complexity of science (Jones, 2009). Moreover, far from viewing the wide range of jobs in which STEM Ph.D. recipients place as evidence of oversupply, they view it as evidence of demand-pull for STEM knowledge across a myriad of economic sectors, with STEM being a new, multi-purpose training (Mathur et al., 2015; Meyers et al., 2016). This camp views these facts as support for, if anything, training more researchers.[3]

[1] https://news.microsoft.com/2008/03/12/bill-gates-testimony-before-the-committee-on-science-and-technology-u-s-house-of-representatives/

[2] https://www.nature.com/articles/d41586-019-03439-x

[3] They often also emphasize that innovation drives long-term economic growth (Romer, 1990)and that increasing the number of researchers yields social benefits beyond the effects on the labor market for researchers.

Thus, though advocates of these polar positions point to a similar set of facts, they provide starkly different interpretations. Intuitively, most of the debate has focused on quantities, but economic logic implies that one cannot distinguish supply from demand solely from equilibrium quantities. We apply this economic logic along with newly constructed data to weigh the supply-push "worst of times" and the demand-pull "best of times" narratives. We find that Ph.D. recipients are neither universally experiencing the best nor worst of times. Rather, labor market conditions and outcomes vary markedly across fields. Indeed, Ph.D. recipients from some fields have a reasonable chance of getting faculty positions at universities or placing in industry jobs with relatively high earnings that use their specialized human capital. In other fields, a high fraction of Ph.D. recipients takes postdoc positions and/or enter industry jobs with relatively low earnings that appear further from their areas of expertise.

We construct our data by first linking graduate students from UMETRICS to their Ph.D. dissertations in ProQuest. This allows us to determine each Ph.D. recipient's degree year and field. We then link each Ph.D. recipient to their labor market outcomes using the universe of W2 tax records from the Internal Revenue Service (IRS). This allows us to track post-degree employment and earnings. Specifically, we track earnings during the first three years after a Ph.D. recipient earns their degree as well as whether their initial placement is in one of three mutually exclusive job types: faculty members at a university, postdocs at a university, or industry.[4]

We begin by analyzing academic placements by computing the share of Ph.D. recipients from each field who take jobs as university faculty. While recognizing that faculty positions vary in quality, we take a large share of faculty placements as evidence of strong demand for faculty in a field. In other words, a high fraction of faculty placements is one characteristic of a field experiencing the best of times. We find large disparities across fields, suggesting considerable variation in demand for faculty. On the high end, there appears to be strong demand for faculty (relative to supply) in the Social Sciences, Medicine, and Comm/Info Sciences, which place about 35.4%, 24.4%, and 24.3% of Ph.D. recipients in faculty positions (Figure 2). On the low end, there appears to be weak relative demand for faculty in Chemistry, Geosciences, and Biology, which place about 2.5%, 7.7%, and 9.1% in faculty positions.

We next turn to placements in postdoc positions at universities. In our data, fields with a relatively high fraction of Ph.D. recipients taking postdoc positions include Biology (58.2%), Eco/Envr Sciences (57.9%), and Geosciences (52.8%) (Figure 2). On the low end, are Engineering (22.5%), Physics (37.8%), and Social Sciences (41.4%). Though low earnings

[4]Our "industry" category includes all placements in the private sector, government, and NGOs.

levels in fields where postdocs are prevalent are consistent with low demand relative to supply,[5] they are also consistent with greater training requirements. However, the logic of human capital theory implies that earnings *growth* should be higher in fields where postdocs are prevalent if additional training is required. By contrast, we find that fields in which postdocs are prevalent tend to have not only lower earnings levels but also lower earnings growth. Specifically, a 10 percentage point increase in the share of Ph.D. recipients from a field that place in postdoc positions is associated with a $3,660 reduction in year 3 earnings (Figure 3a) and a 7.7 percentage point reduction in 3-year wage growth (Figure 3b). Thus, it seems likely that fields where postdocs are prevalent are experiencing the worst of times.[6]

Of course, universities are not the only source of demand for Ph.D. recipients and it is possible that weak demand from universities is offset by strong demand from industry (or that strong demand from industry attracts people to PhD programs). Turning to industry placements, we distinguish between the demand-pull "best of times" and the supply-push "worst of times" views in two ways. First, we examine Ph.D. recipients' earnings. If supply is pushing researchers out of academia and into industry, we should observe weak labor market outcomes, including relatively low earnings. By contrast, if demand is pulling researchers into industry then we ought to see strong labor market outcomes, including relatively high earnings. Second, we directly evaluate "match quality", constructing a dissertation-industry "relevance" measure by comparing a Ph.D. recipient's dissertation text to the text of the universe of patents from the United States Patent and Trademark Office (USPTO), linking patents to assignee firms, and using tax data on the universe of U.S. business establishments to aggregate the relevance measure to the industry level. This allows us to quantify, for each Ph.D. recipient, the relevance (or "match quality") of their specialized human capital (as reflected in their dissertation) to each industry.[7] Though for some fields (e.g. Social Sciences), our patent-industry relevance measure may understate actual relevance, it is likely that our measure is well-suited for most STEM fields we analyze, especially engineering and biology. Moreover, we validate our dissertation-industry relevance measure by showing that it predicts labor market outcomes for Ph.D. recipients. A one standard deviation increase in the relevance score between a Ph.D. recipient's dissertation and a 6-digit industry is associated with a 0.012 (15.66% relative to baseline) percentage point increase in the probability that

[5] One source of excess supply could be bad information (see Levitt (2010) and Ganguli et al. (2020)).

[6] Independent of the share of people entering postdocs, there may be private return to people taking postdocs, however, Kahn and Ginther (2017); Davis et al. (2022) find that returns (in the form of earnings) to a postdoc are quite low.

[7] This analysis of job matches is informative for two reasons. First, if Ph.D. recipients are taking jobs that do not require their specialized human capital, then it seems likely they could have obtained similar positions with less and/or different investments. Second, there is evidence that research-trained individuals view the ability to conduct research as a valuable job attribute (Roach and Sauermann, 2010; Stern, 2004).

they place in that industry (Table 2) and 3.4% higher earnings (Table 3).

We find that, of Ph.D. recipients placing in industry, those coming from Mathematics, Engineering, and Physics have the highest average initial earnings at $105k, $99k, and $82k. In contrast, those coming from Biology, Eco/Envr Sciences, and Comm/Info Sciences have the lowest average initial earnings at $58k, $59k, and $60k. With respect to match quality, we find that Ph.D. recipients from Engineering, Physics, and Chemistry have the highest average dissertation-industry relevance, while Agriculture, Comm/Info Sciences, and Social Sciences have the lowest. Overall, the field-level relationship between earnings and relevance is high (R^2=0.74 to 0.77), and a 1 standard deviation increase in the relevance score is associated with a $12,091 increase in earnings.

To summarize our results, we create a field-level composite ranking that assesses fields based on the quality of their academic and industry placements, and provides a more holistic understanding of which fields are, in a relative sense, experiencing the best and worst of times. Topping the list of placement outcomes are Physics, Engineering, and Mathematics, which appear to be experiencing the best of times. In contrast, our ranking suggests that Biology, Eco/Envr Sciences, and Agriculture appear to be experiencing the worst of times. Of course, the "best" and "worst" of times are relative rankings across fields of Ph.D. recipients and do not capture the fact that across fields Ph.D. recipients experience stronger labor market outcomes than most workers with less formal schooling. Moreover, we do not view this ranking as the final word on labor markets for Ph.D. recipients, and encourage other work to construct alternative metrics that offer additional perspective. However, using data on actual job placements – along with earnings and a new measure of match quality – is an informative leap toward determining which fields are experiencing relatively strong demand for the specialized human capital of their Ph.D. recipients.

2 Data

We combine a variety of detailed administrative data to conduct our analysis. We start with graduate students from the UMETRICS data and link them to their dissertations in ProQuest, allowing us to identify field, degree year, and dissertation titles and abstracts. We then link these Ph.D. recipients to their post-degree labor market outcomes using tax data (W2 records and the Business Register). Finally, we use USPTO patent data, along with a patent-firm bridge, to construct a new dissertation-industry "relevance" score. This section summarizes our data and methods. Additional details are provided in Appendix B.

UMETRICS Graduate Students UMETRICS is administrative data from universities. We use data on grant transactions at 24 major research universities (64 campuses), which collectively account for more than one-third of federally funded academic R&D.[8] Crucially, the data include information on the job titles of individuals receiving payments from these grants, allowing us to identify graduate students (Ikudo et al., 2019). These graduate students comprise our core sample, and we link them to a variety of additional information to determine their field and degree year, measure the relevance of their specialized human capital to different industries, and track their labor market outcomes (including job placements and earnings).

Dissertations, Fields, and Specialized Human Capital We link each UMETRICS graduate student to their dissertation in Proquest. As indicated, this allows us to determine the degree year as well as the field of each Ph.D. recipient, which is crucial because much of our analysis takes place at the field level. As discussed below, this match also enables us to use the text of each Ph.D. recipient's dissertation and compute its similarity to all patents in the USPTO. Linking these patents to assignee firms, allows us to measure the relevance of a Ph.D. recipient's specialized human capital (as reflected in their dissertation) to different industries, which sheds light on how well-matched each individual is to the firm or industry in which they are actually employed after receiving their Ph.D. We can only link patents granted between 2000 and 2015 to assignee firms (see below), so we restrict our sample to dissertations with a degree year between 2004-2015. This allows us to measure, for each dissertation, the similarity to patents issued in the four years before the dissertation year and the dissertation year. We impose this restriction to measure the relevance of a dissertation to an industry's recent patent portfolio while avoiding patents that are issued after a Ph.D. recipient may have been hired.[9]

Labor Market Outcomes To track labor market outcomes for the UMETRICS Ph.D. recipients, we use a Protected Identification Key (PIK) to link them to a variety of confidential data at the U.S. Census Bureau.[10] First, we link them to the universe of W2 tax

[8] We use the 2018 Q4 release of UMETRICS. For additional work using UMETRICS, see Chang et al. (2019), Ikudo et al. (2019), Buffington et al. (2016), Lane et al. (2013), Lane et al. (2015), Weinberg et al. (2014), Zolas et al. (2015), and Ross et al. (2022).

[9] One might prefer to use the application date to the award date for patents, but Census's bridge from patents to assignees is based on award dates (2000-2015), meaning that an analysis based on application dates would generate a sample with uncertain properties.

[10] Using identifying information such as name and birth date, the Census Bureau assigns individuals a PIK which is a unique, internal, person-level identifier. This is done through the Person Validation System (PVS), which is a probabilistic match. Once an individual is assigned to a PIK, we can link them to a variety of confidential Census data.

records (2005-2018), from which we identify the firms (EINs) at which they were employed as well as their earnings.[11] Next, we use the EINs to link each firm to the Business Register (BR), the universe of business establishments in the United States (DeSalvo et al., 2016). This allows us to determine the industry of the firm that employs each Ph.D. recipient in our sample.[12]

We use the information on the firm's industry as well as a list of EINs from the Integrated Postsecondary Education Data System (IPEDS) to identify universities and determine whether a Ph.D. recipient's initial post-degree job placement is in academia or industry.[13] Within the subset of Ph.D. recipients that place in academia, we use earnings information to impute whether the job was a faculty or postdoc position.[14] Thus, each Ph.D. recipient's initial placement is in one of three mutually exclusive job types: faculty members at a university, postdocs at a university, or industry (which includes the private sector, government, and NGOs). In addition to placements, we also track earnings during the first three years after a Ph.D. recipient earns their degree, subsetting our sample to Ph.D. recipients with positive earnings and non-missing industries in all years 1-3 following their degree year.

Patents and Dissertation-Industry Relevance As noted, a goal is to measure how closely related a Ph.D. recipient's dissertation is to the patenting portfolio of a firm or an industry. To do this, we identify the 1,000 USPTO patents that are most closely related to each dissertation on the basis of a similarity score computed using dissertation and patent text. We then use a patent-firm bridge from Census (Goldschlag and Perlman, 2017) to link patents to assignee firms and their accompanying industry codes, which allows us to aggregate our similarity score to the industry level and to measure the "relevance" of each dissertation to every industry. We detail the construction of this similarity measure in Section 3.

Final Sample Our final sample of individuals is a set of UMETRICS graduate students who 1) are assigned exactly one PIK, and thus can be linked to employment outcomes at

[11] Though, for convenience, we use the term "firm" throughout the paper, W2 records technically link individuals to a federal tax identification number (EIN) of their employer. Since a single firm can have multiple EINs, firms and EINs are not synonymous (see Figure 2 of (DeSalvo et al., 2016)). However, this distinction is not critical for our purposes because we only use the EIN as an intermediate variable through which we identify the industry in which a Ph.D. recipient works.

[12] See Appendix Section B.3 for details on how we assign EINs to industries using the Business Register (BR).

[13] IPEDS is maintained by the National Center for Education Statistics (NCES).

[14] We use the Survey of Doctoral Recipients (SDR) to determine average postdoc salaries by field. We then classify a job placement as a faculty position if it is at least 1 standard deviation above the mean and classify it as a postdoc otherwise. We note that this approach to identifying postdocs implies low initial earnings and high earnings growth for postdocs.

Census; 2) link 1-to-1 to a ProQuest dissertation, and thus their specialized human capital can be observed and compared to the patenting portfolio of firms and industries; and 3) have a dissertation that is relevant to at least 1 patent.[15] This gives us a final sample of 12,450 Ph.D. recipients. We focus on 11 STEM fields in our analysis because they receive the most federal funding and are thus the largest. Moreover, our dissertation-industry relevance measure is based on innovation intensity reflected in US patents, which is a useful measure given the relevance of patenting in STEM fields.

Summary Statistics Table 1 shows key summary statistics for this sample – both dissertation characteristics and employment outcomes – for each of the three initial placement sectors. Nearly half (5,800 or 46.6%) of Ph.D. recipients *initially* place in industry positions. Another 40.2% (5,000) initially place in postdoc positions and the remaining 13.3% (1,650) initially place in faculty positions.

In terms of the PhD recipient characteristics, those who *initially* place in faculty positions tend to be older when they receive their degree (Age at degree) and tend to belong to earlier cohorts (Degree year) than those who place in postdoc and industry positions, which implies that the share of people placing as faculty declines over time. The dissertations of Ph.D. recipients who place in industry are relevant to more patents and have higher cosine similarity scores (see Section 3) than the dissertations of those who place in academia. Thus, Ph.D. recipients who place in industry appear to have specialized human capital that is closer to the patenting frontier. However, these differences in dissertation-patent measures, across placement types, are not statistically significant.

In terms of employment outcomes, Ph.D. recipients who place in industry have the highest starting earnings at $83,540 followed by those who place in faculty positions at $80,780. Though the means are similar, the standard deviation is about twice as high for those who place in industry positions. Postdocs earn by far the least, starting at $36,130, although that is in part by construction. Three years after receiving their degree, Ph.D. recipients who *initially* placed in industry experience rapid earnings growth of 26.0% to $105,300. Those who initially placed in faculty positions experience much more modest growth of 3.8% to $83,810. By year three, the standard deviation of earnings for those in industry is nearly 4 times that of those in faculty positions. Though average earnings growth for those who initially placed as postdocs is strong at 40.4%, their low starting earnings imply that, three years later, the $50,720 they earn is still considerably less than those who initially placed in faculty or industry positions.

[15] If the similarity score for a dissertation-patent link is zero, the link is discarded and the dissertation will be linked to fewer than 1,000 patents. In the extreme, if all the patents have a zero similarity score to the dissertation, the dissertation will not be linked to any patents. See data appendix B.4 for additional details.

3 Measuring Dissertation-Industry Relevance

This section describes the construction of our new dissertation-industry "relevance" score, which we use to directly assess the match quality between a Ph.D. recipient and their industry of employment. We use the raw text of doctoral dissertations (title and abstract) in ProQuest to represent a Ph.D. recipient's specialized human capital. We use USPTO patents to represent the frontier of knowledge (innovation) valued and used by industries. We then use natural-language processing (See Kelly et al. (2021), Koffi (2021), and Biasi and Ma (2022) for other applications in economics) to measure the textual similarity between each dissertation and every patent. We then use the patent-firm bridge to link each patent to its assignee firm and accompanying industry, which enables us to aggregate the dissertation-patent scores to the dissertation-industry level. The rest of this section details the construction of our new relevance measure and validates it by demonstrating its ability to predict real labor market outcomes such as placement and earnings.

3.1 Dissertation-Patent Similarity

Term Frequency-Inverse Document Frequency Vectors To measure the similarity of patents to dissertations, we employ methods from natural language processing. Specifically, we use term frequency-inverse document frequency (TF-IDF), which identifies the frequency of a word in a document relative to its occurrence in an entire corpus. We then use cosine similarity measures to make comparisons across corpora (i.e., compare dissertations to patents). We start by computing the term frequency (TF) for each term t and document d (a set of terms) from corpus D (i.e., a set of documents).[16] The TF measures how often term t occurs in document d (relative to all terms in the document):

$$TF(t,d) = \frac{c(t \in d)}{c(d)}$$

where $c(t \in d)$ is the number of times term t appears in document d and $c(d)$ is the total number of terms in d.

Next, we use document frequency to measure how many documents in a corpus D contain the term t. We use inverse document frequency (IDF) to reweight these terms according to the frequency with which they occur in the entire corpus, which reflects their informativeness. Specifically, for a given term t, the IDF for corpus D is:

[16] In our case, documents are either dissertations or patents and the corpora are the set of all ProQuest dissertations or the set of all USPTO patents.

$$IDF(t, D) = log\left(\frac{|D|}{\sum_{d \in D} \mathbb{1}[c(t \in d) > 0]}\right)$$

where $|D|$ is the total number of documents in the corpus D and the denominator is the number of documents in D where the term t appears at least once. Intuitively, a document's use of a term that is widely used throughout the corpus is less informative than (and is thus down-weighted relative to) its use of a term that rarely appears in the corpus.[17]

Finally, for each term t and document d from corpus D, we take the product of $TF(t, d)$ and $IDF(t, D)$ to compute the term frequency-inverse document frequency:

$$TFIDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

This measure reflects how important a term t is to document d in corpus D. The TF-IDF increases as document d uses the term t more often as a fraction of the total number of terms used. If the document does not use a term at all, then $TF(t, d)$ and thus $TFIDF(t, d, D)$ are zero. Thus, the TF-IDF for term t in a document is higher for a term used intensively in that document and/or for using a term that is not used by many other documents in the corpus to which the document belongs.

Textual Similarities Between Dissertations and Patents Using the TF-IDF measures, we compute cosine similarity, a basic measure of linguistic proximity between documents, to construct a similarity score between dissertations and patents. Let A be the corpus of dissertations and B be the corpus of patents. Let $a \in A$ and $b \in B$ be a specific dissertation and patent (respectively) to be compared. Let $t_1, ..., t_n$ represent the n terms that appear in either a or b. We define the vectors V_a and V_b, which stack the TF-IDF measures for dissertation a and patent b, as

$$V_a = [TFIDF(t_1, a, A), TFIDF(t_2, a, A), ..., TFIDF(t_n, a, A)]$$

$$V_b = [TFIDF(t_1, b, B), TFIDF(t_2, b, B), ..., TFIDF(t_n, b, B)]$$

The elements of these vectors are indexed by the n terms that appear in either the dissertation or patent, so both vectors have length n. This conformability allows us to compute the cosine

[17] For example, the words "method", "system" and "process" are typically used to describe a patented invention but less often used in a dissertation. Thus, these words will be receive less weight (i.e. a smaller IDF) in the USPTO patent corpus than the ProQuest dissertation corpus.

similarity between dissertation a and patent b, which is defined as:

$$S(a,b) = \frac{V_a}{\|V_a\|} \cdot \frac{V_b}{\|V_b\|}$$

where $\|V_a\|$ and $\|V_b\|$ are the Euclidean norms of V_a and V_b. The measure ranges from 0 (no terms in common) to 1 (exactly the same terms), with higher values corresponding to higher degrees of similarity between the two documents.

Dissertation and Patent Data The data used to construct the cosine similarity scores are the universe of patent data from 2000 to 2015 (USPTO PatentsView) and all ProQuest doctoral dissertations from 2004 to 2015. For each ProQuest dissertation abstract, we compute its cosine similarity to each patent and retain the 1,000 patents with the highest similarity scores (though some dissertations match to fewer than 1,000 patents with positive cosine similarity scores). We drop patents with a zero cosine similarity.

Table A1 shows three examples of dissertation-patent pairs and the corresponding (qualitative) TF-IDF cosine similarity scores. We have chosen high, middle, and low cosine similarity score pairs for comparison purposes. The TF-IDF cosine similarity score is high between a dissertation on wireless communication and a patent on wireless communication. The similarity score is relatively high between a dissertation on one treatment of rat spinal cord injury and a patent on another treatment for nervous system injury. In contrast, the similarity score is low between the dissertation on wireless communication and the patent on the treatment of the nervous system.

3.2 Aggregating to the Industry-Level

Our goal is to create a full set of measures for how similar each Ph.D. recipient's dissertation is to the patenting portfolio of every industry for all pairwise combinations of dissertations and industries (i.e., dyads). To do so, we aggregate our dissertation-patent cosine scores to dissertation-industry scores, a process that we describe here and which is illustrated in Figure 1.

After attaching each graduate student to their dissertation and identifying the 1000 most relevant patents to each dissertation (see above), we subset to patents in the five years up to and including the dissertation year and connect each patent to its assignee firm. This creates a data set with unique observations at the dissertation-patent-firm level. We then collapse over the patents assigned to each firm to obtain the average cosine similarity score for each dissertation-firm pair. Thus, each dissertation-firm pair inherits a similarity score based on the average of several dissertation-patent similarity scores. We next use the Business Register

to assign a dominant industry (NAICS code) to each firm and collapse over the firm in each industry to obtain the average similarity score for each dissertation-industry. In the end, we have dyadic data with a score measuring the relevance of each Ph.D. recipient's dissertation to each industry. Note that the dyadic match is zero for a graduate student-industry pair if none of the 1000 patents that are most relevant to the graduate student's dissertation are assigned to that industry.

We compute these dissertation-industry similarity scores at the 2-digit, 4-digit, and 6-digit levels of industry aggregation. Higher levels of NAICS codes correspond to more detailed industry classifications, so as we move from 2- to 6-digit codes, we are able to observe the similarity between a Ph.D. recipient's dissertation and the patent portfolio of increasingly detailed industries.

3.3 Validating the Relevance Measure

This section uses our dissertation-industry dyad data to validate our new dissertation-industry relevance measure by examining the extent to which it predicts placements and earnings, two key labor market outcomes. At the outset, we caution against a causal interpretation of these estimates. With respect to placements, Ph.D. recipients likely write dissertations to competitively position themselves for their desired jobs. With respect to earnings, it is plausible that Ph.D. recipients who obtain jobs to which they are well-matched have unobserved characteristics associated with higher earnings. Yet our goal is to validate our measures rather than to estimate the causal effects of industry relevance.

We first examine whether the relevance measure predicts the specific industry (NAICS code) in which a Ph.D. recipient actually places. To do this, we focus on people who place in industry (i.e., non-academic placements) and regress an indicator for whether an individual places in a specific industry on the dissertation-industry relevance measure. Table 2 shows estimates for initial job placements (upper panel A) and placements three years after graduation (lower panel B). Each column is a different specification. All regressions include industry (NAICS) and individual fixed effects. Column 2 adds controls for the number of patents linked to each individual and column 3 adds degree year fixed effects and field fixed effects. Column 4 includes all controls. As noted, we construct our dissertation-industry similarity scores at the 2-digit, 4-digit, and 6-digit levels of industry aggregation, so we run regressions at each of these aggregation levels. Thus, each cell in Table 2 contains the coefficient on the relevance measure from a separate regression.

The probability that a Ph.D. recipient places in an industry is strongly increasing in the relevance of that industry to the graduate student's dissertation, both initially and three

years after receiving their degree. For initial placements, a one standard deviation increase in the relevance score is associated with an increase of 0.384, 0.072, and 0.035 percentage points in the probabilities of placing in a given 2-, 4-, and 6-digit industry. These increases are 9.11%, 24.69%, and 45.92% of the mean probability of placement.[18] For placements three years after graduation, a one standard deviation increase is associated with increases of 0.411, 0.078, and 0.031 for 2-, 4-, and 6-digit industries, which are 9.87%, 26.90%, and 41.28% of the means.

Controlling for the number of patents linked to a dissertation attenuates the estimates, especially as industries become more detailed, which is not surprising because it is effectively another measure of industry relevance. The estimates are also attenuated when we control for degree year and field. Nevertheless, even when all controls are included, we still see a strong relationship between industry relevance and placement. For initial placements, a one standard deviation increase in relevance is associated with an increase of 0.175, 0.027, and 0.012 percentage points in the probabilities of placing in a given 2-, 4-, and 6-digit industry, which are 4.20%, 9.41%, and 15.97% relative to the unconditional means. For placements three years after graduation, the associated increases are 0.216, 0.047, and 0.017 percentage points, which are 5.19%, 16.12%, and 22.12% of the means. The relationship between dissertation-industry relevance and placement tends to be stronger for placements three years after graduation than for initial placements. This implies that more relevant placements are more durable and/or that people move toward more relevant industries from those that are less relevant.

Next, we examine the relationship between the relevance score (for initial job placement) and the earnings of Ph.D. recipients. To do this, we select the industry in which the Ph.D. recipient initially places, which reduces our dyadic data structure to a single observation for each Ph.D. recipient and the industry in which they place. Then we regress log earnings on the industry relevance score for the Ph.D. recipient's initial job placement. Table 3 shows these estimates for initial earnings (upper panel A) and earnings three years after graduation (lower panel B). As with placements, each column is a different specification, all of which include industry (NAICS) fixed effects, degree year fixed effects, and field fixed effects. Column 2 adds controls for the number of patents linked to each individual and

[18]The regression coefficients for initial earnings in Column 1 of Table 2 are 0.0280, 0.0089, and 0.0069 for 2-, 4-, and 6-digit industries. The standard deviations of the relevance scores at these industry aggregation levels (unconditional on placement) are 0.1371, 0.0804, and 0.0500 (Column 3 of Table A2), so a one standard deviation increase in the relevance scores are associated with increases in placement probabilities of 0.0280*0.1371= 0.0038, 0.0089*0.0804=0.000716, and 0.0069*0.0500=0.000345. The unconditional probabilities of randomly placing in a given 2-, 4-, and 6-digit industry (Column 1 of Table A2) are 1/24=0.0417, 1/345=0.0029, and 1/1,331=0.0007513, so relative to the means, the increases in placement probabilities are 0.0038/0.0417=0.0911, 0.000716/0.0029=0.2469, and 0.000345/0.0007513=0.4592.

column 3 adds demographic controls, including gender, race, ethnicity, place of birth, and age at degree. We run regressions at the 2-, 4-, and 6-digit levels of industry aggregation.

A higher relevance score between a Ph.D. recipient's dissertation and the industry of their initial job is strongly related to earnings, both initially and three years after receiving their degree. Indeed, a one standard deviation increase in the relevance score is associated with 3.37%, 2.05%, and 2.04% higher initial earnings at the 2-, 4-, and 6-digit levels of industry aggregation.[19] Three years after graduation, a one standard deviation increase in the relevance score is associated with 4.57%, 2.74%, and 2.87% higher earnings.

Controlling for the number of patents linked to each dissertation tends to increase the estimates and controlling for demographic characteristics tends to attenuate the estimates. When all covariates are included, a one standard deviation increase in the relevance score at the 2-, 4-, and 6-digit levels of industry aggregation is associated with earnings increases of 3.39%, 3.32%, and 3.41% initially and increases of 4.78%, 4.36%, and 5.00% three years after graduation.

In sum, we view the strong relationship between our dissertation-industry relevance score and real economic outcomes – both placements and earnings – as promising validation of the measure. In the following sections, we use this measure to help assess the quality of industry job placements for Ph.D. recipients from different fields.

4 Analysis of STEM Career Outcomes

This section analyzes the placements of Ph.D. recipients from different fields to determine which are experiencing (in a relative sense) the best and worst of times. We start with faculty placements, from which we infer the demand for faculty in each field. We then move to postdoc placements, evaluating whether fields with a high fraction of postdocs experience relatively stronger wage growth (evidence of the need for extra training) or are simply facing weak demand for their specialized human capital. We then turn to industry placements, evaluating their quality using earnings and our new relevance score, which turn out to be closely related. After analyzing all three job types, we create a composite index ranking fields in terms of job placement quality. This index allows us to simply but more holistically summarize which fields are, in a relative sense, experiencing the best and worst of times.

[19] For initial earnings, the regression coefficients are 0.270, 0.151, and 0.149 for 2-, 4-, and 6-digit industries (Column 1 of Panel A in Table 3). The standard deviations of the relevance scores at these industry aggregation levels (conditional on placement) are 0.1249, 0.1355, and 0.1372 (Column 5 of Table A2), so one standard deviation increase in the relevance scores is associated with increases in earnings of 0.270*0.1249=0.033723, 0.151*0.1355=0.02046, and 0.149*0.1372=0.02044. The calculations are similar for earnings three years after degree receipt, using coefficients in column 3 of Panel A in Table 3.

4.1 Faculty Placements

We start by analyzing the share of Ph.D. recipients in each field who take jobs as faculty, as postdocs, and in industry. Figure 2 shows placement shares in the three position types, sorted by the share of faculty placements.[20]

A large share of faculty placements is an indicator of strong demand from universities for researchers in a field.

We see that the share of placements in each sector varies considerably across fields. Overall, there appears to be relatively strong demand for faculty in the Social Sciences (35.4%), Medicine (24.4%), and Comm/Info Sciences (24.3%). In contrast, demand for faculty is weak, relative to supply, in Chemistry (2.5%), Geosciences (7.7%), Biology (9.1%), and Engineering (10.2%). The remaining fields – Eco/Envr Sciences (14.1%), Mathematics (14.5%), Agriculture (15.7%), and Physics (16.1%) – are intermediate cases.

It is notable that the two largest fields – Biology and Engineering – are both at the lower end in terms of faculty placements. However, as we will see, these two fields are quite different in terms of postdoc and industry placements. These differences highlight the usefulness of our approach for simultaneously analyzing all three placement types and the value of a composite job placement ranking, which we construct in Section 4.4, to describe the overall, domestic labor market facing U.S. Ph.D. recipients in each field.

4.2 Postdoc Placements

We next turn to postdoc placements. Figure 2 shows that there is considerable variation, across fields, in the share of Ph.D. recipients taking postdoc positions. On the high end are Biology (58.2%), Eco/Envr Sciences (57.9%), and Geosciences (52.8%). On the low end are Engineering (22.5%), Physics (37.8%), and Social Sciences (41.4%). The remaining fields are Medicine (42%), Chemistry (45.3%), Agriculture (46.1%), Mathematics (47.3%), and Comm/Info Sciences (50.1%).

Unlike faculty placements, it is less clear *ex-ante* what a high fraction of postdoc placements means in terms of the relative supply and demand in a field. In other words, it is unclear whether a high fraction of postdoc placements in a field reflects the best or worst of times in that field. On the one hand, a high fraction of postdocs might indicate weak demand relative to supply. On the other hand, there may be strong demand in a field, but the content in the field may be so complex that it cannot be fully acquired during graduate

[20]The largest fields in our sample are Engineering (n=4,100) and Biology (n=2,800), which comprise 35.5% and 24.2% of the total. These are also the two largest fields in the Survey of Earned Doctorates, where 25% and 20% of all Ph.D. recipients are from Engineering and Biology. See Table A here: https://ncses.nsf.gov/surveys/earned-doctorates/2021

school and so further training as a postdoc is required. For instance, it may be that the complexity of biological processes means that postgraduate training is more important in Biology and Medicine than in other fields.

Human capital theory helps adjudicate whether a high fraction of postdoc placements reflects a need for additional human capital investments in a field or whether a high share of postdocs is more likely to indicate relatively weak demand. If postdocs are prevalent in Biology or other fields because additional human capital investment is particularly important in those fields, then earnings should grow relatively rapidly for people entering postdocs in those fields and be high later on. In contrast, if postdocs are prevalent in a field because the supply in a field is large relative to demand, earnings should continue to be low even as careers progress.[21]

Figure 3a shows that, for those whose initial placement is in a postdoc, earnings are lower in the third year after graduation in fields where the postdoc share among non-faculty placements is higher (the markers are sized according to the size of each field).[22] The relationship is remarkably strong with an R^2 of 0.83. To account for differences in initial earnings across fields, Figure 3b shows the relationship between earnings growth and the share of initial placements in postdocs. While postdocs exhibit strong earnings growth in all fields (in part due to how we identify them), as with the wage level, wage growth is also lower in fields with more postdoc placements. Again the relationship is quite strong, with an R^2 of 0.698 Notably, the two largest fields – Engineering and Biology – are extreme cases. Engineering has one of the smallest fractions of Ph.D. recipients taking postdoc positions and has the highest earnings/earnings growth among postdocs. Biology has one of the largest fractions of Ph.D. recipients taking postdoc positions and also has among the lowest earnings/earnings growth. More generally, fitted lines through the scatterplots suggest that a 10 percentage point increase in the share of graduates taking postdocs in a field among non-faculty placements is associated with year three earnings that are lower by \$3,660[23] and wage growth that is lower by 7.7 percentage points.

[21]Note that our goal is not to estimate the private return to postdocs – one would expect that individuals who take postdocs obtain human capital, including knowledge, connections, and qualifications, that increase their subsequent earnings (although existing evidence is surprisingly weak (Kahn and Ginther, 2017; Davis et al., 2022)). Rather, our goal is to estimate wage growth for the average person taking a postdoc in each field and compare average wage growth in fields where postdocs are prevalent to those where they are less common, to estimate the importance of supply and demand versus the importance of acquiring additional human capital beyond Ph.D. training.

[22]In some fields such as Biology, Ph.D. recipients often take multiple back-to-back postdoc positions that cumulatively last longer than three years. Ideally, we would track longer-term earnings growth, but our earnings data only extend through 2018. At the same time, postdocs are intended to be of modest duration and long postdocs may themselves be indicative of weak demand.

[23]This is a 7.2% reduction relative to the \$50,720 mean year 3 earnings for those who initially placed in postdoc positions.

Since both earnings (three years post-degree) and earnings growth are lower in fields with a higher fraction of postdocs, it seems likely that at least part of the reason for a large share of postdoc placements in some fields is high relative supply and not greater training requirements. Thus, overall, we interpret a large fraction of postdoc placements as evidence of weak demand (relative to supply) in a field and a high share of postdoc placements as an indicator that a field is experiencing the worst of times.

4.3 Industry Placements

Figure 2 shows widely varying shares, across fields, of Ph.D. recipients who take positions in industry. At the high end, 67% of Ph.D. recipients in Engineering place in industry jobs followed by Chemistry (52%) and Physics (46%). At 23%, the Social Sciences have the lowest share of industry placements, which suggests that the specialized human capital possessed by social scientists may be less relevant to industry. Comm/Info Sciences (25%) and Eco/Envr Sciences (28%) have the next lowest shares of Ph.D. recipients taking industry jobs.

In Section 4.1, we assumed that a high share of faculty placements represents strong (relative) demand for faculty in a field. In Section 4.2, we drew on insights from human capital theory to suggest that a high share of postdoc placements in a field likely indicates a weak relative demand for that field's specialized human capital. It is not immediately clear whether a high share of industry placements indicates a high relative demand in industry for the specialized human capital in a field or a high supply relative to other sources of demand. This is, of course, a fundamental insight from supply and demand analysis - that supply and demand shifts cannot be distinguished using equilibrium quantities without additional information. Traditionally, economists have turned to price data to distinguish supply from demand, and we do look at earnings data. In addition, we try to measure the quality of industry jobs using our new dissertation-industry relevance score (Section 3).

Table 4 shows, by field, the starting wage, 3-year wage, and dissertation-industry relevance score for Ph.D. recipients that place in industry. Those in Mathematics, Engineering, and Physics have the highest average starting earnings at $105,400, $98,520, and $81,800. At the low end are Ph.D. recipients in Biology, Eco/Envr Sciences, and Comm/Info Sciences, with average starting earnings of $58,170, $59,200, and $60,240. Thus, engineers (the highest-paid field) start out earning 81.1% more than biologists (the lowest-paid field).

Three years after receiving their degree, Ph.D. recipients in Mathematics, Engineering, and Physics remain the highest paid, with average earnings of $146,700, $124,100, and $101,100. Those in Biology and Eco/Envr Sciences remain at the bottom, averaging $78,040 and $66,480, respectively. However, Agriculture – with average year-3 earnings of $73,630 –

has displaced Comm/Info Sciences near the bottom of the earnings distribution. Unsurprisingly, the relationship between initial and 3-year earnings is very strong, with a correlation of 0.747. Taken together, these results suggest that mathematicians, engineers, and physical scientists may be drawn into industry by demand, while biologists and ecological/environmental scientists may be pushed into industry by supply.

Of course, earnings are only one job attribute valued by Ph.D. recipients. There is also evidence that they value jobs in which they make use of their specialized human capital (Roach and Sauermann, 2010; Stern, 2004). Our new dissertation-industry relevance measure allows us to directly assess how closely matched a Ph.D. recipient's specialized human capital (measured from their dissertation) is to the patenting portfolio of the industry in which they are employed.

Column 3 of Table 4 shows that Ph.D. recipients in Engineering, Physics, Chemistry, and Mathematics have the highest average relevance scores at 0.2813, 0.2272, 0.2111, and 0.2099. Thus, most of the fields at the top of the industry earnings distribution are also the fields where people are most closely matched to their industry. At the bottom of the relevance score distribution are Ph.D. recipients in Agriculture, Social Sciences, and Comm/Info Sciences, which have average scores of 0.1540, 0.1663, and 0.1671. These fields are also lower in the industry earnings distribution.

Figure 4 plots the relationship between initial and year 3 earnings measures and the dissertation-industry relevance score, confirming the patterns in Table 3. The top earning fields – Engineering, Mathematics, and Physics are among the top four fields (along with Chemistry) in terms of the relevance score. The lowest earning fields – Eco/Envr Sciences and Biology – are clustered at the lower end of the relevance score range. The regression line in Figure 4 indicates that a 1 standard deviation increase in the relevance score is associated with a \$12,092 increase in initial earnings for those who place in industry, which is a 14.5% increase from the mean of \$83,540. The relationships are quite strong with an R^2 of .77 and .74 for initial and year 3 earnings respectively.

Overall, earnings and our dissertation-industry relevance score paint a consistent picture. They both indicate that in terms of the quality of industry placements, Engineering, Mathematics, and Physics are experiencing the best of times while Biology, Eco/Envr Sciences, and Agriculture are experiencing the worst of times. While patenting may not capture relevance equally for all fields (e.g., the social sciences), the strong relationship between relevance and earnings suggests that our relevance measure reasonably reflects the value of Ph.D. recipients to industry.

Though the measures used to evaluate the quality of industry job placements – earnings and the relevance score – are consistent with each other, they may fail to capture certain job

attributes valued by Ph.D. recipients. Indeed, it is possible that Ph.D. recipients who place in high-paying jobs with low relevance scores are nevertheless doing work that depends on their specialized human capital. For instance, physicists are known to get jobs on Wall Street using their mathematics and programming knowledge. To probe whether our earnings and relevance score measures are failing to capture important job attributes, we examine in Table 5 the specific narrow (6-digit) industries into which Ph.D. recipients place. We categorize these placements by match quality and pay level: (1) well-matched high-paid industries, (2) well-matched low-paid industries, (3) poorly-matched high-paid industries, and (4) poorly-matched low-paid industries.[24]

Ph.D. recipients who place in highly paid industries, whether well-matched or not, seem likely to be using their specialized human capital. Four of the top five industries for both categories are Research and Development in the Physical, Engineering, and Life Sciences (except Biotechnology); Custom Computer Programming Services; Engineering Services; and Internet Publishing and Broadcasting and Web Search Portals. By contrast, Ph.D. recipients that place in low-pay and poorly matched industries seem less likely to be relying heavily on their specialized human capital. Indeed, the top three narrow industries – Temporary Help Services, Facilities Support Services, and Employee Leasing Services – are not clearly connected to the specialized human capital Ph.D. recipients spent years accumulating in graduate school. The fourth-ranked narrow industry – Elementary and Secondary Schools – likely has a high social return and may draw on the specialized human capital of Ph.D. recipients, but it is likely these jobs could have been obtained with much less human capital investment. The number of Ph.D. recipients that place in well-matched, but low-paid narrow industries is quite low, suggesting again that earnings and the relevance score are related. Overall, our analysis of specific narrow industries does not suggest that our earnings and relevance score measures are missing important job characteristics valued by Ph.D. recipients.

4.4 Composite Ranking

This section combines our results on faculty placements (Section 4.1), postdoc placements (4.2), and industry placements (Section 4.3) into a simple, transparent composite ranking that summarizes which fields are experiencing the best and worst of times. We first create three separate rankings of the 11 fields based on: 1) the share of Ph.D. recipients that

[24]First, we split the subset of Ph.D. recipients who place in industry into two groups based on whether they are above or below the mean dissertation-industry relevance score for their field. Second, we classify all industries into two categories – high-paid and low-paid – based on the mean W2 wage of all employees in the U.S. with a bachelor's degree and above (i.e., not only the Ph.D. recipients in our sample) in that industry. Education levels are drawn from Census's Individual Characteristics File (ICF) and earnings are drawn from W2 records.

place in faculty positions, 2) the share of Ph.D. recipients that place in postdocs (reverse coded) conditional on not placing in a faculty position, 3) the average initial earnings of Ph.D. recipients that place in industry positions, and 4) the average dissertation-industry relevance score for Ph.D. recipients that place in industry. The first measure captures the quantity of faculty placements. The second captures the probability of being in a postdoc, which enters negatively given our estimates that postdocs are associated with high supply relative to demand rather than exceptionally large training needs. The last two measures capture the quality of industry placements. These individual rankings are displayed in the first three columns of Table 6.

We then take weighted averages of the four individual rankings to create a set of composite rankings to capture the overall quality of job placements for the Ph.D. recipients in each field. First, we equally weight each individual ranking (assigning a weight of .25 to each). Second, because we have 2 measures of the quality of industry placements, to avoid over-weighting outcomes for those who place in industry, we give weight of 1/3 to faculty placements, 1/3 to postdoc placements, and weights of 1/6 to each of the industry placement variables. Lastly, we rank the sectors in accordance to the share of people entering each sector in the entire population, giving half of the industry weight to each of the two industry variables (i.e., so that the total weight they receive equals the share of people placing in industry. These composite rankings are displayed in Table 6.

Overall, the composite rankings are consistent with our separate analyses. In particular, Ph.D. recipients in Physics, Engineering, and Mathematics are experiencing "the best of times", while those in Biology, Eco/Envr Sciences, and Agriculture are experiencing "the worst of times".

Notably, there appear to be multiple paths a field can take to enjoy high-quality overall job placements. Some fields – such as Physics – benefit from both a relatively high fraction of faculty placements as well as relatively high earnings and relevance scores, suggesting strong demand from both universities and industry. Alternatively, fields like Engineering place a relatively low share of Ph.D. recipients in faculty positions (suggesting relatively weak demand from universities), but have strong demand from industry as suggested by the relatively high industry earnings and relevance scores, providing a strong overall labor market for those in Engineering. Thus, for some fields, demand from industry can make up for a shortfall in the university of demand (alternatively, demand from industry may drive enrollments).

As noted, our composite rankings have several limitations. First, they are relative rankings across fields of Ph.D. recipients. They do not capture the fact that, across all fields, these individuals experience much stronger labor market outcomes than typical workers with

less formal schooling. Second, all rankings that summarize a variety of information into a single score will inevitably lack some nuance. Alternative data on Ph.D. placements and alternative metrics measuring job characteristics may provide additional insights. Lastly, the weights we have applied are largely arbitrary. In the end, we do not view these rankings as the final word on cross-field labor markets for Ph.D. recipients. However, actual job placements, earnings, and relevance scores are critical measures of job quality that co-move in intuitive ways and provide an informative path toward understanding which fields are experiencing relatively strong demand for the specialized human capital of their Ph.D. recipients.

5 Conclusion

Returning to our initial question – is it the best or worst of times for highly-trained researchers? Our analysis suggests that this question is fundamentally the wrong question to ask. As the beginning of Dickens' novel suggests, it is simultaneously a good time for some fields and a bad time for others. Indeed the two largest fields – Engineering and Biology – illustrate these poles. While only about 10% of graduates in either of these fields go into faculty positions straight out of graduate programs, far more biologists (58%) take postdoc positions while far more engineers go into industry (67%). While the large share of biologists taking postdocs may indicate a need for further training given the complexity of biological phenomena (and the same might be argued for other fields where postdocs are prevalent), our results suggest that an imbalance between supply and demand is more likely to be at play – if human capital accumulation were the driving factor one would expect to see rapid earnings growth in Biology and other fields where postdocs are common. By contrast, we show that year-3 earnings and earnings growth are both low in fields where postdocs are common. Moreover, a novel measure of the relevance to industry of the Ph.D. recipient's specialized knowledge indicates that it is in the fields where knowledge is least relevant to industry that postdocs are most common.

While it is beyond the scope of our analysis, it is perhaps useful to ask why Ph.D. recipients in some fields might be experiencing the worst of times. We conjecture that perverse incentives built in the research system (Teitelbaum, 2008; Stephan, 2012a) combined with a lack of information about long-term career prospects is likely a factor (Levitt, 2010; Ganguli et al., 2020) that may have been exacerbated by the expansion of biomedical Ph.D. programs during the doubling of the NIH budget (Zerhouni, 2006; Blume-Kohout and Clack, 2013). Indeed, informed observers are arguing for and promoting efforts to increase transparency in the life sciences (and also in the humanities) in part because of the perceived lack of

opportunities (Blank et al., 2017; Benderly, 2018).[25]

While we have pointed to fields where labor markets for STEM doctorates are weaker than others, it is important to bear in mind that, regardless of field, labor market outcomes for STEM doctorates are considerably stronger than most workers. Our estimates should be viewed as helping to fine-tune supply and demand rather than suggesting that STEM research education does not have private and, presumably, even higher social value.

[25]See https://cgsnet.org/resources/for-current-prospective-graduate-students/ and https://nglscoalition.org/.

References

Alberts, Bruce, Marc W Kirschner, Shirley Tilghman, and Harold Varmus (2015), "Opinion: Addressing systemic problems in the biomedical research enterprise." *Proceedings of the National Academy of Sciences*, 112, 1912–1913. 2

Benderly, BL (2018), "A trend toward transparency for ph. d. career outcomes? science." 22

Biasi, Barbara and Song Ma (2022), "The education-innovation gap." Technical report, National Bureau of Economic Research. 9

Blank, Rebecca, Ronald J. Daniels, Gary Gilliland, Amy Gutmann, Samuel Hawgood, Freeman A. Hrabowski, Martha E. Pollack, Vincent Price, L. Rafael Reif, and Mark S. Schlissel (2017), "A new data effort to inform career choices in biomedicine." *Science*, 358, 1388–1389, URL https://www.science.org/doi/abs/10.1126/science.aar4638. 22

Blume-Kohout, Margaret E and John W Clack (2013), "Are graduate students rational? evidence from the market for biomedical scientists." *PLoS One*, 8, e82759. 21

Buffington, Catherine, Benjamin Cerf, Christina Jones, and Bruce A Weinberg (2016), "Stem training and early career outcomes of female and male graduate students: Evidence from umetrics data linked to the 2010 census." *American Economic Review*, 106, 333–38. 6, 40

Chang, Wan-Ying, Wei Cheng, Julia Lane, and Bruce Weinberg (2019), "Federal funding of doctoral recipients: What can be learned from linked data." *Research Policy*, 48, 1487–1492. 6, 40

Cyranoski, David, Natasha Gilbert, Heidi Ledford, Anjali Nayar, and Mohammed Yahia (2011), "Education: the phd factory." 2

Davis, James C, Holden A Diethorn, Gerald R Marschke, and Andrew J Wang (2022), "A machine learning approach to identifying postdocs in lehd data." 4, 16

DeSalvo, Bethany, Frank Limehouse, and Shawn D Klimek (2016), "Documenting the business register and related economic business data." *US Census Bureau Center for Economic Studies Paper No. CES-WP-16-17*. 7, 41

Ganguli, Ina, Patrick Gaulé, and Danijela Vuletić Čugalj (2020), "Biased beliefs and entry into scientific careers." 4, 21

Goldschlag, Nathan and Elisabeth Perlman (2017), "Business Dynamic Statistics of Innovative Firms." Working Papers 17-72, Center for Economic Studies, U.S. Census Bureau, URL https://ideas.repec.org/p/cen/wpaper/17-72.html. 7, 43

Gould, Julie (2015), "How to build a better phd." *Nature News*, 528, 22. 2

Heggeness, Misty L, Kearney TW Gunsalus, Jose Pacas, and Gary McDowell (2017), "The new face of us science." *Nature*, 541, 21–23. 2

Ikudo, Akina, Julia I Lane, Joseph Staudt, and Bruce A Weinberg (2019), "Occupational classifications: A machine learning approach." *Journal of Economic and Social Measurement*, 44, 57–87. 6, 40

Jones, Benjamin F (2009), "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" *The Review of Economic Studies*, 76, 283–317. 2

Kahn, Shulamit and Donna K Ginther (2017), "The impact of postdoctoral training on early careers in biomedicine." *Nature biotechnology*, 35, 90–94. 4, 16

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2021), "Measuring technological innovation over the long run." *American Economic Review: Insights*, 3, 303–320. 9

Koffi, Marlène (2021), "Innovative ideas and gender inequality." Technical report, Working Paper Series. 9

Lane, Julia, John King, and Lou Schwarz (2013), "The creation of new administrative data." *Available at SSRN 2213916*. 6, 40

Lane, Julia I, Jason Owen-Smith, Rebecca F Rosen, and Bruce A Weinberg (2015), "New linked data on research investments: Scientific workforce, productivity, and public value." *Research policy*, 44, 1659–1671. 6, 40

Levitt, David G (2010), "Careers of an elite cohort of us basic life science postdoctoral fellows and the influence of their mentor's citation record." *BMC medical education*, 10, 1–7. 4, 21

Mathur, Ambika, Frederick J Meyers, Roger Chalkley, Theresa C O'Brien, and Cynthia N Fuhrmann (2015), "Transforming training to reflect the workforce." 2

Meyers, Frederick J, Ambika Mathur, Cynthia N Fuhrmann, Theresa C O'Brien, Inge Wefes, Patricia A Labosky, S Duncan D'Anne, Avery August, Andrew Feig, Kathleen L Gould, et al. (2016), "The origin and implementation of the broadening experiences in scientific training programs: an nih common fund initiative." *The FASEB Journal*, 30, 507. 2

Powell, Kendall (2015), "The future of the postdoc." *Nature*, 520, 144–148. 2

Roach, Michael and Henry Sauermann (2010), "A taste for science? phd scientists' academic orientation and self-selection into research careers in industry." *Research policy*, 39, 422–434. 4, 18

Romer, Paul M (1990), "Endogenous technological change." *Journal of political Economy*, 98, S71–S102. 2

Ross, Matthew B, Britta M Glennon, Raviv Murciano-Goroff, Enrico Berkes, Bruce A Weinberg, and Julia I Lane (2022), "Women are credited less in science than men." *Nature*, 608, 135—-145. 6

Stephan, Paula (2012a), "Perverse incentives." *Nature*, 484, 29–31. 21

Stephan, Paula E (2012b), *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA. 2

Stern, Scott (2004), "Do scientists pay to be scientists?" *Management science*, 50, 835–853. 4, 18

Teitelbaum, Michael S. (2008), "Structural disequilibria in biomedical research." *Science*, 321, 644–645, URL https://www.science.org/doi/abs/10.1126/science.1160272. 21

Wagner, Deborah, Mary Lane, et al. (2014), "The person identification validation system (pvs): applying the center for administrative records research and applications'(carra) record linkage software." Technical report, Center for Economic Studies, US Census Bureau. 40

Weinberg, Bruce A, Jason Owen-Smith, Rebecca F Rosen, Lou Schwarz, Barbara McFadden Allen, Roy E Weiss, and Julia Lane (2014), "Science funding and short-term economic activity." *Science*, 344, 41–43. 6, 40

Zerhouni, Elias A. (2006), "Nih in the post-doubling era: Realities and strategies." *Science*, 314, 1088–1090, URL https://www.science.org/doi/abs/10.1126/science.1136931. 21

Zolas, Nikolas, Nathan Goldschlag, Ron Jarmin, Paula Stephan, Jason Owen-Smith, Rebecca F Rosen, Barbara McFadden Allen, Bruce A Weinberg, and Julia I Lane (2015), "Wrapping it up in a person: Examining employment and earnings outcomes for ph. d. recipients." *Science*, 350, 1367–1371. 6, 40

Table 1: PIK/Dissertation Summary Statistics by Initial Placement

	Initial Placement					
	Faculty		Post-Doc		Industry	
	Mean	SD	Mean	SD	Mean	SD
PhD Recipient Characteristics						
Degree year	2010	3.15	2011	3.06	2011	3.093
Age at degree	33.26	6.369	30.88	4.658	29.87	4.059
Number of patents matched	188.9	61.93	191.2	55.57	202.1	61.19
Mean cosine similarity	0.214	0.112	0.192	0.108	0.234	0.119
Median cosine similarity	0.201	0.112	0.180	0.107	0.221	0.119
Max cosine similarity	0.400	0.152	0.374	0.158	0.421	0.157
Employment Outcomes						
Year 1 earnings	80,780	25,810	36,130	14,410	83,540	52,140
Year 3 earnings	83,810	32,640	50,720	28,670	105,300	129,500
ln(Year 1 earnings)	11.26	0.282	10.32	0.839	11.12	0.790
ln(Year 3 earnings)	11.25	0.482	10.62	0.839	11.32	0.787
Observations (rounded)	1,650		5,000		5,800	

Notes – This table displays the means and standard deviations of PhD recipient characteristics and employment outcomes for our analysis sample of Ph.D. recipients. Postdocs are identified as those whose first-year earnings were less than one standard deviation above the mean earnings in their field (in a given degree year). Year 3 earnings by sector is the mean earnings of Ph.D. recipients three years after receiving their degree based on the sector of their *initial* placement.

Table 2: Industry Relevance and Placement Outcomes

	(1)	(2)	(3)	(4)
Panel A: Placement in Year 1				
2-Digit NAICS				
Relevance Score	.0280***	.0245***	.0142***	.0128***
	(.0024)	(.0036)	(.0023)	(.0036)
4-Digit NAICS				
Relevance Score	.0089***	.0024**	.0061***	.0034***
	(.0005)	(.0009)	(.0005)	(.0009)
6-Digit NAICS				
Relevance Score	.0069***	.0004	.0046***	.0024***
	(.0003)	(.0006)	(.0003)	(.0006)
Panel B: Placement in Year 3				
2-Digit NAICS				
Relevance Score	.0300***	.0271***	.0167***	.0158***
	(.0024)	(.0037)	(.0023)	(.0038)
4-Digit NAICS				
Relevance Score	.0097***	.0046***	.0070***	.0058***
	(.0005)	(.0009)	(.0005)	(.0009)
6-Digit NAICS				
Relevance Score	.0062***	.0013*	.0048***	.0033***
	(.0003)	(.0006)	(.0003)	(.0006)
Industry FE	X	X	X	X
Individual FE	X	X	X	X
# of Patents		X		X
Degree year FE			X	X
Field FE			X	X
Obs. 2-Digit NAICS	139,200	139,200	139,200	139,200
Obs. 4-Digit NAICS	2,001,000	2,001,000	2,001,000	2,001,000
Obs. 6-Digit NAICS	7,719,800	7,719,800	7,719,800	7,719,800

Notes – Each cell reports the coefficient from a regression of an indicator variable for placement in an industry on the relevance of the dissertation to that industry. The unit of observation is a person-industry dyad and every person has an observation for each industry (i.e., the dyadic structure is "balanced"). The placement outcome equals 1 if the Ph.D. recipient placed in the industry and equals 0 otherwise. The sample is limited to people who place in industry, so each Ph.D. recipient places in one, and only one, industry. The relevance score is the average cosine similarity between the text of the recipient's dissertation and the text of patents assigned to the firms in that industry. The upper panel reports results for the placement in the first year after degree receipt and the lower panel reports results for the placement in the 3rd year after degree receipt. There are a total of 12,450 Ph.D. recipients/dissertations, 24 2-digit industry codes, 345 4-digit industry codes, and 1331 6-digit industry codes.

Table 3: Industry Relevance and Earnings

	(1)	(2)	(3)
Panel A: Earnings in Year 1			
2-Digit NAICS			
Relevance Score	.270***	.307***	.272***
	(.060)	(.066)	(.065)
4-Digit NAICS			
Relevance Score	.151*	.269***	.245**
	(.063)	(.076)	(.075)
6-Digit NAICS			
Relevance Score	.149*	.265**	.249**
	(.066)	(.082)	(.081)
Panel B: Earnings in Year 3			
2-Digit NAICS			
Relevance Score	.366***	.420***	.383***
	(.061)	(.067)	(.065)
4-Digit NAICS			
Relevance Score	.202**	.335***	.322***
	(.065)	(.078)	(.078)
6-Digit NAICS			
Relevance Score	.209**	.375***	.365***
	(.071)	(.090)	(.089)
Industry FE	X	X	X
Degree year FE	X	X	X
Field FE	X	X	X
# of Patents		X	X
Demographic Covariates			X
Obs.	5,800	5,800	5,800

Notes – Each cell reports the coefficient from a regression of log earnings on the dissertation-industry relevance score. The unit of observation is a person. The relevance score is the average cosine similarity between the text of the Ph.D recipient's dissertation and the text of patents assigned to the firms in the industry in which they initially placed. Earnings in year 3 is the earnings of Ph.D. recipients three years after receiving their degree (related to their year-1 placement). The sample is restricted to people whose initial placement is in industry.

Table 4: Industry Earnings and Relevance

Field	Year 1 Earnings	Year 3 Earnings	Relevance
Agriculture	62,920	73,630	0.154
Biology	58,170	78,040	0.1819
Chemistry	74,530	89,630	0.2111
Comm/Info Sciences	60,240	82,770	0.1671
Eco/Envr Sciences	59,200	66,480	0.1721
Engineering	98,520	124,100	0.2813
Geosciences	79,470	88,880	0.1826
Medicine	61,440	78,050	0.1752
Mathematics	105,400	146,700	0.2099
Physics	81,800	101,100	0.2272
Social Sciences	73,090	82,840	0.1663
Observations	5,800	5,800	5,800

Notes: Columns 1 and 2 show the earnings, by field, one year and three years since degree for those who initially placed in industry. Column 3 shows the mean relevance scores by field. The relevance score is the average cosine similarity between the text of the recipient's dissertation and the text of patents assigned to the firms (in the industry in which they placed).

Table 5: Common Industry Placements by Pay and Relevance

NAICS	Industry	PIK Count
Well-matched, high paid		
541712	Research and Development in the Physical, Engineering, and Life Sciences (except Biotechnology)	500
541511	Custom Computer Programming Services	300
541330	Engineering Services	200
519130	Internet Publishing and Broadcasting and Web Search Portals	100
325412	Pharmaceutical Preparation Manufacturing	100
Other	Other	1,300
Well-matched, low paid		
622110	General Medical and Surgical Hospitals	40
336111	Automobile Manufacturing	20
Other	Other	80
Bad-matched, high paid		
541712	Research and Development in the Physical, Engineering, and Life Sciences (except Biotechnology)	200
541511	Custom Computer Programming Services	100
541330	Engineering Services	80
621111	Offices of Physicians (except Mental Health Specialists)	70
519130	Internet Publishing and Broadcasting and Web Search Portals	60
Other	Other	1,700
Bad-matched, low paid		
561320	Temporary Help Services	80
561210	Facilities Support Services	80
561330	Employee Leasing Services	60
611110	Elementary and Secondary Schools	40
541940	Veterinary Services	30
Other	Other	550

Notes – This table shows the top 5 industries (6-digit NAICS level) in which the most PhD recipients placed for each pair of match quality and pay level. We construct these 4 categories by first splitting the subset of Ph.D. recipients who place in industry into two groups based on whether they are above or below the mean industry relevance score for their field; and then classifying all industries into two categories – high-paid and low-paid – based on the mean W2 wage of all employees in the U.S. with a bachelor's degree and above (i.e., not only the Ph.D. recipients in our sample) in that industry. The "Other" category collapses other industries that contain too few people, for the sake of disclosure restrictions.

Table 6: Index of Placement Outcomes

Field	%Faculty Rank	%Postdoc Rank (reverse)	Industry Earnings Rank	Industry Relevance Rank	Composite Rank (equal weight)	Composite Rank (1/3,1/3,1/6,1/6)	Composite Rank (placement weight)
Engineering	8	1	2	1	2	2	1
Physics	4	2	3	2	1	1	2
Mathematics	6	7	1	4	3	5	3
Social Sciences	1	3	6	10	4	3	4
Chemistry	11	5	5	3	6	6	5
Medicine	2	4	8	7	5	4	6
GeoSciences	10	9	4	5	7	9	7
Agriculture	5	6	7	11	8	6	8
Comm/Info Sciences	3	8	9	9	8	6	9
Eco/Envr Sciences	7	10	10	8	10	10	10
Biology	9	11	11	6	11	11	11

Notes – The first columns of the table show the four criteria we used to build the indices: 1) the rank of the share of faculty placements, 2) the rank of the share of postdoc placement, 3) the rank of industry earnings, and 4) the rank of industry relevance. The last three columns show three types of indices based on these four criteria. First, we equally weight each individual ranking (assigning a weight of .25 to each). Second, because we have 2 measures of the quality of industry placements, to avoid over-weighting outcomes for those who place in industry, we give weight of 1/3 to faculty placements, 1/3 to postdoc placements, and weights of 1/6 to each of the industry placement variables. Lastly, we rank the sectors according to the share of people entering each sector in the entire population, giving half weight to each of the two industry variables so that the total weight they receive equals the share of people placed in industry (i.e., the faculty rank weight = 1650/12450, the postdoc rank weight = 5000/12450, the industry earnings rank weight = 2900/12450, and the industry relevance rank weight = 2900/12450).

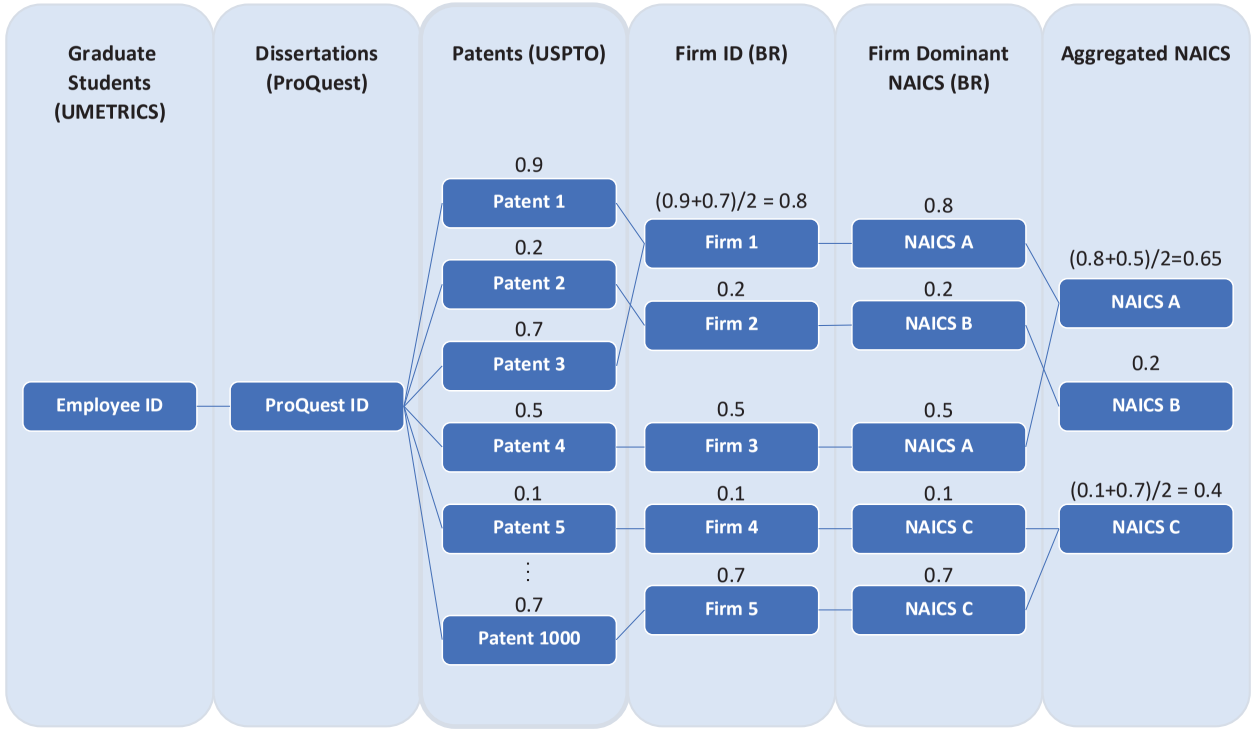


Figure 1: Illustration of Data Linkages

Notes – This flowchart illustrates a stylized example of how we construct our measure of the relevance of each industry to each graduate student's dissertation. We first attach each graduate student to their dissertation and identify the 1,000 most relevant patents to each dissertation based on the cosine similarity. In this example, the dissertation has a cosine similarity of 0.9 with Patent 1, 0.2 with Patent 2, 0.7 with Patent 3, and so on. Next, we connect each patent to its assignee firm and collapse over the patents assigned to each firm to obtain the average cosine similarity score for each dissertation-firm pair. Thus, each dissertation-firm pair inherits a similarity score based on the average of one or more dissertation-patent similarity scores. In this example, Patent 1 and Patent 3 are both assigned to Firm 1, and so the average similarity score between the dissertation and Firm 1 is (0.9+0.7)/2=0.8. Since Patent 2 is the only patent assigned to Firm 2, the average similarity score between the dissertation and Firm 2 is 0.2. The average similarity scores between the dissertation and other assignee firms are computed similarly. We next use the Business Register to assign a dominant industry (NAICS code) to each firm and collapse over the firm in each industry to obtain the average similarity score for each dissertation-industry pair. In this example, Firm 1 and Firm 3, which have similarity scores of 0.8 and 0.5 with the dissertation, are both assigned to NAICS A. Thus, the similarity between the dissertation and NAICS A is (0.8+0.5)/2=0.65. Since Firm 2 has a similarity score with the dissertation of 0.2 and is the only firm in NAICS B, the similarity between the dissertation and NAICS B is 0.2. The similarity score between the dissertation and other NAICS codes (industries) is computed similarly. In the end, we have balanced dyadic data with a score measuring the relevance of each Ph.D. recipient's dissertation to each industry (Ph.D. recipients whose dissertation does not link to any patents assigned to a firm in a given industry have similarity score of zero with that industry).

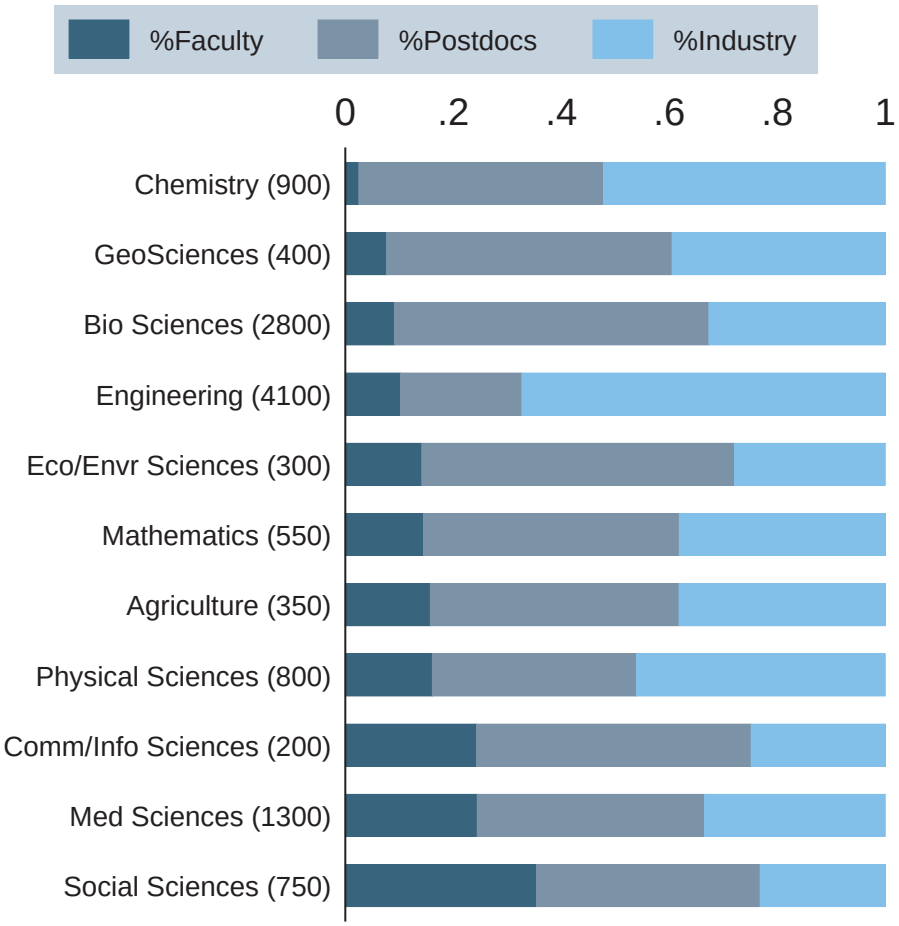
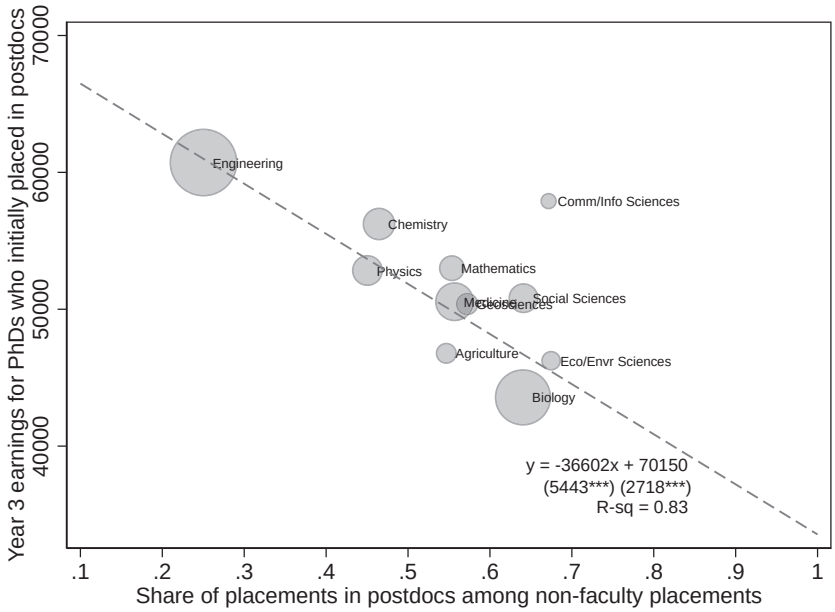
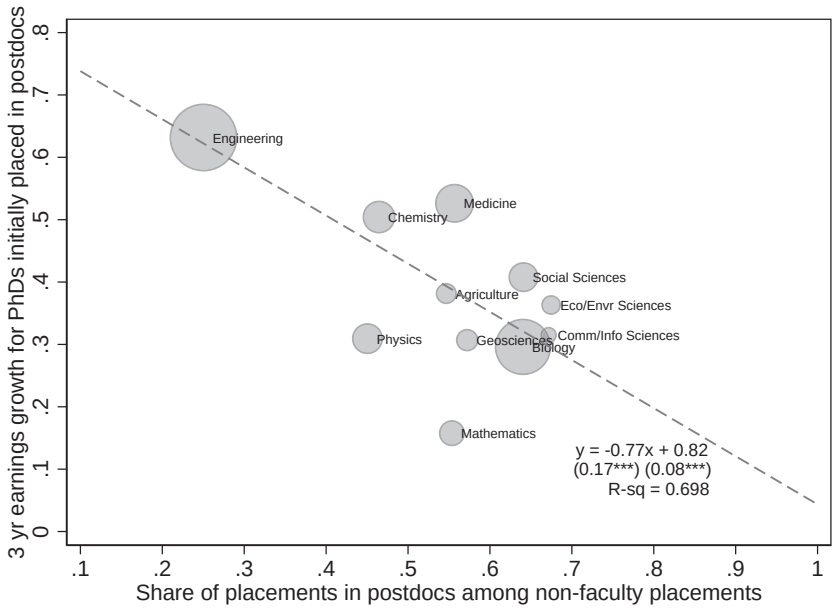


Figure 2: Share of Initial Placements in Each Sector

Notes – Fields are sorted by the share of faculty placements. The numbers in parentheses show the (rounded) numbers of observations.



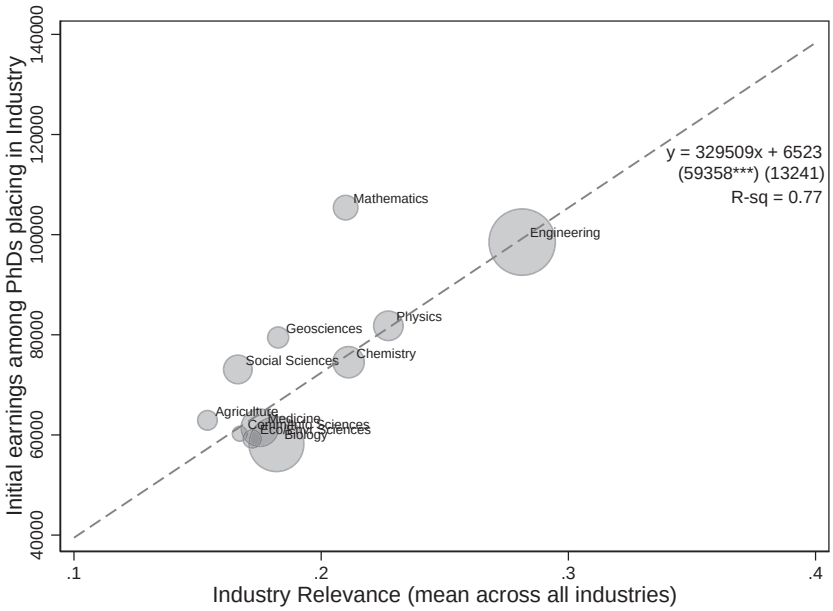
(a) People who take postdocs in fields with lots of postdocs earn less even after 3 Years



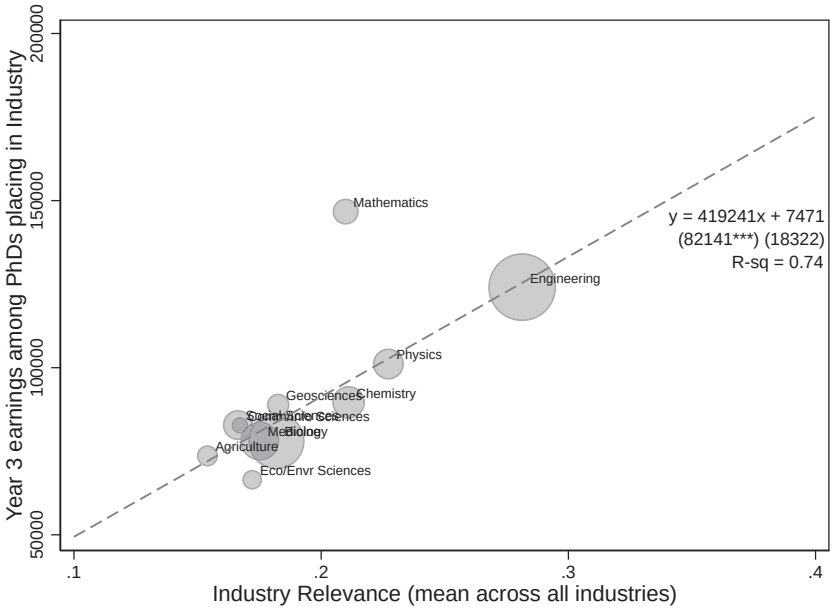
(b) People who take postdocs in fields with lots of postdocs have lower wage growth

Figure 3: Postdoc Placement Rate and Earnings

Notes – The figures plot earnings in year 3 (Panel (a)) and earnings growth between years 1 and 3 (Panel (b)) for people who place in postdocs initially by field plotted against the share of placements in postdocs among non-faculty placements. The dashed lines are the fitted lines from regressions weighted by field size. Standard errors and significance levels are shown in parentheses. *** $p<0.001$.



(a) Initial earnings



(b) Year 3 earnings

Figure 4: Industry relevance is strongly related to earnings

Notes – The figures plot earnings in years 1 and 3 for PhDs who place in industry against the relevance of PhD recipients in each field to industry. The relevance score is the average cosine similarity between the text of the recipient's dissertation and the text of patents assigned to the firms in that industry. The dashed lines are the fitted lines from regressions weighted by field size. Standard errors and significance levels are shown in parentheses. *** p<0.001.

A Additional Tables and Figures

Table A1: Sample Dissertations, Patents, and Similarities

Proquest	**US Patent**	**Similarity**
ProQuest #AAI1432716. Multiuser TDMA channel estimation. Wireless communication systems require efficient utilization of the limited available spectrum. There are various methods in which division of spectrum between users have been done till date, including frequency division...	U.S. Patent No. 7110378. Channel aware optimal space-time signaling for wireless communication over wideband multipath channels. A method and system is described for more optimally managing the usage of a wideband space-time multipath channel. The wideband space-time multipath channel is decomposed into a plurality of orthogonal sub-channels...	high
ProQuest #AAI3319598. Therapeutic potential of radial glial RG3.6 cells in rat spinal cord injury. Spinal cord injury (SCI) triggers a cascade of pathophysiological changes that lead to secondary tissue damage after the mechanical insult. Early after SCI, cells are disrupted and excitotoxic amino acids (e.g. glutamate) are released. Inflammatory cytokines and chemokines are quickly induced...	U.S. Patent No. 9717804. Regenerating functional neurons for treatment of disease and injury in the nervous system. Methods for producing new neurons in the brain in vivo are provided according to aspects of the present invention which include introducing NeuroD1 into a glial cell, particularly into a reactive astrocyte or NG2 cell...	medium
ProQuest #AAI1432716. Multiuser TDMA channel estimation. Wireless communication systems require efficient utilization of the limited available spectrum. There are various methods in which division of spectrum between users have been done till date, including frequency division...	U.S. Patent No. 9717804. Regenerating functional neurons for treatment of disease and injury in the nervous system. Methods for producing new neurons in the brain in vivo are provided according to aspects of the present invention which include introducing NeuroD1 into a glial cell, particularly into a reactive astrocyte or NG2 cell...	low

Notes–This table shows similarity scores for two pairs of sample Proquest dissertation text and USPTO patent text.

Table A2: Dyad Summary Statistics: Relevance of Dissertations to All Industries and to those in which PhD Recipients Place

	Unconditional on Placement			Conditional on Placement	
	(1) NAICS count	(2) Mean Cosine	(3) SD	(4) Mean Cosine	(5) SD
2-Digit NAICS	24	0.1237	0.1371	0.1986	0.1249
4-Digit NAICS	345	0.0258	0.0804	0.1670	0.1355
6-Digit NAICS	1,331	0.0093	0.0500	0.1537	0.1372

Notes – Column 1 presents the number of 2-, 4-, and 6-digit NAICS codes in which at least one Ph.D. recipient placed. Columns (2) and (3) report the mean and standard deviation of the mean cosine similarity score for each dissertation and each industry (unconditional on whether the person placed in that industry). Columns (4) and (5) report the means and standard deviations of the mean cosine similarity score for the industry in which the person placed. Across all three levels of NAICS codes, the mean cosine similarity scores is higher for the industry in which the person places than across all industries (i.e., unconditional on placement), implying that the dissertation-industry relevance is related to the industry in which PhD recipients place.

Table A3: Placement Shares, Relevance, and Wages

	Observations	Industry	Faculty	Postdocs	Industry	Faculty	Postdocs
		Panel A: Placement			**Panel B: Industry Relevance**		
Agriculture	350	0.382	0.157	0.461	0.154	-	-
Biology	2,800	0.327	0.091	0.582	0.182	-	-
Chemistry	900	0.522	0.025	0.453	0.211	-	-
Comm/Info Sciences	200	0.249	0.243	0.509	0.167	-	-
Eco/Envr Sciences	300	0.279	0.141	0.579	0.172	-	-
Engineering	4,100	0.673	0.102	0.225	0.281	-	-
Geosciences	400	0.395	0.077	0.528	0.183	-	-
Medicine	1,300	0.335	0.244	0.421	0.175	-	-
Mathematics	550	0.382	0.145	0.473	0.210	-	-
Physics	800	0.461	0.161	0.378	0.227	-	-
Social Sciences	750	0.232	0.354	0.414	0.166	-	-
		Panel C: Initial Earnings			**Panel D: Year 3 Earnings**		
Agriculture	350	62,920	66,990	33,860	73,630	73,330	46,780
Biology	2,800	58,170	71,860	33,610	78,040	75,300	43,560
Chemistry	900	74,530	72,140	37,380	89,630	69,730	56,230
Comm/Info Sciences	200	60,240	81,320	44,030	82,770	81,390	57,900
Eco/Envr Sciences	300	59,200	66,410	33,920	66,480	67,080	46,240
Engineering	4,100	98,520	88,030	37,210	124,100	94,010	60,720
Geosciences	400	79,470	75,370	38,550	88,880	78,220	50,380
Medicine	1,300	61,440	79,180	33,120	78,050	84,350	50,550
Mathematics	550	105,400	88,560	45,780	146,700	85,340	53,000
Physics	800	81,800	71,420	40,350	101,100	74,990	52,830
Social Sciences	750	73,090	88,290	36,090	82,840	86,110	50,810

Notes: Panel A displays the fraction of Ph.D. recipients who place in industry, faculty, and post-doc positions one year after receiving their Ph.D. Panel B displays the mean (over Ph.D. recipients whose initial placement is in industry) of cosine similarity score between each recipient's dissertation text and the text of the patents assigned to firms in the industry in which the recipient placed. Panels C and D show the mean wages of Ph.D. recipients one year and three years after receiving their degree based on the sector of their year-1 placement.

B Data Appendix

This section details each of our data sources and the links between them. We begin with graduate students in the UMETRICS data, whom we link to their Proquest dissertations. We focus on this group because we have, in the form of their dissertation, rich information about their degree year, field of study, and the specialized human capital at the end of their graduate training. We then link these Ph.D. recipients to their labor market outcomes using tax data (W2 records, the Business Register), allowing us to track their post-degree employment and earnings. In particular, we are able to determine whether they place in one of three mutually exclusive job types: faculty, postdoc, or industry. Finally, we use patent data from the U.S. Patent and Trademark Office (USPTO), along with a confidential patent-firm bridge at the U.S. Census Bureau, to construct a new dissertation-industry "relevance" score. This score allows us to directly measure how similar a Ph.D. recipient's specialized human capital is to the patenting portfolio of firms and industries, and sheds light on how well-matched each individual is to the industry in which they are actually employed after receiving their Ph.D.

B.1 UMETRICS Graduate Students

UMETRICS is an administrative dataset housed at the Institute for Research on Innovation and Science (IRIS) at the University of Michigan. It provides information on all grant transactions at 24 major research universities (64 campuses), which collectively account for more than one-third of federally funded academic R&D.[26] The data capture transactions from university grants to individuals providing labor (e.g. faculty, post-docs, students) and vendors providing goods and services (e.g. lab space, IT services, microscopes, lab animals). The UMETRICS data also include information on the job titles of individuals receiving payments from grants, which allows us to flag graduate students (Ikudo et al., 2019), the focus of this paper.

Using personally identifiable information (PII) from UMETRICS, such as name and (partial) birth date, Census staff probabilistically link UMETRICS employees to a Protected Identification Key (PIK), which is an internal person-level identifier at Census (Wagner et al., 2014). Once an individual is assigned to a PIK, we can link them to a variety of confidential data at Census (in our case, we link them to demographic and tax information – see below). We restrict our sample to UMETRICS employees who receive a single PIK.

B.2 Dissertations, Fields, and Specialized Human Capital

A crosswalk between the ProQuest's Dissertation and Theses Database and UMETRICS employees is available from IRIS. We use this crosswalk to link each UMETRICS graduate student to their dissertation, and thus further restrict our sample to graduate students who both receive a single PIK (see previous section) and who are linked to their ProQuest dissertations. Unfortunately, some PIKs link to multiple dissertations and some dissertations

[26] We use the 2018 Q4 release of UMETRICS data. For additional work using UMETRICS, see Chang et al. (2019), Ikudo et al. (2019), Buffington et al. (2016), Lane et al. (2013), Lane et al. (2015), Weinberg et al. (2014), Zolas et al. (2015).

are attributed to multiple PIKs. To simplify the analysis, we only keep the PIKs that match, 1-to-1, to a ProQuest dissertation. By confining our analysis to these 1-to-1 links, people, PIKs, and dissertations can be thought of interchangeably.

By restricting to graduate students who are assigned a PIK and a dissertation, we can measure the textual similarity between their dissertation and all patents in the USPTO and their match to the industry in which they place. Linking these patents to assignee firms, allows us to measure the relevance of a Ph.D. recipient's specialized human capital (as reflected by their dissertation) to different industries. This will shed light on how well-matched each individual is to the firm or industry in which they are actually employed after receiving their Ph.D. ProQuest also provides information on each Ph.D. recipient's degree year and field of study, which is crucial since much of our analysis takes place at the field level. We restrict our analysis to dissertations in a STEM field.

Since we can only link patents granted between 2000 and 2015 to their assignee firms (see Section B.4), we restrict our sample to dissertations with a degree year between 2004-2015. This is because we only use dissertation-linked patents granted during the four years prior to a dissertation year and the dissertation year itself, which ensures we are measuring similarity around the dissertation date (not after or long before) and ensures patent portfolios are not pulled toward dissertation topics after hiring Ph.D. recipients.

B.3 Labor Market Outcomes

As discussed, to track the labor market outcomes of the UMETRICS Ph.D. recipients, we use the Protected Identification Key (PIK) to link them to a variety of confidential data at the U.S. Census Bureau. First, we link them to the universe of W2 tax records for the years 2005-2018 (the years that were available at Census at the time of analysis), from which we identify the firms using the Employer Identification Numbers (EINs) at which they were employed as well as their earnings.[27] Next, we use the EINs to link each firm to the Business Register (BR), the universe of business establishments in the United States (DeSalvo et al., 2016). This allows us to determine the industry of the firms (technically, EINs) that employ each Ph.D. recipient in our sample.[28]

We use the information on firm industry (NAICS 611300 - "Colleges, Universities, and Professional Schools") as well as a list of EINs from the Integrated Postsecondary Education

[27]We use each PIK's dominant (i.e., highest paying) job. Using the dominant job means that each PIK is linked to one, and only one, EIN for a given year. When multiple jobs are tied for the highest earnings, we ensure that each PIK is paired with a single EIN (each year) by randomly breaking ties. Focusing on the dominant job simplifies the analysis and the disclosure of results.

[28]To do this, we assign a single dominant NAICS code to each firm (technically, EIN) in the W2 tax records using information from the County Business Patterns Business Register (CBPBR). The CBPBR contains establishment-year level data, and each establishment (in each year) has a NAICS code along with information on payroll and employment. Crucially, each establishment also has an accompanying EIN, which enables us to link industry, payroll, and employment information to the W2 EINs. Focusing on a two-year window around the W2 year, we assign to each W2 EIN the NAICS code with the largest payroll. If two or more NAICS codes have the same payroll, ties are broken using employment. If two or more NAICS codes have the same payroll and employment, ties are broken using establishment counts. If two or more NAICS codes have the same payroll, employment, and establishments, remaining ties are broken randomly. We focus on a two-year window (for a total of 5 years) to reduce the noise of yearly links.

Data System (IPEDS), which is maintained by the National Center for Education Statistics (NCSES), to identify universities and determine whether a Ph.D. recipient's initial post-degree job placement is in academia or industry. Note that IPEDS provides EIN information on nearly all U.S. universities, not just UMETRICS universities. This is important because, while a Ph.D. recipient may place at their UMETRICS university, they are more likely to place at a different, non-UMETRICS academic institution.

Within the subset of Ph.D. recipients that place in academia, we use earnings information to impute whether the job was a faculty or postdoc position. Specifically, we identify a position as being a postdoc if the Ph.D. recipient has W2 earnings below a cutoff specific to each degree year and field. To estimate the earnings cutoff, we use the Survey of Doctoral Recipients (SDR) to compute the mean and standard deviation of the earnings of postdocs in each degree year and field. We then classify a job placement as a faculty position if it is at least 1 standard deviation above the mean earnings of postdocs in that field and classify it as a postdoc position otherwise.

Thus, each Ph.D. recipient's initial placement is in one of three mutually exclusive job types: faculty members at a university, postdocs at a university, or industry (which includes the private sector, government, and NGOs). In addition to placements, we also track earnings during the first three years after a Ph.D. recipient earns their degree, subsetting our sample to Ph.D. recipients with positive earnings and non-missing industries in all years 1-3 following their degree year.

B.4 Patents and Dissertation-Industry Relevance

As noted, we want to measure how closely related a Ph.D. recipient's dissertation is to the patenting portfolio of each industry. We create this dissertation-industry dyadic data structure in a series of steps.

Initial Placement First, we link UMETRICS graduate students to their dissertations and use W2 records to determine the firms at which they work (along with their earnings) after they receive their degree (see Sections B.2 and B.3).

Relevance of Dissertations to Patents Second, we identify, for each Ph.D. recipient's dissertation abstract, the 1,000 most closely related patents in terms of text similarity (using methods described in Section 3 of the main text). If the cosine similarity for a dissertation patent link is zero, the link is discarded and if fewer than 1,000 patents have a positive cosine similarity to a dissertation, the dissertation will be linked to fewer than 1,000 patents. In the extreme, if no patents have a nonzero similarity score, the dissertation will not be linked to any patents and that graduate student is excluded from the analysis. Since the Census patent-firm bridge is for years 2000 to 2015 (see next section), patents granted outside this range cannot be assigned to a FIRMID. Thus, we subset to patents granted between 2000 and 2015. As noted in Section B.2, we also subset to patents granted 0-4 years before the paired dissertation's degree year. This restriction ensures we are measuring similarity around the dissertation date (not way before or after) and ensures patent portfolios are not pulled toward dissertation topics after hiring Ph.D. recipients.

Relevance of Dissertations to Firms We wish to use the links and similarity scores for dissertations and patents to construct a measure of how closely related each dissertation is to assignee firms and the industries in which these firms operate. Our third step uses a confidential patent-firm bridge derived from the Business Dynamics Statistics of Patenting Firms (BDSPF) at Census (Goldschlag and Perlman, 2017) to link patents granted in the years 2000-2015 to assignee firms and then aggregate the similarity score for a given dissertation-patent pair to a similarity score for dissertation-firm pairs. To simplify the analysis, we limit the patent-firm links to those where the patent is assigned to one and only one firm. By linking the dissertation-patent pairs (and their accompanying cosine similarity scores) to the patents' assignee firm, we can create a data set with unique observations at the dissertation-patent-firm level. We then collapse over the patents to obtain the average cosine similarity score for each dissertation-firm pair. Thus, each dissertation-firm pair inherits a similarity score that is comprised of a combination of several dissertation-patent similarity scores.

Relevance of Dissertations to Industries In our final step, we create dyadic data with dissertation-industry pairs. These data allow us to study how closely a Ph.D. recipient's specialized human capital matches the patenting portfolio of each industry, defined by the North American Industry Classification System (NAICS). We obtain NAICS codes for every U.S. business establishment from the Business Register (BR), which we use to assign a "dominant" NAICS code to each assignee firm on the basis of payroll, employment, and establishment counts. This process is similar to how we assign NAICS codes to each W2 EIN. However, instead of using the EIN as the firm identifier, we use an internal Census identifier called FIRMID, which allows us to link firms to the patent-firm bridge (which uses FIRMID as the firm identifier – see footnote 28).[29] Once a dominant NAICS code is linked to each firm (assignee firms with missing dominant NAICS codes are dropped), we collapse over firms to obtain a cosine similarity score for each dissertation-industry pair. Thus, each dissertation-industry pair inherits a similarity score from dissertation-firm pairs which inherited their scores from dissertation-patent pairs. At the end of this step, we have a dyadic data structure where each Ph.D. recipient is linked to every NAICS code and each pair has a similarity score as well. We can aggregate to any level of NAICS. In the paper, we aggregate to the 2-, 4-, and 6-digit levels.

B.5 Sample Restriction Summary

This section provides a summary of the sample restrictions imposed to arrive at our final sample. We begin with the dissertation-patent links, which we use to generate a cosine similarity score between each dissertation and each patent. The observations are uniquely defined by a dissertation-patent pair, and each dissertation is linked to at most 1,000 patents. We then make the following restrictions, which can be separated into dissertation/PIK side restrictions and patent side restrictions. Combined, these restrictions result in a final (rounded) sample of 12,450 Ph.D. recipients.

[29] We drop from our sample patents that are linked to firms that cannot be assigned to a NAICS code. We use this information to measure the research and development portfolios of firms.

PIK-Dissertation Restrictions

1. Subset dissertations to those that can be linked, 1-to-1, with a PIK.
2. Subset dissertations to those with a degree year between 2004-2015. (The Census patent-firm bridge is for years 2000 to 2015 and we only use dissertation-linked patents issued 0-4 years prior to when a dissertation is completed.)
3. Subset dissertations to those in a STEM field.
4. Subset PIKs to those with positive earnings in all years 1-3 following their degree year.
5. Subset PIKs to those who place at EINs with non-missing industries in all 1-3 years following their degree year.

Patent Restrictions

1. Subset to patents granted between 2000 and 2015. As indicated, the Census patent-firm bridge is for years 2000 to 2015, and so patents granted outside this range cannot be assigned to a FIRMID.
2. Subset to patents granted 0-4 years before the paired dissertation's degree year. Ensures we are measuring similarity around the dissertation date (not after or long before) and ensures patent portfolios are not pulled toward dissertation topics after hiring Ph.D. recipients
3. Subset to patents that have a single assignee firm.
4. Subset to patents to those whose assignee firm as a non-missing NAICS code.