NOTES

CHINESE GRADUATE STUDENTS AND U.S. SCIENTIFIC PRODUCTIVITY

Patrick Gaulé and Mario Piacentini*

Abstract—The migration of young Chinese scientists to undertake graduate studies in U.S. universities is arguably one of the most important recent episodes of skilled migration. Using a new data set covering around 16,000 Ph.D. graduates in 161 U.S. chemistry departments, we show that Chinese students have a scientific output during their thesis that is significantly higher than other students. In fact, conditional on acceptance into the same programs, Chinese students perform about as well as the awardees of the NSF doctoral fellowship program. These results shed new light on the benefits of student migration on scientific productivity of destination countries.

I. Introduction

IMMIGRANTS from China are a large fraction of science and engineering Ph.D. graduates educated in the United States. Of around 30,000 Ph.D. students graduating in 2006, more than 4,300 (14.3%) were Chinese citizen (NSF, 2009). Recent Ph.D. graduates from U.S. universities are more likely to have done their undergraduate studies at Tsinghua University or Peking University than at the University of California, Berkeley, or any other institution (Mervis, 2008). As of the 2000 census, 8.9% of doctorate holders in U.S. science and engineering occupations were born in China (NSF, 2007).

Using a new original data set covering around 16,000 Ph.D. graduates in 161 U.S. chemistry departments, we show that Chinese students have a scientific output during their thesis that is significantly higher than other students. In fact, conditional on acceptance into the same programs, Chinese students perform about as well as the awardees of the NSF doctoral fellowship program—America's best and brightest in science and engineering.

We relate the productivity differential between the Chinese and other students to a selection effect, as the majority of Chinese migrants come from a very restricted set of extremely selective Chinese universities. The relatively high quality of Chinese students migrating to the United States is consistent with the theoretical conclusion of the Roy model on a positive link between positive self-selection and migration costs (first shown in Borjas, 1987, and further discussed in Jasso & Rosenzweig, 1990). This paper is thus related to the extensive literature on the economic impacts of immigration (Borjas, 1999). This literature emphasizes the importance of the quality of immigrants for cost-benefit analysis, but it is generally unable to measure quality beyond a few limited proxies such as schooling levels.

Earlier research on immigration and scientific productivity has focused on the role of foreign-born star scientists in U.S. science and on the propensity of natives and migrants to publish during their career. Both Levin and Stephan (1999) and Hunt (2009) show that foreign-born U.S. scientists tend to outperform natives on a number of measures of scientific productivity.

Our results suggest that the immigration of Chinese students substantially expands the pool of talent available to the American scientific research entreprise. This advantage has, however, to be balanced against a number of possible counterarguments, including the fact that migration of foreign graduate students may crowd out native students (Borjas, 2004) or decrease incentives for natives to engage in scientific careers (Borjas, 2006; Freeman, 2009).

II. Data Construction

We identify Ph.D. students using Proquest Dissertations and Abstracts. This bibliographic database lists abstracts of completed Ph.D. theses with the name of the student, the university, and year of graduation, as well as the field and the name of the advisor. It also includes links to the full text of the theses, which is useful because theses from certain universities include additional bibliographical information on students (MacGarvie, 2007).

Proquest data provide good coverage of Ph.D. graduates for U.S. universities and recent years. However, we do not directly observe country of birth or of undergraduate education. To address this limitation, we use a technique similar to that pioneered by Kerr (2008a, 2008b) in his study of ethnic patent inventors. This technique relies on the fact that names implicitly contain information about the origin of individuals, and ethnicity can be reasonably well inferred by matching names to lists of ethnic names. Using the same approach, we constructed lists of Chinese last and first names and used them to code students as Chinese.

To verify the quality of the results obtained with our ethnic name matching algorithm, we manually coded CVs for universities that require students to report biographical information in theses. We find that 88% of students coded as Chinese had received their undergraduate degrees in China (and a further 5% in Taiwan). Conversely, our algorithm identified 96% of students that did their undergraduate studies in China as Chinese.

As a benchmark group to the Chinese, we also identify NSF graduate fellows in our data. The NSF graduate research fellowship is a highly prestigious award for American science and engineering students. Freeman, Tanwin, and Hanley (2005) refer to the program as "supporting the best and brightest in science and engineering."

We construct scientific output measures by matching our list of students to publication data from Scopus. This bibliographical database has the advantage of including affiliation data for each author. To minimize errors in the matching process, we used the fact that papers authored by chemistry students are almost invariably written with their advisor as coauthor. A publication is matched to a student if nine criteria are successfully met: one author of the publication needs to have the (a) last name of the student, (b) first initial of the student, (c) correct departmental affiliation of the student, (d) correct university affiliation of the student, and one of the coauthor of the paper has to have (e) the last name of the advisor, (f) the first initial of the advisor, (g) the correct departmental affiliation of the advisor, and (h) the correct university affiliation of the advisor. Finally, the paper has to be published (i) no earlier than three years prior to the graduation of the student and no later than the year of graduation.

Received for publication June 9, 2010. Revision accepted for publication January 5, 2012.

* Gaulé: CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Academy of Sciences of the Czech Republic, Prague, Czech Republic; Piacentini: OECD.

We thank Ajay Agrawal, Pierre Azoulay, Dominique Foray, Bill Kerr, Josh Lerner, Fiona Murray, Paula Stephan, Yanbo Wang, two referees and seminar participants at the NBER Summer Institute, NBER productivity lunch, the University of Louvain, the COST workshop on the economics of science (June 2009, Pisa), and the second international conference on migration and development (September 10–11, 2009, Washington DC) for insightful comments, discussions, and advice. Maria Eugenia Sagon and Cecilia Lazzarini provided excellent research assistance. Patrick Gaule acknowledges support from the Swiss science foundation and the Innovation Policy and the Economy Working Group, National Bureau of Economic Research.

TABLE 1.—DESCRIPTIVE STATISTICS ON SCIENTIFIC OUTPUT

	Chinese Students ($n = 2,380$)		NSF Fellows ($n = 336$)		All Other Students ($n = 13,357$)	
	Mean	SD	Mean	SD	Mean	SD
First-authored publications						
Number	1.06	1.44	1.33	1.52	0.80	1.16
JIF weighted	5.36	8.24	7.98	9.75	4.12	6.79
Cites to	33.39	131.41	50.81	153.31	22.81	76.77
All publications						
Number	1.99	2.37	2.32	2.36	1.49	1.91
JIF weighted	10.50	14.77	14.15	15.45	7.82	11.43
Cites to	73.44	268.32	92.88	202.51	46.77	134.03

TABLE 2.—REGRESSION RESULTS

	First-Authored Publications			All Publications		
	Number	JIF Weighted	Cites to	Number	JIF Weighted	Cites to
A. Specifications without School or Advisor Fixed Effects						
Chinese student	0.280***	0.259***	0.417***	0.283***	0.283***	0.481***
	(0.030)	(0.034)	(0.086)	(0.026)	(0.031)	(0.080)
NSF doctoral fellow	0.450***	0.597***	0.735***	0.373***	0.519***	0.612***
	(0.059)	(0.065)	(0.162)	(0.052)	(0.058)	(0.118)
B. School Fixed-Effects Specifications						
Chinese student	0.308***	0.322***	0.518***	0.300***	0.339***	0.583***
	(0.038)	(0.044)	(0.118)	(0.033)	(0.040)	(0.119)
NSF doctoral fellow	0.348***	0.284***	0.004	0.253***	0.156**	−0.172
	(0.050)	(0.059)	(0.204)	(0.055)	(0.078)	(0.166)
C. Advisor Fixed-Effects Specifications						
Chinese student	0.268***	0.270***	0.332***	0.235***	0.259***	0.354***
	(0.029)	(0.034)	(0.075)	(0.025)	(0.029)	(0.076)
NSF doctoral fellow	0.331***	0.310***	0.099	0.248***	0.198***	0.003
	(0.060)	(0.067)	(0.192)	(0.053)	(0.060)	(0.119)

Panel A specifications are estimated using a Poisson regression with robust standard errors. Panels B and C specifications are estimated by Poisson quasi-maximum likelihood. All specifications include year of graduation fixed effects, and subfield fixed effects. The various productivity measures used as dependent variable are based on the papers published between three years before graduation and the year of graduation. $N = 16{,}073$. Robust standard errors in parentheses. *$p < 0.1$, **$p < 0.05$, and ***$p < 0.01$.

III. Descriptive Statistics

We have 16,073 students, of whom 2,385 (14.8%) are identified as Chinese. The share of Chinese students does not exhibit much variation over time but considerable variation across universities. New York University, Southern Illinois University, and the University of New Mexico have more than 50% of Chinese students, while Berkeley, the University of Colorado, Arizona State, the University of Oregon, and the University of Texas at El Paso have less than 5%. The fraction of Chinese students is markedly higher in lower-tier schools (schools with lower chemistry R&D budgets). Chinese students represent 10.2% of students in tier 1 schools, 13.2% in tier 2 schools, 16.3% in tier 3 schools, and 20.6% in tier 4 schools.

Scientific output can be measured by counting the number of publications or adjusting for the quality of the journal (publications weighted by journal impact factors, hereafter referred to as JIF-weighted publications) or by counting the number of cites these publications receive. Moreover, in chemistry, and in the physical and life sciences more generally, first authorship has a special meaning and is typically used to recognize the junior scholar who made the largest contribution to the paper. Thus, it is interesting to distinguish between first-authored publications and all publications. Descriptive statistics by type of student for the six resulting output measures are displayed in table 1.

IV. Estimation and Results

We regress various measures of scientific output on a dummy for Chinese students and a dummy for NSF fellows. We estimate most regressions with a quasi-maximum likelihood conditional fixed-effects Poisson model (Hausman et al., 1984). This model has several desirable properties, including consistency of the coefficient estimates independent of any assumption on the conditional variance as long as the mean is correctly specified (Wooldridge, 1997) and consistency in the standard errors even if the data-generating process is not Poisson. This estimator can also be used for fractional and nonnegative variables (Santos Silva & Tenreyro, 2006), such as quality-adjusted first-authored publications in our case. We implement this in Stata with the "xtqmlp" procedure written by Tim Simcoe.[1]

We use three types of specifications. In the first type (table 2, panel A), we have as controls only year of graduation fixed effects and subfield fixed effects.[2] Next, we introduce school fixed effects and so compare students who are enrolled in the same program (table 2, panel B). Finally, we replace the school fixed effects by advisor fixed effects (table 2, panel C). An attractive feature of specifications with advisor fixed effects is that by comparing students who have the same advisor, we compare students who must be doing a very similar type of science.

The coefficient on Chinese student is positive and significant at 1% in all specifications. Comparing the Chinese student coefficient horizontally across output measures, we tend to have larger values for citations than for the number of publications or JIF-weighted publications. Since citations are usually thought to be a better proxy for the

[1] Available for download at http://people.bu.edu/tsimcoe/code/xtpqml.txt.

[2] These are subfields of chemistry (e.g., biochemistry, organic chemistry) as coded by Proquest Dissertations and Abstracts.

quality of research, this suggests that the productivity advantage of Chinese students is stronger in the quality dimension than in the quantity dimension. Comparing the Chinese student coefficient vertically across specifications, we find very similar coefficients. Thus, the productivity differential of Chinese students does not seem to be explained by school or advisor or team characteristics.

The Chinese student coefficient takes values between 0.259 and 0.583. Thus, Chinese students are between 22% (1-exp[−0.259] and 44% (1-exp[−0.583]) more productive than other, non-NSF fellow, students, depending on the specification. To interpret that magnitude, it is useful to note that this difference is about the same as that between NSF fellows and other students in the specifications with school fixed effects. In other words, conditional on being admitted in the same program, Chinese students seem to be doing as well as the recipients of the highly selective prestigious NSF fellowships.

V. Discussion

Why do the Chinese perform so well? Our preferred explanation is a selection effect. U.S. education enjoys an excellent reputation in China and attracts the brightest and most motivated Chinese students. Despite the fact that U.S. universities are admitting large numbers of Chinese students, it is nevertheless considerably more difficult for a Chinese than for a native to be accepted into a U.S. Ph.D. program. Evidence from Attieh and Attieh (1997) suggests that top U.S. universities give substantial preference to U.S. citizens in their admission decisions. While this may reflect an underlying preference for natives, it could also be an optimal response to difficulties encountered in evaluating the applications of Chinese students (lack of familiarity with schools, grading systems, and reference letter writers).

Precisely because of these difficulties, an undergraduate degree from one of the top Chinese university is a de facto requirement for entry into a U.S. Ph.D. program. Chinese graduate students overwhelmingly come from a set of extremely selective Chinese universities. Around 10 million high school finishers take the national college entrance exam, but only three thousand are admitted into the two most prestigious schools, Peking University and Tsinghua University, which are thus more selective than the most exclusive U.S. institutions. In fact, the majority of MIT undergraduates would not have had standardized test scores high enough to be admitted into the undergraduate programs of these two Universities.[3]

Besides this positive selection story, two other plausible arguments could explain the productivity advantage of the Chinese. The first argument points to incentive effects due to the higher preferences of the Chinese for an academic career. The second explanation emphasizes a cultural predisposition of the Chinese for higher effort at work.

According to the first argument, a career in academia and postdoctoral training in particular may be relatively more attractive to Chinese students, thus increasing the incentives to publish during the Ph.D. In particular, immigration considerations may be relevant, as maintaining valid visa status in the United States is easier when undertaking postdoctoral training than when working in industry.[4]. However, data from the 2001 Survey of Earned Doctorates indicate that Chinese Ph.D. students in chemistry are more likely to be planning to go to industry directly from graduate school than others (35.8% of those born in China plan to work in industry versus only 26.2% of those born anywhere else) and no more likely to be planning to follow postdoctoral training (37.4% of the Chinese plan to follow postdoctoral training, versus 39.7% of all the other students) (Stephan, 2010).

In support of the second argument, anecdotal evidence suggests that Chinese graduate students work harder and spend more time in the laboratory. Again, complementary evidence from other data sources suggests that this cultural predisposition for higher effort might be of second-order importance. In fact, a survey of postdocs found only small differences between Chinese and Americans in terms of hours worked (50.5 hours per week versus 49.8; Sigma Xi, 2005).

[3] The median math SAT score of MIT undergraduates is 770, which is lower than the top centile cutoff. Only 3% of Chinese entrance test takers scoring in the top centile are admitted into Peking University and Tsinghua University.

[4] From the perspective of immigration law, postdoctoral training is not considered as work. Most postdoctoral fellows are on visitor (J1) rather than on work (H1B) visas. The latter, but not the former, are subject to a yearly cap.

VI. Conclusion

The contribution of this paper is to show that Chinese students enrolled in U.S. Ph.D. programs have a strong record of publications during their graduate studies. We argue that this excellent performance of the Chinese with respect to other students is most likely explained by the fact that they come from a restricted number of very selective elite institutions in China. Given that the graduate student is the workhorse of the modern laboratory, the influx of talented students is bound to enhance the productivity of U.S. universities. This benefit to host universities has to be balanced against potential negative effects, in particular, the fact that migration may decrease incentives for natives to engage in scientific careers (Borjas, 2006).

Since the costs of training graduate students are mainly borne by the host country in the form of research and teaching assistantships, it is important to consider the stay rates of Chinese students. According to estimates derived by Finn (2007) using Social Security data, the stay rate for Chinese doctorate recipients is around 92% after five years from obtaining the title, the highest observed for any country in 2005. However, the current high stay rates cannot be taken for granted given the steadily raising skill premium in China and the aggressive recruiting policies of Chinese universities.

An important limitation of our study is that our research is based on an early measure of productivity—publication during doctoral studies—and does not address post-Ph.D. outcomes. We do not know if Chinese students continue to outperform non-Chinese students after their training period. On the one hand, initial differences in scientific productivity tend to persist and might even amplify over time. On the other hand, postgraduate occupations may require skills that the Chinese immigrants may lack. Further evidence on productivity and mobility choices after the Ph.D. is thus central to a better understanding of to what extent Chinese student migrants contribute to U.S. scientific productivity.

REFERENCES

Attiyeh, G., and R. Attiyeh, "Testing for Bias in Graduate School Admissions," *Jouranl of Human Resources* 32 (1997), 524–548.

Borjas, George, "Self-Selection and the Earnings of Immigrants," *American Economic Review* 77 (1987), 531–553.

——— "The Economic Analysis of Immigration," in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics* (New York: Elsevier, 1999).

——— "Do Foreign Students Crowd Out Native Students from Graduate Programs?" NBER working paper 10349 (2004).

——— "Native Internal Migration and the Labor Market Impact of Immigration," *Journal of Human Resources* 41 (2006), 221–258.

Finn, Michael G., "Stay Rates of Foreign Doctorate Recipients from U.S. Universities, 2005" (Oak Ridge, TN: Oak Ridge Institute for Science and Education, 2007).

Freeman, Richard B., Chang Tanwin, and Chiang Hanley, "Supporting 'The Best and Brightest' in Science and Engineering: NSF Graduating Research Fellowships," NBER working paper 11623 (2005).

——— "What Does Global Expansion of Higher Education Mean for the US?" NBER working paper 14962 (2009).

Hausman, Jerry, Bronwyn, Hall, and Zvi Griliches, "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52 (1984), 909–938.

Hunt, Jennifer, "Which Immigrants Are Most Innovative and Entrepreneurial? Distinctions by Entry Visa," NBER working paper 14920 (2009).

Jasso, Guillermina, and Mark R. Rosenzweig, "Self-Selection and the Earnings of Immigrants: Comment," *American Economic Review* 80 (1990), 298–304.

Kerr, William R., "Ethnic Scientific Communities and International Technology Diffusion," this REVIEW 90 (2008a), 518–537.

——— "The Ethnic Composition of US Inventors," HBS working paper 08-006 (2008b).

Levin, Sharon G., and Paula E. Stephan, "Are the Foreign Born a Source of Strength for U.S. Science?" *Science* 285 (1999), 1213–1214.

MacGarvie, Megan, "Using Published Dissertations to Identify Graduates' Countries of Origin," mimeograph, Boston University (2007).

Mervis, Jeffrey, "US Graduate Training: Top Ph.D. Feeder Schools Are Now Chinese," *Science* 321 (2008), 185.

NSF, "Why Did They Come to the United States? A Profile of Immigrant Scientists and Engineers," (Arlington, VA: National Science Foundation, Division of Science Resources Statistics, 2007).

——— "Science and Engineering Doctorate Awards: 2006" (Arlington, VA: National Science Foundation, Division of Science Resources Statistics, 2009).

Santos Silva, J. M. C., and Silvana, Tenreyro, "The Log of Gravity," this REVIEW 88 (2006), 641–658.

Sigma Xi, "The Sigma Xi Postdoc Survey" (2005), http://postdoc.sigmaxi.org/.

Stephan, Paula E., "Tabulations from Survey of Earned Doctorates," Discussion at NBER Summer Institute (2010).

Wooldridge, Jeffrey M., "Quasi-Likelihood Methods for Count Data," in H. Pesaran and P. Schmidt (eds.), *Handbook of Applied Econometrics* (Oxford: Blackwell, 1997).