NBER WORKING PAPER SERIES

PRODUCING HUMANITIES PHDS AMONG BAS AT DOCTORAL INSTITUTIONS

Todd R. Jones
Ronald G. Ehrenberg

Working Paper 22372
http://www.nber.org/papers/w22372

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2016

Cornell Higher Education Research Institute (CHERI) receives financial support from the Andrew W. Mellon Foundation, but the conclusions reported in this paper are strictly our own. The use of NSF data does not imply NSF endorsement of the research, research methods, or conclusions contained in this paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Producing Humanities PhDs among BAs at Doctoral Institutions
Todd R. Jones and Ronald G. Ehrenberg
NBER Working Paper No. 22372
June 2016
JEL No. I23

ABSTRACT

This paper investigates which attributes of a Carnegie PhD-level institution predict the share and number of its undergraduate humanities BA recipients that will go on to earn a humanities PhD. We use restricted-access individual-level Survey of Earned Doctorates data from the National Science Foundation (NSF) to determine both where and when PhD recipients received their BA. We use a truncation-correction methodology to account for problems inherent with studying PhD recipients, who often will receive their PhD after the data end. Using OLS, negative binomial regression, and an analysis similar to that of a prior, related paper, we find robust relationships between PhD production and student test scores, instructional expenditures per student, and the number of highly-ranked humanities PhD programs an institution has.

Todd R. Jones
Field of Economics
Cornell University
404 Uris Hall
Ithaca, NY 14853
trj25@cornell.edu

Ronald G. Ehrenberg
Cornell Higher Education Research Institute
271 Ives Hall East
Ithaca, NY 14853-3901
and NBER
rge2@cornell.edu

1. Introduction

While numerous studies have addressed the determinants of PhD students' times to degree and completion rates, only a few of these have been directed towards the humanities.[1] Despite long reported poor job market prospects for new humanities PhDs, graduate programs in the humanities have continued, sometimes with smaller cohort sizes, to churn out PhD students in the humanities. These programs also educate undergraduate students and our focus in this paper is on whether the scope and quality of a doctoral institution's PhD programs in the humanities influences the likelihood that humanities undergraduate students at the institution go on to receive PhDs in the humanities. Put another way, we are interested in the dual role that humanities departments at doctoral universities play in producing new PhDs and in generating undergraduate student interest in going on for doctoral degrees, typically at other institutions.

Using restricted access data from the National Science Foundation's *Survey of Earned Doctorates (SED*), we calculate that for students who received undergraduate degrees during the 1980-91 period, the probability that an undergraduate humanities major at a Carnegie Category bachelor's college went on to receive a PhD in the humanities is slightly higher than the probability that a undergraduate humanities major at a Carnegie Category doctoral university went on to receive a PhD in the humanities. However, because of the differences in the sizes of the two types of institutions, the ratio of those receiving PhDs in the humanities who were undergraduates at Carnegie Category doctoral universities as compared to those who received PhDs in the humanities and were undergraduates at Carnegie bachelor's institutions was almost 2.5. If one restricts the calculation to USNWR top 20 institutions in both categories, the ratio rises to almost 2.7.[2] Therefore, knowledge of the characteristics of the doctoral institutions that

---

[1] See, for example, Ronald G. Ehrenberg et al (2010), Ronald G. Ehrenberg et. al. (2007). Jeffrey Groen et al (2007), and Jeffrey Groen (2016).

[2] Because times to degree in PhD programs in the humanities can extend to over 20 years and the restricted access version of the SED to which we were granted access ends in 2011, we limit these calculations to the 1980-1991 period and use a maximum time-to-degree of 20 years to avoid issues of truncation (people who will eventually receive PhDs but are not yet observed in the data to have done so). In our empirical work, accounting for truncation is an important part of our analyses. These calculations were performed by dividing the number receiving humanities PhDs by the number of undergraduate humanities majors, meaning that some individuals from non-humanities majors entered into the numerator. We also only consider those institutions that produced at least one PhD during this time frame.

lead to more undergraduates going on to receive PhDs in the humanities is of importance in understanding the determinants of PhD production in the humanities.

Lemke (2009) is among the first to examine the determinants of undergraduate (BA) institutions in "generating" PhD recipients. His study computed a PhD production (generation) rate by dividing publicly-available data on the number of PhDs in all fields earned by alumni of selective liberal arts colleges by the number of BAs earned by the students at these institutions. A shortcoming with his approach is that while it *is* known where a PhD recipient received her BA, it *is not* known when she received it. Lemke assumed a 6-year time-to-degree (TTD) window from BA to PhD, which is quite reasonable for many STEM fields. However, this assumption is not in line with the data from the SED on humanities doctorates times to degree – it is much too short. Figure 1 shows that there is a great deal of heterogeneity in TTD in the humanities. A rather large fraction of individuals earning humanities doctorates take 15 or more years to receive a PhD.

Using restricted-access, individual-level *Survey of Earned Doctorates* data allows us to utilize the actual matches of where and when each individual graduated from her BA institution. As noted above, our focus is on PhD production in the humanities from Carnegie Category doctoral institutions. Over all the years our data cover, 58% of humanities PhDs come from students who received their bachelor's degree from Carnegie PhD institutions. This number was steady over the 10-year period 2002-2011.[3]

In the study of the PhD production rate, there is an inherent truncation problem where not all PhDs from a given BA cohort are observed because some individuals complete their PhD after the conclusion of the data. The truncation problem increases in severity for later cohorts. Using the SED data allows us to address this problem using truncation correction methodology

[3] For the entire time frame, these numbers are 18% for Master's institutions and 23% for BA institutions. For the 10-year window, these are 19% and 23%, respectively. These are computed after making the sample restrictions (other than on Carnegie classification) described in Section 3.

where early cohorts' patterns of time-to-degree are used to predict the number of PhDs that later cohorts will produce.

We use two main models to explore the relationship of numerous variables to PhD production, each model employing a different outcome variable. The first model uses the PhD production rate, the share of humanities bachelor's graduates that a given doctoral university produced in a year that ultimately receive a PhD in the humanities. After correcting for truncation, we use OLS to examine which features of the doctoral university and it undergraduate student body predict the share of humanities graduates who ultimately receive PhDs in the humanities. The second model employs negative binomial count models and uses the number of humanities PhDs produced from undergraduate humanities degree recipients as the dependent variable.

We present several different versions of each model. Across both main models, one of the variables that seems to matter in the production of PhDs is the number of humanities departments ranked in the top 10 of the National Research Council (NRC) 1995 rankings.[4] Put simply, the number of high quality doctoral programs in the humanities at a doctoral university is an important predictor of whether humanities undergraduates at the institution go on for PhD study in the humanities.

We find generally consistent results across our range of models, sample selections, and various versions of truncation-models. Key variables that are robustly associated with either a higher PhD production rate or higher PhD level (controlling for the number of humanities BAs) include incoming first-year students 75th percentile entrance test scores, instructional expenditures per student, and the number of humanities departments ranked in the top 10 in the 1995 NRC rankings.

---

[4] For comparison, in the appendix we also display results using an approach similar to that of Lemke.

This paper proceeds as follows. We first discuss PhD production, then proceed to describe the data, present summary statistics, methodology, results, and conclude.

2. PhD Production

We follow Lemke (2009) in assuming that characteristics of the doctoral institution and its undergraduates produce (generate in his terminology) PhDs. Among the variables we include in our analyses are the following: First, we include the student-faculty ratio because greater student interaction with instructors may be tied to a greater propensity to pursue doctoral studies. Second, we include the institution's 75th percentile test scores for first-year incoming students as a measure of its undergraduate students' academic ability. We also include several measures of the size and composition of the student body: the total number of students (including graduate students), the percentage of the students that are undergraduates, the percentage of undergraduates that is female, and the percentage of undergraduates that come from underrepresented minority groups. Because the importance of the humanities on campus could also play a role in generating student interest in going on for PhD study in the humanities, we also include the percentage of bachelor's degrees that were received in the humanities.

Expenditures on instruction and research may be associated with producing PhDs, so we include instructional expenditures per student and research expenditures per full-time faculty member. If the extent to which doctoral university faculty interact with undergraduate students differ between private and public universities, whether the institution is public or private may influence the production of PhD students.

Finally, we are in interested in whether having highly-ranked humanities doctoral programs is associated with having more humanities undergraduates who go on to earn a humanities PhD. Having more highly-ranked programs is presumably correlated with having more faculty at the cutting edge of their field, and, to the extent that these faculty interact with undergraduate students, this may influence students' decisions to obtain a PhD in the humanities. Thus, we include a set of variable indicating the number of humanities' doctoral programs a doctoral university has in each quality tier of programs (e.g. 1 to 10 and 11 to 25).

3. Data

We make use of several datasets in our analysis. The primary dataset is the National Science Foundation *Survey of Earned Doctorates* (SED). The SED is a survey administered to recipients of doctoral degrees at the culmination of their studies. The response rate to the SED is very high, being over 90% in each year since it started.[5] These data come at the individual-level and contain information on all PhD recipients whose bachelor's degree (henceforth denoted by BA) was obtained between 1980 and 2005 and who received a PhD by 2011.[6] Here, the year refers to the first year of the academic year; e.g., 1980-1981 is referred to as 1980.[7] We construct TTD by computing the length of time between BA and PhD receipt. Our measure of TTD thus includes any time pursuing other degrees, working, etc. before completing the doctoral degree.

As the focus is on humanities PhD recipients, we consider only the subset of fields which we classify as being in the humanities.[8] We drop observations that are missing BA institution or year of BA or PhD receipt. In addition, we focus our analysis on doctoral institutions as defined in the Carnegie 2010 classifications – those with Carnegie Codes 15 (Research Universities (very high research activity)), 16 (Research Universities (high research activity)), and 17 (Doctoral/Research Universities). We also do not consider BA institutions that are for-profit or are outside of the 50 states or Washington DC.

In order to be included in the sample, the BA institution must have produced at least one humanities PhD in our data. We add a BA year observation which takes on value 0 if there were no PhDs produced at any time from that BA year and BA institution.

---

5 See http://www.nsf.gov/statistics/infbrief/nsf06312/ - accessed March 4, 2016. Being over 90% is at least since 2006, the date of the referenced report.

6 There are a relatively small number of 2006 BA recipients, which we do not include in our analysis.

7 The SED data indicates the month and year of the completion of both the BA and the PhD. We consider completions occurring between July-December of year x and January-June of year x+1 as corresponding to school year x. If the month of completion is missing, we use the year variable as the year of completion.

8 We consider the following SED degree fields as being in the humanities: 700-799 with the exception of 773 (Archaeology). We additionally include 652 (Area/Ethnic/Cultural/Gender Studies), 676 (Linguistics), and 984 (Theology/Religious Education). Some definitions are very uncommon or have changed over the years; we do not make extra considerations for such cases.

We also use the *IPEDS Completions Survey* data to compute the number of BA degrees granted by each doctoral university each year, both overall and in the humanities. [9] This is used to compute the humanities PhD production rate, which is used in the OLS models. The fraction of BA degrees that are in the humanities is also calculated from this data. One point to note is that the Completions Survey switched from being under HEGIS to being under IPEDS in 1987. This resulted in the IPEDS code (UnitID) replacing the HEGIS code (FICE). It is not the case that there is not always a 1:1 mapping from FICE to UnitID – a single FICE can correspond to multiple UnitIDs. We handle this by combining the number of BA completions at the FICE level, even for 1987 and after. This results in an inflated number of BAs for some institutions. For the instructional variables, we merge at the UnitID level with the exception of the test score data.

We obtain additional explanatory variables, which are matched at the BA institution-BA year level, from a number of sources. The SED data is used to create a public institution indicator. We use the IPEDS Fall Enrollment dataset to obtain the total number of students. Specifically, we sum full-time students, having a weight of 1, with part-time students, having a weight of 0.4.[10] The Fall Enrollment dataset is also used to compute the share of students who are undergraduates, the share of undergraduates who are female, and the share of undergraduates who are from underrepresented minority groups.[11] These variables also take into accounts the above weights. We divide total students by full-time faculty counts from the IPEDS Fall Staff Survey[12] in order to compute the student-faculty ratio.

---

[9] The SED and Completions Survey both classify students into discipline categories, but the system is not the same. We were able to largely classify subjects in similar ways. One exception, however, is that archaeology is included with anthropology in the Completions Survey. For this reason, we do not classify anthropology or archaeology as humanities for either the SED or Completions Survey. We also consider only first reported BA major, which is potentially problematic if the student received two (or more) majors, one in the humanities and one outside the humanities.

[10] The weight of 0.4 for part-time students is very similar to the weight used in the IPEDS "Calculation of FTE Students." See https://nces.ed.gov/ipeds/glossary/index.asp?id=854 - Accessed February 20, 2016.

[11] Due to problems with subcategories not adding up correctly in early years, when computing certain of these variables, we did not use certain subcategories (such as FT or PT UG Unclassified). When computing race, the sum of the race categories were often less than expected; we assume that the race categories do not include the non_res variable.

[12] Full-time faculty counts come from the sum of the "Faculty, full-time men" and "Faculty, full-time women" variables.

Data from the Delta Cost Project (Lenihan, 2012) allows us to compute instructional expenditures per student and research expenditures per full-time faculty. The former is obtained by dividing instructional expenditures by full-time students; the latter comes from dividing research expenditures by full-time faculty.[13] From the Annual Survey of Colleges Standard Research Compilation,[14] we construct a measure of the academic quality of students at each institution, the 75th percentile test score, a weighted average of the reported ACT and SAT 75th percentiles scores.[15]

The National Research Council (NRC) has periodically produced rankings of doctoral programs across a range of disciplines. We use the 1995 rankings[16] and restrict our attention to the fields in the humanities.[17] We use the "scholarly quality of program faculty" ranking. We construct variables that count how many departments a given university had which were ranked

---

13 Specifically, instructional expenditures are from the instruction01 variable ("Expenditures for instruction – current year total") and research expenditures are from the research01 variable ("Expenditures for research – current year total").

14 The score data includes two identifies, its own ID and FICE. However, there is not always a 1:1 mapping between the two. After cleaning the test score data, in cases where the FICE code maps to multiple IDs, we select one ID to use based first on the one that seems to match the institution name the best, second based on the one that has non-missing test scores, and finally based on the one with the highest test scores. If ID maps to multiple FICE, we use the FICE that is in the main dataset for matching. Finally, there are cases where there is no match between datasets via FICE, but the institution name is associated with a different FICE code in the score data. We manually make these matches. Finally, we manually match a number of cases by institution name if the FICE codes do not match between datasets. We dropped the data from 1983 as it is unreliable.

15 The composite SAT score, the sum of an institutions verbal and math 75th percentile scores, was converted to its ACT equivalent using the conversion found at https://web.archive.org/web/20160130223549/http://act.org/solutions/college-career-readiness/compare-act-sat/ - accessed April 27, 2016. We rounded up to the ACT score if the SAT score was outside one of the listed ranges. For example, if the range was 1290-1320 for a 29 and 1330-1350 for a 30, we converted a score of 1325 to 30. We made note of the re-centering of the SAT scores in 1995 and have converted all scores prior to 1995 to their current day equivalent (separately for math and verbal) using the SAT I Individual Score Equivalents conversion table from the CollegeBoard (see http://research.collegeboard.org/programs/sat/data/equivalence/sat-individual - accessed April 27, 2016). The chart presented single conversion numbers. We rounded to the higher number. For example, a verbal score of 690 is re-centered to 750 and 700 is re-centered to 760. We re-centered 695 to 760. For cases where there was a test score for SAT, but not ACT, we used the SAT score entirely and vice versa. As there was a great deal of missing data on earlier years of percentage taking each test, we backfill years missing this data with the earliest reported percentages. We also fill in some of the missing score data using linear interpolation.

16 Because we only use 1995 rankings, these variables are constant within institution (across BA year). There is a high correlation between the rankings that were published in 1995 (Goldberger et al. 1995) and those published in the 1980s (Jones et al. 1982). The 1980s ranking were for fewer institutions, but for departments that appeared in both years (and after re-ranking them to reflect only the departments in consideration), the correlations of the number of departments at an institution in each rankings interval were as follows : 1 to 10 (.942), 11 to 25 (.888), 26 to 50 (.867), and 51+ (.805). If one does not restrict to departments in both, somewhat smaller, but still high correlations are observed. This way is not perfect, but yields the following: 1 to 10 (.917), 11 to 25 (.842), 26 to 50 (.797), and 51+ (.593). The NRC evaluation of doctoral programs conducted in the first decade of the 21st century (Ostricker et. al. 2011) did not provide point estimates of program rankings and thus could not be used in our analysis.

17 The NRC fields that we use for humanities are: art history, classics, comparative literature, English language and literature, French language and literature, German language and literature, linguistics, music, philosophy, religion, Spanish and Portuguese language and literature, and history.

in the top 10, between 11 and 25, between 26 and 50, and greater than 50.[18] We point out that there is heterogeneity in the total number of departments ranked in each field of study. For example, many more departments are ranked in English Language and Literature than in French Language and Literature.

We face the challenge of missing data. For many of the control variables, we use a linear extrapolation to fill in data for years that are bookended by two other years with non-missing data. There are several variables that contain missing data that is not bookended because of when the survey started. In these cases, we backfill using the first possible year of data (for example, we fill in 1980-1983 with the 1984 value).[19] After this process, if an observation is still missing, we code it as 0 and create a missing data dummy that takes on value 1 for missing and 0 otherwise. Universities not included in the 1995 NRC are coded as having zero programs in each of our quality intervals and we do not include a missing dummy variable for such institutions.

4. Summary Statistics

Summary statistics are presented in Table 1.[20] The unit of observation is a Carnegie Category doctoral university; each university contributes multiple observations, one for each year. Means and standard deviations are conditional on the observation not being missing (after filling in some of the missing observations as described in Section 3). The fraction missing for each variable is found in the final column. Most of the variables have a small percentage missing. The test score variable and expenditures variables have a larger percentage missing. As pointed out above, we do not consider institutions that do not have a department ranked in the

[18] We do not restrict the departments to those institutions that appear in our data. The bands are not equally spaced; results should be interpreted accordingly.

[19] More specifically, we fill in full-time faculty of years 1980-1986 with 1987; test scores of 1980-1984 with 1985; and instructional and research expenditures of 1980-1986 with 1987. We note that this is not ideal as it does not take into account trends, only takes care of cases that have a non-missing value for the first year of data, and results in a lack of variation over these several years. We compute student-faculty ration, instructional expenditures per student, and research expenditures per faculty after backfilling.

[20] Observations that do not contribute to regression results because e.g., they did not have any humanities BAs, are included in the summary statistics.

1995 NRC data as being missing. The 48% reported here indicates 48% of the doctoral universities did not have a single department that appeared in the NRC data.[21]

The mean student-faculty ratio is 18. Mean incoming freshman 75th percentile ACT scores are fairly high, at 27. The average institution has a student body of 13,400 students, 76 percent of whom are undergraduates. Of the undergraduates, about half are female, and about a sixth are minority. The average instructional expenditures per student is $8,000, while the average research spending per faculty member is about $50,000, with a rather large variance. Of the BA graduates, about 12% come from the humanities. Finally, the average number of departments in the NRC rankings is shown broken up into several bands. Note that the total number of institutions in the NRC data is not extremely large, so there are not as many departments ranked in the 51+ interval.

It sometimes happens that an individual who receives a BA in a field other than one in the humanities goes on to earn a humanities PhD. As described in Section 5.2, there are reasons both to consider and to not consider such individuals in the analysis. We see that 6,651 humanities PhD recipients earned a non-humanities BA and 1,581 have missing data on their major. For comparison, 31,667 majored in the humanities. Table 2 shows which broadly-defined fields such individuals earned a BA in. The largest category is the social sciences, followed by education. Relatively few individuals who received humanities PhDs majored in computer, math, or physical science as undergraduates.

5. Empirical Strategy

We first describe the truncation-correction that we use to account for PhD degrees that are not observed in the data due to the data not extending long enough into the future. A simple model of truncation is presented in the following section; descriptions of two additional truncation models (linear and quadratic) appear in the appendix. Following the discussion of the truncation-correction methodology, we describe the estimating models employed to examine the

[21] In results not shown, for each NRC range, there are a large number of 0s. This declines in prevalence from the 1-10 to 11-25, from 11-25 to 26-50, and from 26-50-51+.

importance of attributes of the undergraduate institution in producing humanities PhDs, namely OLS and negative binomial regression.

5.1. Truncation Correction

There is a large amount of truncation in our outcome data. Specifically, we do not observe all of the PhDs that are produced from a given BA cohort because time-to-degree (TTD) varies by individual and is sometimes greater than the number of years we observe in the data for an individual.[22] For example, if the BA year for a given institution is 1999, we will only observe PhDs completed by 2011, when the SED data ends. Long TTD is particularly prevalent in the humanities. Figure 1 displays the distribution of TTD using our sample of humanities PhD received by recipients of BAs at Carnegie PhD institutions. We restrict the BA years to be 1980-1986 in this table so as to allow at least 25 years of observation post-BA. We also do not show TTD of 25 years or greater. We see a great amount of variation in TTD.

We address the truncation problem by estimating the total number of PhDs that are produced by an institution for each BA year. We use a maximum TTD of 20 years.[23] In other words, we consider only PhDs that take 20 or less years to complete after the BA graduation year.

The main idea with the truncation correction methodology is to first estimate model parameters from the earlier data, the BA years for which we observe at least 20 years of SED data, and second to apply these parameters to the more recent years in order to estimate the number of PhDs produced for the latter years.

To obtain the estimating sample, we restrict the data to institution-BA years (1980-1991) in which we observe all of the following 20 years in the data. For the 1980 class, we consider only PhD receipt until 2000 and do not consider PhDs obtained after this year even if they appear in the data.

---

[22] Another reason for this is that, although response to the Survey of Earned Doctorates is very high, it is not 100%.

[23] We also present results for a maximum TTD of 25 years in the Appendix.

Using the estimating sample aggregate the data so that it is at the TTD-level. In other words, we calculate the total number of PhDs granted after 4, 5,…,20 years from the BA year (regardless of what the BA year was). Thus, we have 16 total observations. We then compute the cumulative probability of receiving a PhD by each TTD. If 65% of students graduated within 12 years post-BA, then the cumulative probability for 12 years is 0.65. Because we are only considering up to 20 years, the cumulative probability of the 20^{th} year is 1.00. We save these probabilities to be used later.

For the institution-BA years (1992-2005) for which we do not observe the entire 20 TTD years post-BA, we create a running total of the number of PhDs granted by 2011, the last year we observe PhD receipt. We note the maximum possible number in the data from BA year for each institution-BA year by subtracting the BA year from 2011 (for BA year 1999, this is 12 years). We then multiply the total number of PhDs received by 2011 (after 12 years) by the inverse of the probability corresponding to the 12^{th} year that we estimated previously. If the number of PhDs received after 12 years at a given institution-BA year is x and the probability of receiving a degree by 12 years (calculated previously) is 0.65, we obtain the estimated total number of PhDs after 20 years for this institution-BA year by multiplying x by (1/0.65).

For institution-BA years for which we do not observe any receipt of PhD in year 2011, we still apply the percentage based on the maximum possible number TTD years in the data. For example, if the BA year is 1999 and we observe a total of 5 PhDs granted from 1999 until 2009, but none in 2010 or 2011, we still compute the estimated PhDs based on a 12-year TTD instead of 10. For institution-BA years for which we do not observe any PhDs granted by 2011, we assign them a value of 0 PhDs produced. For the estimating sample, we use the actual number of PhDs produced after 20 years.

5.2 Estimation

OLS

We divide the actual (estimating sample) or predicted number of PhDs produced from a BA institution-BA year by the number of humanities BA graduates for that BA year to compute the PhD production rate. [24] We regress this on the explanatory variables (and the missing dummies). In all of our OLS specifications, we weight the observation by the number of humanities BAs the institution produced in the BA year.[25] We cluster standard errors at the FICE level.

We split our sample in several different ways. Only a subset of institutions had at least one doctoral department ranked in the NRC 1995 data. We present results both restricting and not restricting to these institutions.

We also define the numerator of the PhD production rate, the count of PhD recipients, in two different ways. The first version only considers those humanities PhD recipients for whom we also observe as having completed a BA in the humanities. The second version considers all recipients, regardless of BA major. Because the denominator of the PhD production rate is the number of BAs in the humanities, the first may be more natural. However, there are a fairly small number of observations that have missing data on which major they earned (see Table 2). Assuming some, but not all, of these cases received a humanities BA, we will understate humanities PhD production among only humanities BAs in the first definition, which restricts to those for which we observe having completed a humanities BA. The second definition will overstate it as it also includes both those with non-humanities BAs and missing BAs. However, this second version is the right one in terms of quantifying the total number of BAs who go on to earn a humanities PhD, not taking into account in what they majored.

---

[24] If an observation is missing data on the number of humanities BAs or if this number is 0, then this observation will have a missing share variable and will not be included in the regression. In the negative binomial model, it will be included if it is 0 and will not if it is missing.

[25] It sometimes happens that an institution is predicted to have a very high (even greater than 1) PhD share. It is likely in such cases that the institution awards a very small number of humanities BAs, but produces a disproportionately high number of humanities PhDs. This can be because a relatively large share of the humanities BAs went on the a PhD; it can also be the case that non-humanities BAs go on for a PhD. Such cases can produce a very high predicted PhD share. Because there are very few humanities BAs, these observations will receive a small weight.

Negative Binomial Model

We also present results using the negative binomial count model, which may provide a better model for our data, which are non-negative count data. In this model, we include the truncation-corrected predicted number of PhDs produced (the numerator of the PhD production rate) as opposed to the PhD production rate.[26] We do not weight by humanities BAs, but include this as a control variable. We do not include the fraction of BAs that are in the humanities in this model because interpretation is not straightforward after already controlling for the number of total students and the number of humanities BAs. Due to the nature of the negative binomial model, we exclude any observations with a predicted number of PhDs produced that is strictly less than 0. Results are presented broken up into the same four versions as described in the above OLS section.

6. Results

We first present the estimates resulting from the OLS regressions. This is followed by those obtained using the negative binomial model.

OLS

Table 3 presents the baseline OLS results, the outcome variable being the truncation-corrected PhD production rate - the share of humanities BAs who go on to receive a humanities PhD. For ease of interpretation, we multiply the share by 100. The other percentage variables described below are also multiplied by 100.

As described in section 5.2, we present results in four combinations. The first two columns calculate the share by dividing the number of humanities PhDs *whose first BA major was also observed to be in the humanities* by the total number of humanities BAs. Columns 3 and 4 instead use the number of humanities PhDs *regardless of BA major,* keeping the same

[26] In some cases, the truncation methodology for the linear and quadratic models (in the Appendix) produce negative predicted total PhDs. We exclude such cases from the sample when performing negative binomial regression.

denominator. Columns 2 and 4 restrict the sample to those institutions appearing in the NRC 1995 data, while columns 1 and 3 do not make this restriction.

The mean PhD production rate (at the BA institution-BA year level) for the first specification is shown at the top of the bottom section. This is interpreted as 2.1 percent of humanities BA graduates go on to receive a humanities PhD (after correcting for truncation) within 20 years of their BA. Several of the coefficients in this model are statistically significant. Institutions with higher-ability undergraduate students, as proxied for by the 75 percentile incoming freshman test scores, have a higher PhD production rate. The total number of students at the university is negatively related to the share, as is the share of undergraduates who are female. Instructional expenditures per student is positively correlated with producing humanities PhDs, providing evidence that institutions that devote more resources to teaching produce more PhDs in the humanities. Having PhD programs in the top 10 of the 1995 NRC rankings is associated with a higher share, the magnitude of which is quite large – having one more program in the top 10 is correlated with to a share that is 0.35 higher. The other categories below top 10 also have positive coefficients, with 26-50 and 51+ being significant. Variables that are not statistically significant include student-faculty ratio, percent minority, percent undergraduate, percent humanities, research expenditures per full-time faculty member, and public university status (although the magnitude of this last variable is quite large).

In column 2, which is restricted to institutions in the NRC data, the mean share is slightly higher than in column 1. This suggests that the subset of institutions which appear in the NRC data (about half of all in the sample) have a higher propensity to produce humanities PhDs than those that do not appear in this data. The amount of variation explained by these variables, as seen by R-squared, is also somewhat higher here than in the first column. Similar results are obtained as in column 1, the exception being that the final two NRC categories decrease in magnitude and lose significance.

Columns 3 and 4 are comparable to columns 1 and 2, respectively, with the exception that the share variable is calculated for all BA majors, including those not in the humanities.[27] The results are, for the most part, robust to this redefinition. The PhD production rates are, not surprisingly, higher than in the analogous rates in the first two columns.

Appendix Table A1 shows results using the additional definitions of the truncation-correction described in section A.1. The sample selection in column 2 of Table 3, our preferred specification, is replicated in Table A1, Column 1. This sample selection is used across all 6 columns, each of which use a different version of truncation-correction. The results are largely robust to the different versions.[28]

Negative Binomial Model

The main count model results are displayed in Table 4. There are two important things note. First, the dependent variable is now the 20-year truncation-corrected count (as opposed to share) of PhDs. Second, we now control for the number of humanities BAs (as opposed to using it in the denominator of the dependent variable). The results are presented in an analogous way to those in Table 1. Columns 1 and 3 use all institutions, regardless of whether they appear in the NRC or not, while column 2 restricts to NRC institutions. Columns 1 and 2 restrict the dependent variable to those whose first BA major was in the humanities; columns 3 and 4 do not.

The mean of the dependent variable, the number of PhDs produced is lower in column 1 than in column 3, implying that, on average, around 1.3 individuals whose first major was not in the humanities (or missing) went on to receive a humanities PhD. A similar pattern can be seen between columns 2 and 4. In addition, the NRC-only sample produced considerably more humanities PhDs than those not included in the NRC data. This is because such schools produce more humanities PhDs at a higher rate and/or they are larger.

---

[27] The number of observations differs between columns 1 and 3 because in order for an institution to make it into the sample, it must have produced at least one PhD over the sample window. It happens that some institutions produced only PhDs with non-humanities BAs. Such institutions, with all of the filled in years of 0 PhDs produced, are included in columns 3 and 4.

[28] The main exception is Column 6, which uses the 25-year quadratic correction. This produces a much higher mean share and often has more extreme coefficient estimates, although most of the variables that are significant in the other columns remain so here.

Compared to the OLs results in table 3, the explanatory power is substantially lower, noting that the former is measured using pseudo R-squared and the latter adjusted R-squared.

Not surprisingly, the coefficient on the number of humanities BAs (divided by 1000) is positive, large, and highly statistically significant –after controlling for the total number of students, the greater the pool of humanities undergraduates BAs, the greater the number who will ultimately earn a PhD, on average. We see that many of the variables that were statistically significant in Table 3 continue to be so here, including test scores, instructional expenditures per student, and several of the NRC variables. Total students is significant in column 3. However, its sign is different than in our earlier OLS model. The interpretation of this variable differs between the two models because of the different definition of the outcome variable as well as the fact that we are now including the number of humanities BAs (and excluding the percentage of BAs who are in the humanities) as explanatory variables. The fraction of undergraduates who are minority is negative and sometimes significant, but this coefficient is extremely small and economically not very meaningful. There is a similar finding with percentage of students who are undergraduate, although in the opposite direction, always significant, and somewhat larger in magnitude.

Table A2, Column 1 replicates Table 4 Column 2. The results are again largely similar across the different truncation models.[29]

7. Conclusion

Our major empirical finding is the strong positive association between the number of highly ranked humanities PhD programs at an institution and the number of its undergraduate students that go on for PhDs in the humanities. Strong doctoral programs in the humanities thus seem to contribute to the supply of humanities PhDs both through the number of PhDs they directly generate and through their impact on the number of their undergraduate students that go

[29] Again, Column 6 exhibits somewhat different behavior.

on for PhDs. We also find that higher levels of instructional expenditures per student and higher entering test scores for undergraduate students are both associated with more undergraduates at the institution going on for PhDs in the humanities.

By using the restricted access *Survey of Earned Doctorate* individual observation data, we were able to determine each humanities PhD recipient's time-to-degree and, given the heterogeneity in time to degree, to compute better estimates of the share of bachelor's graduates at an institution in a year that subsequently received PhDs in the humanities than did the early method used by Lemke (2009).[30] Our method required us to address the problem that when using the individual level SED data we may not observe all PhDs produced from bachelor's recipients at an institution in a year because if an individual earns a PhD after our data end, we will not observe him or her as a PhD recipient. This required us to use a truncation methodology to estimate how many PhDs in the humanities would ultimately be received by bachelors' recipients from an institution in a year. While we believe this is an improved methodology, in the appendix, we also present estimates using a variant of Lemke's method assuming a fixed 10 year time-to-degree and find that are our major empirical finding continues to be observed.

[30] As noted earlier Lemke used institutional level data on the number of PhDs in a year received by bachelor's recipients from an undergraduate institution and matched them up to the number of bachelor's degrees granted by the institution a fixed number of years earlier.

References

Ehrenberg, R.G., Jakubson, G. H., Groen, J.A., So, E., and Price, J. (2007). "Inside the Black Box of Doctoral Education: What Program Characteristics Influence Doctoral Students' Attrition and Graduation Probabilities?", Educational Evaluation and Policy Analysis 29 (June):124-150

Ehrenberg, R.G., Zuckerman, H. , Groen, J.A. and Brucker, S.M. (2009). Educating Scholars: Doctoral Education in the Humanities. Princeton University Press, Princeton NJ

Goldberger,M.L., Maher, B.A., and Flattau,P.E. eds. (1995). Research doctoral Programs in the United States: Continuity and Change. National Academy Press, Washington DC

Groen, J. (2016). "The Impact of Labor Demand on Time to Doctorate". Education Finance and Policy Analysis 11 (Winter) 43-69

Groen, J., Jakubon, G.H., Ehrenberg, R.G., Condie, S., abd Liu, Albert Y. (2008). "Program Design and Student Outcomes in Graduate Education". Economics of Education Review 27 (April): 111-124

Jones, L.V. , Lindzey, G. and Coggshall, P.E. eds. (1982). An Assessment of Research – Doctorate Programs in the United States. National Academy Press, Washington DC

Lemke, R. J. (2009). Generating Doctoral Degree Candidates at Liberal Arts Colleges. In Ehrenberg, R. G. and Kuh, C. V., Editors, Doctoral Education and Faculty of the Future, pages pp. 93-108. Cornell University Press, Ithaca, NY.

Lenihan, C. (2012). IPEDS Analytics: Delta Cost Project Database 1987-2010 (NCES 2012-823). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Ostricker, J.P., Holland, P., Kuh, C. and Voytuk, J. eds.(2011). A Data-Based Assessment of Research-Doctorate Programs in the United States. National Academy Press, 2011.

A. Appendix

Section A.1 describes details on the additional truncation models. Section A.2 presents the Lemke-style analysis, which is compared to the results found in Section 6.

A.1. Additional Truncation Models

In the main body of this paper, we presented results using the simple model of truncation correction with a 20-year maximum TTD. Here, we describe two additional truncation correction models, the "linear" model and the "quadratic" model. Selected results of all three models, with a TTD of both 20 and 25 years are presented in Tables A1 and A2.

Linear Truncation Model

The linear model allows for the number of PhDs being produced to follow a linear trend based on how many years the BA-year is from 1980. This is in contrast to the simple model, which does not directly take into account the year of BA, only the maximum number of years possibly observed in the data.

As in the simple model, we first restrict the data to the BA years for which we observe the maximum TTD in the data. For a 20 year maximum TTD, which we will assume throughout this description for expositional purposes, this is BA years 1980-1991. However, instead of aggregating to the TTD-level, we aggregate to the BA year-TTD level. For each BA year, this produces the total number of PhDs earned (across all institutions) after 4, 5,…, 20 years. Again for each BA year, we compute the cumulative probabilities of completing PhDs by each TTD. Call these probabilities $deg\_prob_{YEAR,TTD}$, where *YEAR* denotes BA year and *TTD* is time-to degree (taking on values 4,…,20).[31] We also compute how many years the BA year is from 1980: $time\_1980 = YEAR - 1980$.

For each TTD, we run a separate regression, which is of the form:

$$deg\_prob_{YEAR,TTD} = \beta_0 + \beta_1 time\_1980_{YEAR,TTD} + \epsilon_i$$

[31] For all truncation models, we group TTDs of less than 4 years with 4 years.

We obtain the coefficients and apply them to the years for which we do not observe the maximum TTD in the data. We do this in an analogous way to the simple model. Specifically, we multiple the number of degrees produced after the max number of possibly observed years by the inverse of the sum of the intercept term and the coefficient on the year term multiplied by the number of years since 1980.

Quadratic Truncation Model

The quadratic model is nearly identical to the linear model with the exception that we add a squared *time_1980* term to the regression equation.

A.2. Lemke-Style Regression

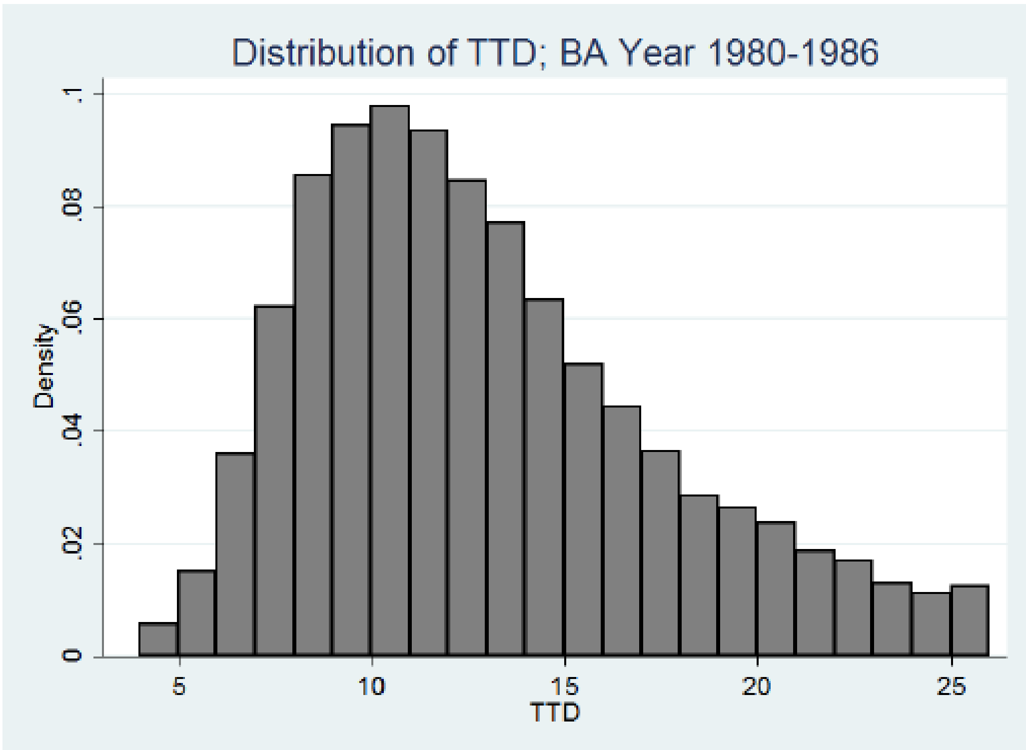
We also present results, for comparison with those presented above, of an analysis conducted similar to, but not exactly the same as, that done in Lemke (2009). Because this author did not have access to individual-level SED data on PhD recipients, public-use data was instead used. This data splits PhD recipients by BA institution, but it does not specify which year these individuals received their BAs. Thus, a 6-year TTD was assumed.

Instead of a 6-year TTD, we use a 10-year TTD. Using our SED data, for each BA institution, we aggregate the number of PhDs received by PhD year, tossing out information on when the BA was received. Institution-PhD years with no PhDs are coded as 0. We also restrict our sample to PhD received in 1990-2011, allowing 10 years from the beginning of our BA data.[32] We then compute a PhD production rate, which is the number of PhDs by PhD year for each BA institution divided by the number of humanities BAs from 10 years before the PhD year. We use the explanatory variables associated with the BA year.

[32] Because our SED data only covers individuals who received their BA in 1980 and later, we observe a relatively small number of PhDs for the early PhD years. The earlier the PhD cohort, the fewer BA cohorts can contribute to the number of PhDs observed. For example, the we would observe up to a TTD of 20 years for PhD year 2000, but only up to a TTD of 10 years for PhD year 1990.

We run the same analysis as presented in the OLS results, making the results in here comparable with those in Tables 3 (and A1). In many ways, the results in Table A3 are very similar to those in Table 3. Similar magnitudes and coefficients are observed. By and large, the same variables are also statistically significant.

Figure 1: Distribution of TTD



Only TTD of 25 years and less is considered. TTD of less than 3 is combined with 4.

Table 1: Summary Statistics

	Mean	SD	Missing
Stud.-Facul. Ratio	17.86	14.31	.001
75 Percentile Score	26.73	3.2	.133
Total Students	13.42	9.11	.001
% UG	76.13	14	.012
% Female	51.03	10.04	.012
% Minority	16.75	18.84	.012
Instr. Exp./Student	7.77	6.97	.057
Rsch. Exp./FT Fclty	50.45	77.39	.122
Public	.64	.48	0
% Humanities	12.4	8.4	.006
NRC 1-10 BA	.84	2.15	.48
NRC 11-25 BA	1.23	2.08	.48
NRC 26-50 BA	1.48	1.91	.48
NRC 51+ BA	1.21	1.13	.48

Observations: 6994. The means and standard deviations are conditional on not being coded as missing. The missing column is presented as fractions. Total students, institutional, expenditures per students, and research expenditures per student are in units of 1,000. The missing column for the NRC variables indicates the percentage that did not have at least one category appear in the NRC rankings for humanities departments.

Table 2: BA Fields of Humanities PhDs with Non-Humanities BA

	(1) N	Pct.	Cum. Pct.
Business	443	5.38	5.38
Communication	707	8.59	13.97
Computer	114	1.38	15.35
Education	877	10.65	26.01
Engineering	356	4.32	30.33
Life Science	528	6.41	36.75
Math	291	3.53	40.28
Physical Science	307	3.73	44.01
Psychology	546	6.63	50.64
Social Science	2215	26.91	77.55
Other	267	3.24	80.79
Missing	1581	19.21	100.00
Total	8232	100.00	

SED major fields are classified as follows: Business: 900-939; Communication: 940-959; Computer: 400-419; Education: 800-899; Engineering: 300-399; Life Science: 0-299; Math: 420-499; Physical Science: 500-599; Psychology: 600-649; Social Science: 650-699, 773; Other: 960-989. Missing indicates that no BA field was indicated.

Table 3: OLS Regression

	(1) Share simple 20	(2) Share simple 20	(3) Share simple 20	(4) Share simple 20
Stud.-Facul. Ratio	0.003	0.002	0.001	-0.001
	(0.005)	(0.008)	(0.007)	(0.010)
75 Percentile Score	0.138***	0.140**	0.193***	0.193***
	(0.037)	(0.045)	(0.043)	(0.053)
Total Students	-0.031**	-0.032*	-0.038**	-0.039*
	(0.012)	(0.014)	(0.014)	(0.016)
% UG	0.013	0.015	0.014	0.016
	(0.011)	(0.013)	(0.013)	(0.016)
% Female	-0.027*	-0.028*	-0.034*	-0.034
	(0.011)	(0.014)	(0.013)	(0.018)
% Minority	0.003	-0.005	0.006	-0.003
	(0.004)	(0.009)	(0.005)	(0.011)
Instr. Exp./Student	0.046**	0.048**	0.051**	0.054*
	(0.015)	(0.016)	(0.018)	(0.021)
Rsch. Exp./FT Fclty	-0.001	0.000	0.000	0.001
	(0.001)	(0.002)	(0.002)	(0.003)
Public	-0.444	-0.436	-0.579	-0.593
	(0.324)	(0.371)	(0.380)	(0.436)
NRC 1-10 BA	0.353***	0.318***	0.404***	0.361***
	(0.071)	(0.080)	(0.090)	(0.101)
NRC 11-25 BA	0.015	-0.019	0.012	-0.032
	(0.064)	(0.077)	(0.077)	(0.094)
NRC 26-50 BA	0.086*	0.051	0.091*	0.045
	(0.038)	(0.045)	(0.044)	(0.054)
NRC 51+ BA	0.092*	0.004	0.125*	0.008
	(0.043)	(0.075)	(0.052)	(0.090)
% Humanities	-0.007	-0.001	-0.027	-0.024
	(0.016)	(0.023)	(0.018)	(0.028)
PhD Prod. Rate Mean	2.062	2.293	2.616	2.885
PhDs w/ Hum BA Only	Yes	Yes	No	No
NRC Only	No	Yes	No	Yes
Year Dummies	Yes	Yes	Yes	Yes
Missing Dummies	Yes	Yes	Yes	Yes
N	6578	3609	6694	3609
Adjusted R-squared	0.42	0.46	0.37	0.40

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Negative Binomial Regression

	(1) # PhD simple 20	(2) # PhD simple 20	(3) # PhD simple 20	(4) # PhD simple 20
# PhD simple 20				
Stud.-Facul. Ratio	0.001	0.004	-0.002	0.003
	(0.003)	(0.003)	(0.002)	(0.002)
75 Percentile Score	0.102***	0.102***	0.106***	0.105***
	(0.020)	(0.023)	(0.018)	(0.021)
Total Students	0.009	0.003	0.017**	0.009
	(0.006)	(0.005)	(0.006)	(0.005)
% UG	0.008*	0.011*	0.009*	0.011*
	(0.004)	(0.005)	(0.004)	(0.005)
% Female	0.005	0.001	0.012	0.001
	(0.007)	(0.007)	(0.007)	(0.007)
% Minority	-0.006**	-0.004	-0.005*	-0.002
	(0.002)	(0.003)	(0.002)	(0.003)
Instr. Exp./Student	0.013*	0.010*	0.012*	0.009*
	(0.006)	(0.005)	(0.006)	(0.004)
Rsch. Exp./FT Fclty	-0.001	-0.001	-0.000	-0.001
	(0.001)	(0.001)	(0.000)	(0.000)
Public	-0.071	-0.080	-0.043	-0.102
	(0.141)	(0.159)	(0.138)	(0.149)
NRC 1-10 BA	0.151***	0.147***	0.160***	0.139***
	(0.028)	(0.024)	(0.029)	(0.022)
NRC 11-25 BA	0.055*	0.065**	0.046	0.057**
	(0.027)	(0.023)	(0.028)	(0.021)
NRC 26-50 BA	0.091***	0.088***	0.081***	0.075***
	(0.025)	(0.022)	(0.024)	(0.019)
NRC 51+ BA	0.138***	0.059	0.156***	0.066
	(0.029)	(0.037)	(0.028)	(0.034)
Hum BA	1.609***	1.175***	1.413***	1.021***
	(0.254)	(0.202)	(0.248)	(0.183)
PhD Count Mean	5.781	9.193	7.125	11.565
PhDs w/ Hum BA Only	Yes	Yes	No	No
NRC Only	No	Yes	No	Yes
Year Dummies	Yes	Yes	Yes	Yes
Missing Dummies	Yes	Yes	Yes	Yes
N	6786	3640	6994	3640
Pseudo R-squared	0.15	0.14	0.15	0.13

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1: OLS Regression; PhDs w/ Hum. BA Only; NRC Only

	(1) Share simple 20	(2) Share simple 25	(3) Share lin 20	(4) Share lin 25	(5) Share quad 20	(6) Share quad 25
Stud.-Facul. Ratio	0.002	0.002	0.001	0.002	-0.001	-0.015
	(0.008)	(0.009)	(0.007)	(0.008)	(0.007)	(0.029)
75 Percentile Score	0.140**	0.155**	0.118**	0.119**	0.119**	0.409***
	(0.045)	(0.048)	(0.041)	(0.044)	(0.041)	(0.110)
Total Students	-0.032*	-0.035*	-0.030*	-0.031*	-0.029*	-0.068*
	(0.014)	(0.015)	(0.012)	(0.013)	(0.012)	(0.032)
% UG	0.015	0.016	0.013	0.014	0.012	0.036
	(0.013)	(0.014)	(0.012)	(0.013)	(0.012)	(0.034)
% Female	-0.028*	-0.028	-0.031*	-0.032*	-0.031*	0.024
	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)	(0.031)
% Minority	-0.005	-0.005	-0.004	-0.004	-0.004	-0.028
	(0.009)	(0.010)	(0.009)	(0.009)	(0.008)	(0.023)
Instr. Exp./Student	0.048**	0.054**	0.036*	0.042**	0.028	0.056
	(0.016)	(0.018)	(0.015)	(0.015)	(0.014)	(0.071)
Rsch. Exp./FT Fclty	0.000	0.000	0.000	0.000	0.001	0.006
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.009)
Public	-0.436	-0.460	-0.412	-0.423	-0.410	-0.668
	(0.371)	(0.400)	(0.343)	(0.360)	(0.338)	(0.779)
NRC 1-10 BA	0.318***	0.338***	0.297***	0.305***	0.295***	0.745***
	(0.080)	(0.086)	(0.073)	(0.076)	(0.073)	(0.210)
NRC 11-25 BA	-0.019	-0.019	-0.016	-0.015	-0.015	-0.005
	(0.077)	(0.082)	(0.071)	(0.073)	(0.071)	(0.191)
NRC 26-50 BA	0.051	0.054	0.047	0.052	0.044	0.051
	(0.045)	(0.049)	(0.042)	(0.044)	(0.041)	(0.107)
NRC 51+ BA	0.004	0.006	0.003	0.005	0.004	0.087
	(0.075)	(0.081)	(0.071)	(0.074)	(0.071)	(0.190)
% Humanities	-0.001	-0.003	0.005	0.004	0.006	-0.065
	(0.023)	(0.024)	(0.022)	(0.023)	(0.023)	(0.048)
PhD Prod. Rate Mean	2.293	2.465	2.131	2.214	2.109	4.132
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Missing Dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	3609	3609	3609	3609	3609	3609
Adjusted R-squared	0.46	0.45	0.48	0.46	0.50	0.53

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2: Negative Binomial Regression; PhDs w/ Hum. BA Only; NRC Only

	(1) # PhD simple 20	(2) # PhD simple 25	(3) # PhD lin 20	(4) # PhD lin 25	(5) # PhD quad 20	(6) # PhD quad 25
main						
Stud.-Facul. Ratio	0.004	0.004	0.004	0.004	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
75 Percentile Score	0.102***	0.102***	0.100***	0.100***	0.100***	0.100***
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
Total Students	0.003	0.003	0.003	0.003	0.003	0.002
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
% UG	0.011*	0.011*	0.011*	0.011*	0.011*	0.011*
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
% Female	0.001	0.001	0.001	0.001	0.001	0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
% Minority	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Instr. Exp./Student	0.010*	0.010*	0.010*	0.010*	0.010*	0.012
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Rsch. Exp./FT Fclty	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Public	-0.080	-0.075	-0.086	-0.082	-0.088	-0.063
	(0.159)	(0.159)	(0.160)	(0.160)	(0.159)	(0.156)
NRC 1-10 BA	0.147***	0.146***	0.147***	0.147***	0.147***	0.149***
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.026)
NRC 11-25 BA	0.065**	0.065**	0.065**	0.065**	0.065**	0.064**
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.024)
NRC 26-50 BA	0.088***	0.088***	0.088***	0.089***	0.086***	0.091***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)
NRC 51+ BA	0.059	0.062	0.058	0.061	0.057	0.069
	(0.037)	(0.037)	(0.038)	(0.037)	(0.038)	(0.039)
Hum BA	1.175***	1.186***	1.165***	1.179***	1.151***	1.278***
	(0.202)	(0.201)	(0.202)	(0.202)	(0.200)	(0.206)
PhD Count Mean	9.193	9.882	8.541	8.875	8.455	25.982
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Missing Dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	3640	3640	3640	3640	3640	3367
Pseudo R-squared	0.14	0.14	0.15	0.14	0.16	0.23

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3: Lemke Regression

	(1) Share Hum	(2) Share Hum	(3) Share Hum	(4) Share Hum
Stud.-Facul. Ratio	0.002	0.003	-0.001	-0.001
	(0.006)	(0.008)	(0.007)	(0.011)
75 Percentile Score	0.146***	0.150***	0.204***	0.212***
	(0.033)	(0.042)	(0.039)	(0.050)
Total Students	-0.032**	-0.033**	-0.038**	-0.041**
	(0.011)	(0.012)	(0.012)	(0.014)
% UG	0.012	0.016	0.013	0.016
	(0.010)	(0.012)	(0.012)	(0.015)
% Female	-0.017	-0.017	-0.021	-0.020
	(0.009)	(0.012)	(0.011)	(0.014)
% Minority	-0.000	-0.008	0.002	-0.008
	(0.004)	(0.009)	(0.005)	(0.011)
Instr. Exp./Student	0.058**	0.067**	0.062*	0.073*
	(0.021)	(0.025)	(0.026)	(0.031)
Rsch. Exp./FT Fclty	-0.001	0.000	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.003)
Public	-0.328	-0.307	-0.426	-0.401
	(0.286)	(0.330)	(0.329)	(0.385)
NRC 1-10 BA	0.339***	0.312***	0.384***	0.350***
	(0.061)	(0.069)	(0.076)	(0.088)
NRC 11-25 BA	0.006	-0.024	0.002	-0.036
	(0.063)	(0.075)	(0.075)	(0.091)
NRC 26-50 BA	0.085*	0.058	0.093*	0.054
	(0.034)	(0.041)	(0.039)	(0.048)
NRC 51+ BA	0.096*	0.025	0.119*	0.016
	(0.042)	(0.073)	(0.050)	(0.087)
% Humanities	-0.011	-0.004	-0.031	-0.029
	(0.015)	(0.022)	(0.017)	(0.027)
PhD Prod. Rate Mean	2.015	2.240	2.541	2.811
PhDs w/ Hum BA Only	Yes	Yes	No	No
NRC Only	No	Yes	No	Yes
Year Dummies	Yes	Yes	Yes	Yes
Missing Dummies	Yes	Yes	Yes	Yes
N	5557	3053	5654	3053
Adjusted R-squared	0.49	0.52	0.42	0.45

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$