I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13868

Overeducation, Major Mismatch, and Return to Higher Education Tiers: Evidence from Novel Data Source of a Major Online Recruitment Platform in China

Yanqiao Zheng
Xiaoqi Zhang
Yu Zhu

NOVEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13868

Overeducation, Major Mismatch, and Return to Higher Education Tiers: Evidence from Novel Data Source of a Major Online Recruitment Platform in China

Yanqiao Zheng
Zhejiang University of Finance and Economics

Xiaoqi Zhang
Southeast University

Yu Zhu
Nanjing University of Finance and Economics, University of Dundee School of Business and IZA

NOVEMBER 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Overeducation, Major Mismatch, and Return to Higher Education Tiers: Evidence from Novel Data Source of a Major Online Recruitment Platform in China*

We develop a novel approach to study overeducation by extracting pre-match information from online recruitment platforms using word segmentation and dictionary building techniques, which can offer significant advantages over traditional survey-based approaches in objectiveness, timeliness, sample sizes, area coverage and richness of controls. We apply this method to China, which has experienced a 10-fold expansion of its higher education sector over the last two decades. We find that about half of online job-seekers in China are two or more years overeducated, resulting in 5.1% pay penalty. However, the effect of overeducation on pay varies significantly by college quality, city type, and the match of college major with industry. Graduates in STEM (Science, Technology, Engineering and Mathematics) or LEM (Law, Economics and Management) from Key Universities are much less likely to be overeducated in the first place, and actually enjoy a significant pay premium even when they are in the situation.

JEL Classification: I23, I26

Keywords: overeducation, online recruitment data, major-industry mismatch, China

Corresponding author:
Yu Zhu
University of Dundee
School of Business
3 Perth Road
Dundee, DD1 4HN
United Kingdom

E-mail: yuzhu@dundee.ac.uk

* Xiaoqi Zhang thanks Ministry of Education Youth Program in Humanities and Social Science (grant number: 20YJC790176) for financial support. We thank Minsheng Weekly for allowing access to the data and participants of the 2019 Conference of Frontiers in Labor Economics in China and the 2018 Conference on China's Reform and Opening Up for comments. All errors are ours.

1. Introduction

There has been a large body of literature on the returns to education in the labor market. Recent studies suggest declining return to education (Sloane, 2003; Elias & Purcell, 2004; Walker & Zhu, 2008). On the supply side, there has been a steady increase of labor force with higher education (HE). For example, in China, over the decade from 1999, annual enrollment into HE increased by almost fivefold (Kang *et al*. 2019). Although the HE system has played a role in China's remarkable economic growth over the last 4 decades, the continuous supply of new graduates in excess of 7 million each year (Ministry of Education, 2019) in the past decade has put growing pressure on the graduate labor market. Indeed, there is abundance of anecdote evidence of growing qualification mismatch in the labor market, with frequent media reports of people with doctorate degrees teaching at primary or secondary schools and many college graduates unable to find jobs that match their majors. However, systematic evidence is still sparse on the matching of the supply with the demand side of the labor market. One way of bringing in the demand side is through the concept of overqualification (Green & Zhu, 2010).

Overqualification refers to a state of disequilibrium, whereby workers possess excess educational qualifications relative to those their jobs require. The concept was first proposed by Eckaus (1964), Berg (1970) and Freeman (1976), where the focus was on the difference of average education levels between the supply side and the demand side. Later research extends this concept to the individual level (Duncan & Hoffman, 1981; Robst, 1994; Quinn & Rubb, 2006; Green & Zhu, 2010). A general consensus is that overeducation leads to lower returns to education and lower job satisfaction, which implies waste in educational investment (Duncan & Hoffman, 1981; Robst, 1994; Cohn & Ng. 2000; Maynard *et al*., 2006; Green & Zhu, 2010).

Hitherto, there is hardly any studies of overqualification in China, despite a 10-fold expansion of the HE sector in China since 1999 producing more than one hundred million college graduates.[1] Behind the flourishing labor market and its enormous scale, however, it is uncertain whether this supply shock has resulted in mismatch and potential waste of education input.

[1] According to People's Daily, more than 8.34 million university students graduated people in summer 2019, up from 0.85 million back in 1999. According to statista.com, the number in 2019 was nearly double as high as the number of degrees earned at all levels of higher education in the United States. Moreover, around 604 thousand master's and doctor's degree students graduated from public colleges and universities in China in 2018. By 2018, more than 45 percent of the newly-added labor forces had a university diploma or degree (see https://www.statista.com/statistics/227272/number-of-university-graduates-in-china/).

In this paper, we focus on whether the boost in higher-educated labor supply is fully utilized by the demand side. In other words, the industry may not have developed or upgraded as fast as the growth of better-educated population, resulting in inefficiency in human capital utilization. As a result, we may see mis-allocation of labors, unemployment, and redundancy in over-pursuit of academic degrees. To better understand status quo and its possible future implication, it is important to study qualification mismatch in the labor market in China.

The lack of labor mismatch research in China is partly due to data availability limitations. We overcome this problem by using data from *zhaopin.com*, one of the major online labor recruitment platforms in China. To fully utilize the largely textual data, we resort to functions in Python, such as the dictionary-building implementation, and *Jieba*, the best Chinese word segmentation module so-far, to extract key words we need for the research. The main advantages of the data are as below. First, recruitment criteria are clearly listed on online recruitment platforms, therefore we can measure whether and to what extent the demand side and supply side are mismatched, with relative **objectiveness**. As a comparison, most previous research uses questionnaires and asks the job holders whether they feel they are overqualified or whether they would be offered the job if they applied for the job today, which is either subjective or under a hypothetical scenario, with possible systematic underestimation or overestimation by a certain group (Hartog, 2000). Second, data based on online recruitment platforms offers unrivalled **timeliness**, as it is almost updated on the go, therefore up-to-date and ready to fetch immediately. Third, compared to the survey-based approach, our approach to data collection can offer larger **sample sizes and greater area coverage**, at very low marginal costs, at least in principle. In fact, our sample data include job seekers from more than 200 cities, making cross-city heterogeneity analysis possible, which could not be done by survey data due to high costs. Fourth, online recruitment platforms are typically more advantageous in the **richness of controls**. For instance, few surveys would collect information on college major and quality (selectivity), industry, occupation, employment history, and (desired) salaries, all at once.

Therefore, our first contribution is in the development of a novel approach to study overeducation by extracting pre-match information from online recruitment platforms using word segmentation and dictionary building techniques via Python. Using a real example, we demonstrate significant advantages of this new approach over traditional survey-based approaches

in **objectiveness, timeliness, sample sizes, area coverage** and **richness of controls**. Moreover, this approach is easily adaptable to countries with similar online recruitment platforms.

Our second contribution is to the empirical literature on overeducation, in the context of the China which has experienced a 10-fold expansion of its HE sector over the last two decades. Due to lack to suitable survey data, there has been no study focusing on the wage or earnings effects of overeducation of graduates in China. This was unfortunate, given the importance of China and the totally unprecedented scale of the expansion of its HE sector one could explore. It turns out that online recruitment platforms offer a perfect opportunity to study overeducation of graduates, as people who receive more years of education are more likely to use the online channel for job seeking, and these people are more likely to suffer from overqualification. To this end, our data is more suitable than some other data that mainly focus on rural part of China. Moreover, with the wide area coverage provided by data from online recruitment platforms, we are even able to do heterogeneity analysis across cities, which could provide insights considering the different economic development status in the vast territory of China, but is probably not possible by using survey data due to high costs.

Although our data is about intended rather than realized job match, our results are remarkably consistent with the existing literature based on survey data, and also robust to alternative specifications of overeducation or the use of Inverse Probability Weighted Regression Adjustment method rather than OLS. We find that about half of online job-seekers in China are two or more years overeducated, resulting in 5.1% pay penalty. Graduates from Key Universities or people living in Tier 1 (Top 4 metropolises) cities are not only less likely to be overeducated, but also have lower pay penalties when they are. Having a subject-relevant degree reduces the probability of overeducation for the IT industry, but not for the Finance Industry. On the other hand, having a relevant degree carries a pay premium for IT but not for Finance Industry. Whereas the returns to years of schooling is two percentage points lower in Finance Industry than in IT Industry, being overeducated yields a pay premium in the former but has no effect in the latter industry. The different patterns in the college major-industry mismatch likely reflect the varying mix of industry-specific versus general human capital across college majors on the one hand, and the relative importance of signaling of credentials to employers by industries on the other. Finally, we find strong evidence of heterogeneity in both the incidence of overeducation and the effect of overeducation on pay: graduates in STEM (Science, Technology, Engineering and Mathematics)

or LEM (Law, Economics and Management) from Key Universities are much less likely to be overeducated in the first place, and actually enjoy a significant pay premium even when they are in the situation.

The remainder of the paper is organized as follows. Section 2 is the review of the relevant literature. Section 3 presents the data. Section 4 describes the data matching strategy. The empirical results and extensions are presented and discussed in Section 5 and 6 respectively. Finally, Section 7 concludes.

2. Literature review

Researchers have long studied the reasons for overeducation, and arrived at the following main explanations. First, there is the personal preference theory proposed by Lazear (1977) that features the differences in utility achieved from learning. Some people get satisfaction from seeking formal education itself without much consideration about its return, while some others do not. Second, there is the career promotion theory by Sicherman & Galor (1990) who propose that being employed at a position that requires less education might be a good opportunity for promotion. While both theories might apply to specific individuals, overeducation in China is more likely to be driven by policy-orientated shocks in both supply and demand side at large. Third, search friction theory (Albrecht and Vroman, 2002) that emphasizes mismatch caused by search cost. But the theory can hardly explain mismatch over a longer period.

There are, however, two other theories that might apply to China. The first one is the signaling theory (Spence, 1973) that advocates that people may over-invest in education, as a signal of their stronger ability, in order to obtain a (better) job, while education itself does not necessarily bring higher productivity. It is possible that graduates may use higher education credentials to send a signal of stronger ability to potential employers, with less consideration given to whether higher education is necessary or redundant. The second theory is the job competition theory (Thurow, 1975) which states that in economic downturns, well-educated people crowd out lower-educated people for some low-end jobs due to lack of job positions. These two theories align with the reality in China, because the majority of users of online recruitment platforms are born in 1980's and 1990's, a period of baby boom in China, and their college entrance year is after 1999 when the

expansion of higher education took place. Combined, this is an extraordinary period in Chinese history. As a result of over-supply, overqualification may occur.

The effect of overeducation on wage has been studied in many countries. Duncan and Hoffman (1981) use PSID data in 1976 and find that the return to education is 2.9% for years over-educated, as compared to 6.3% for years of education required for the job, which translates to a 3.4% pay penalty. Cohn & Ng (2000) find a 3%-11% pay penalty using Hong Kong census data in 1986 and 1991. Similarly results have been found using data from Germany, Mexico, Australia, and Sweden (Bauer, 2002; Quinn & Rubb, 2006; Green *et al.*, 2007; Korpi & Tahlin, 2009).

Due to data limitations, there are only a few studies on overeducation in China. Most of them are published in Chinese journals and highly descriptive in nature, focusing on the incidence of overeducation (e.g. Gao *et al*. 2017), One notable exception is Yin (2016), which is based on the 8 waves of the China Health and Nutrition Survey pooled over 1989-2009. Using various statistical measures, she finds that the incidence of overeducation is between 20 and 30 percent for the sample as a whole. The wage penalty to overeducation is substantial in OLS, but vanishes under the fixed-effect specification. However, college and university graduates only account for less than 10% of the sample. The large difference between OLS and FE results could be due to inability to control for college majors. Luo and Peng (2010) find the percentages of overeducation slightly exceed 50% with an increasing trend in year 2003, 2005, and 2006 using China General Society Survey. Wu and Lai (2010) and Wu and Wang (2018) find the proportion to be 44% and 33.67% using a survey data of Beijing and Kunming (capital city of Yunnan province), respectively. However, their data is limited to a single city. Therefore, our work is among the first to explore of the issue of overeducation in the nationwide graduate labor market in China.

Aside from overeducation, the issue of major mismatch is rarely touched in the literature. Robst (2008) extends the concept of educational mismatch to incorporate both the quantity and type of schooling into a measure of the distance between schooling and work, which he shows to be a key determinant of the wage effect of the educational mismatch. Using the 1993 National Survey of College Graduates, Robst (2007) shows that around 45% of US graduates have a job which is only partially related or unrelated to their college major. Moreover, he finds that graduates from majors that emphasize general skills have a higher likelihood of mismatch, but suffer relatively lower pay penalties for being mismatched. Using the Annual Population Survey from

2006-2017, a report by the UK Office for National Statistics find that 31% of UK graduates are over-educated (ONS 2019). Furthermore, graduates in Science, Technology, Engineering and Mathematics (STEM) subjects are not only less likely to be overeducated, but also suffer lower pay penalties even when conditional on being overeducated.

3. Data

We use resume data from *zhaopin.com*, one of the largest online recruitment websites in China. *zhaopin.com* was established in 1994, and its business covers the vast majority of cities in China. Its resume data contains variables such as age, gender, education experience, work experience, job intention (including desired salary), work status (working or out of work), self-view, residence city, work place, and *hukou* affiliation. To fully utilize the largely textual data, we resort to functions in Python, such as *Jieba*, the best Chinese word segmentation module to date, to extract key words we need for the research, and the dictionary-building implementation, to build the dictionary of education requirements by the demand side. For this study, we choose a random sample that comprises job seekers of working age. Specifically, the sample consists of men between age 18 and 65, and women between age 18 and 60, looking for a full-time job, and that relevant variables are not missing. We randomly selected 20,000 resumes in June 2017, with 17,810 resumes remained after cleaning, of which 8371 are men and 9439 are women. Unlike LinkedIn, resume data from *zhaopin.com* can only be seen by potential employers, and is *not* available to the public.[2] Therefore, the platform of *zhaopin.com* serves merely as a job market instead of social media or social network. This guarantees that job seekers produce their resumes without concerns about signaling undesirable traits (e.g., ambition) to friends or acquaintances or potential dating mates. In this way, our work may add to the literature by observing job seekers in a secure and private setting, compared to resume platform with social network function.

Data on the demand side is also from *zhaopin*.com, which has millions of recruitment listings. Each listing contains job description (including job responsibilities, requirements on education/ experience/ skill, salary offers and benefits), company introduction (including company name, firm size, address, associated industry) and other supplemental information. For this study, we

[2] The data is officially acquired from a database initiated by Minsheng Weekly, which holds a random subsample of the resume database of *zhaopin.com*. The data is accessible for research purpose, but one has to apply for permission via their official website http://www.cnbo.tv or http://www.msweekydata.com, or email address cnbotv@163.com.

choose a random sample containing 16,000 pieces of recruitment posts published between June and August 2017, with 15,901 pieces remain after cleaning.

One might doubt the representativeness of our data. To tackle this, we compare in Table 1 the distribution of 4 key variables of our sample to the 2016 China Family Panel Studies (CFPS_2016) that are designed to be nationally representative. It is worth emphasizing that we focus on the better educated population, consistent with population selection in previous literature on overeducation (Green and Zhu, 2010; Wu and Wang, 2018, etc.). In this respect, survey data like CFPS has a well-educated subsample size that are quite small (1,355) compared to the size of a typical sample from a recruitment platform (17,810).

1) *Age distribution*. In the CFPS_2016 data, people with high school education or higher are typically around 30 years old, while their poorer-educated counterparts are well above 40 years old. The average ages of the subgroups with high school education or higher coincide well with our sample data from *zhaopin.com*, which in a way proves the representativeness of our data, for the subgroups with high school education or higher.
2) *Education distribution*. The education distribution of *zhaopin.com* is similar to the well-educated subsample of CFPS. Specifically, the relative percentage between subpopulations with some college, bachelor, master, and doctorate degrees from the CFPS data (51.60%: 42.26%: 5.61%: 0.53%) are similar to that in our *zhaopin.com* sample (40.55%: 48.07%: 1.02%: 0.03%).
3) *Monthly income distribution*. The reported income levels in CFPS are much lower than the average proposed salaries from *zhaopin.com*. However, it is frequently documented that survey data like CFPS suffer from non-response and underreporting of income, especially by top income people (Gustafsson *et al*., 2014; Zhang & Zhao, 2019; Li *et al*., 2020). Moreover, the non-response rate exceeds 77.11% regarding monthly income in the CFPS sample, making its income data highly biased. As such, our data sample has an additional advantage in revealing income levels closer to the truth, without concerns of underreporting or non-response.
4) *Occupation distribution*. Most (>99%) of the observations in CFPS have missing values for occupation. In other words, CFPS and similar survey datasets may not server as a good source to study overeducation issues in the labor market. This is indeed one of the reasons why there is lack of relevant research in China. In this sense, our sample offers a perfect source with its unrivalled richness of controls, and yet most importantly, without loss of representativeness.

Table 1: Comparison between CFPS and our data

	CFPS full sample		CFPS subsample with monthly income reported			CFPS subsample with high school+ education	Our sample from *zhaopin.com*		
Education qualifications	**Percent**	**Average age**	**Percent**	**Average age**	**Average income**	**Percent**	**Percent**	**Average age**	**Average income**
Jr High School or below	75.47%	48.96	75.47%	44.82	2938.83		0.71%	31.13	8616.35
Sr High School	14.34%	39.63	16.34 %	33.32	3155.13		9.63%	31.85	8354.73
Some College	5.99 %	33.96	8.73%	30.61	3716.04	51.60%	40.55%	31.20	9361.64
Bachelor	3.78%	33.84	7.15%	29.95	4756.41	42.26%	48.07%	30.92	12192.38
Master	0.39%	32.91	0.95%	31.37	5982.18	5.61%	1.02%	33.17	16426.47
Doctor	0.03%	31.44	0.09 %	32.43	5266.67	0.53%	0.03%	40.20	30500.00
Total	100%	45.96	100%	40.03	3343.83	100%	100%	31.17	10639.85
N	34,992		8,009			1,355	17,810		

Note: Average income for the CFPS whole sample is not calculated because 77.11% of the population did not report income, which could cause substantial bias. Income unit in yuan/month.

4. Matching strategy

Regarding qualification match, there are three methods commonly used in the literature. The first method is to ask employees about education backgrounds required for the job (Duncan & Hoffman, 1981; Hartog & Oosterbeek, 1988; Galasi, 2008). But there are subjective biases in this approach. In addition, it is also sensitive to the choice of words and sentences, for example, "what kind of education is needed to get your current job" and "what kind of education is needed to be competent for your current job" will lead to different answers (Leuven & Oosterbeek, 2011). The second method is to use the dictionary of occupational requirements in the general education development (GED) (Eckaus, 1964). However, the disadvantage of this method is the low updating frequency of the dictionary and the high updating cost, and the fixed education level required by each occupation, so the flexibility of using this method is low (Hartog, 2000). The third method is based on the average education level of the existing labor force in each occupation (Verdugo & Verdugo, 1989; Kiker *et al.*, 1997). The problem with this statistical method is that it reflects the result after the labor market has completed matching rather than the requirements of employers, so it is considered to be weaker than the first two methods (Leuven & Oosterbeek, 2011). This paper uses the real-time recruitment post database of *zhaopin*.com, which can effectively avoid subjective bias, mis-diction and outdatedness, and can reflect the needs of employers, so it effectively overcomes the limitations of the above three methods.

To match the datasets from the supply side and the demand side, our first step is to create a dictionary for the requirements of the demand side. The Python Dictionary object provides a **key:value** indexing facility where the values in the dictionary are indexed by keys. To create a dictionary, the crux is to establish **keys**. In this study, we use both industry and city as the **keys** for the dictionary of the labor demand side. In terms of specific operation, we extract the numerical information of the recruiter, such as the requirements of years of education, take the average value of the requirements over each combination of industry and city, and then compile a dictionary with requirements indexed by industry and city. That is, the **key:value** pair here refers to **(industry, city): requirements**. Our method is similar to the idea behind the dictionary of General Educational Development (GED) (Eckaus, 1964), but more time-effective and considers regional differences.

Once the dictionary is built, the second step is to assign each job seeker to a specific requirement value. Job seekers often search for more than one industry in more than one city. In fact, they often go through multiple rounds of interviews with multiple companies and receive multiple job offers, and make the final choice based on multiple factors such as benefits, prospects, distance from home and so on. Due to the existence of multiple intentions, strict matching is not possible. We handle this by taking the average of the multiple requirements corresponding to the multiple entries of the intentions. After the assignment, each job seeker gets a value for the required years of education. Then we can get the number of years over-educated by subtracting years required from her actual years of education. See figure 1 for the flowchart of the matching process.

While we do not observe the realization of the job match, we capture the pre-match status, which may well predict the match results, to the extent that both the supply and demand sides reveal their true preferences. It is worth emphasizing that *zhaopin.com* is the most efficient online job seeking platform in China, in terms of the response and interview rate and customer service in a highly competitive market.[3]

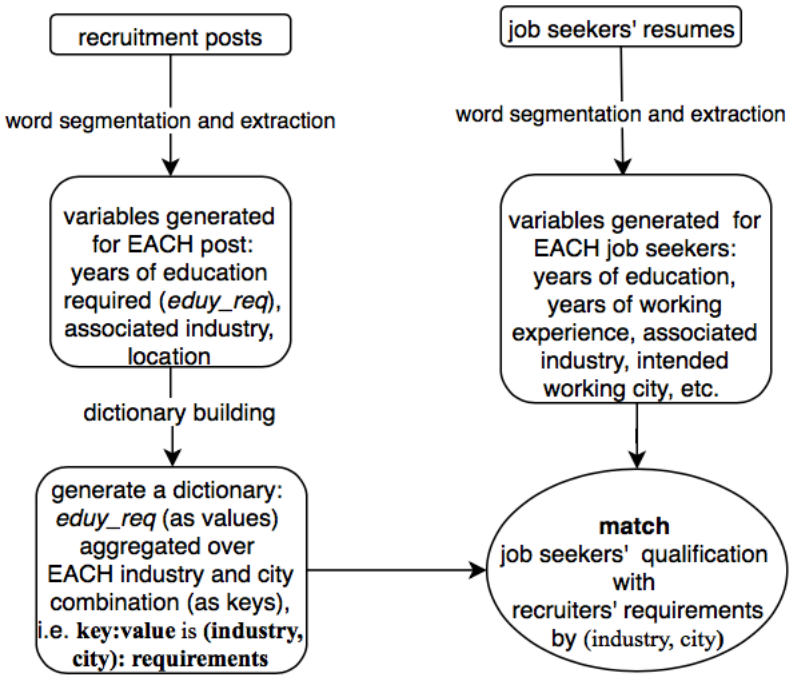

Figure 1: Flowchart of the Matching Process

[3] Recruiting firms need to pay about 20 yuan to view each resume. Although job seekers do not pay to use the platform, unserious job seekers are routinely removed by the AI system, to ensure that the platform keeps attracting recruiters.

5. Empirical Results

Table 2 shows the mismatch of education between supply side and demand side. Nearly half of job seekers hold a Bachelor's degree, while 40% hold some vocational college degree. This means that college graduates account for almost 90% of jobseekers on the platform.[4] In comparison, only about 20% of recruiting posts require a college degree require a bachelor's degree, while another 40% require some college. Moreover, over one third of vacancies requires no degree. This shows that overeducation is prevalent among online job-seeking platforms.

Table 2: Comparison of education supply and demand

Education qualifications	Supply (Job-seekers)	Demand (Recruiting firms)
Junior High School	0.71%	-
High School	9.63%	7.54%
Some College	40.55%	41.83%
Bachelor	48.07%	19.23%
Master	1.02%	0.67%
Doctor	0.03%	0.02%
Unspecified	-	30.72%
Total	100%	100%
N	17,810	15,901

Furthermore, we document education mismatch at the individual level. As shown in Figure 2, the peak corresponds to 2-2.5 years of over-education. Most of the population locate on the right side of zero years, meaning most applicants are over-educated rather than under-educated. Besides, the majority of job seekers in the sample fall in the range between 2.5 years under-educated and 5 years over-educated.

[4] In China, high school graduates may take the college entrance examination. Depending on their performances in the exam, they will be assigned to universities which last 4–5 years, leading to a Bachelor's degree, or vocational colleges which last 3 years, leading to a college diploma. The former is further divided into Key or Ordinary Universities.

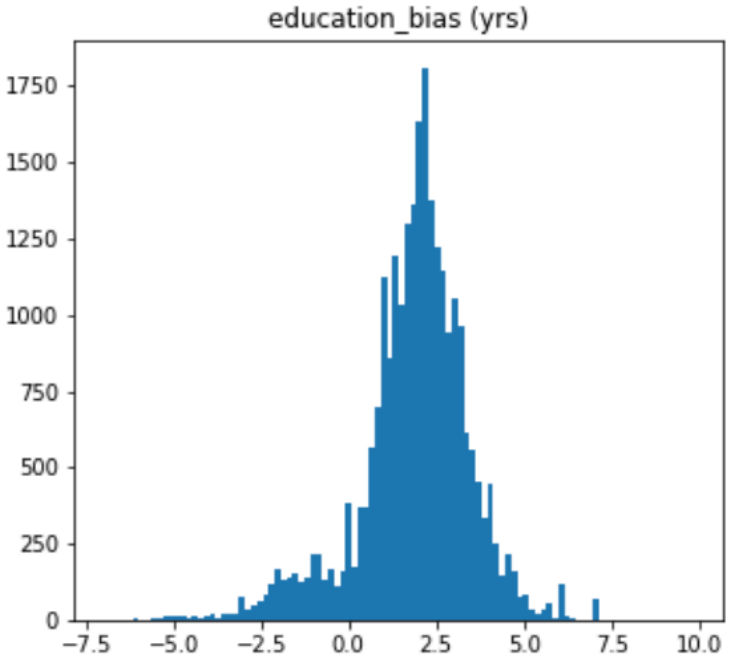

Figure 2: Distribution of Excess Years of Schooling

Table 3 shows the mismatch of industries between labor supply and labor demand. As can be seen, 43% of job seekers list IT industry as their intended industry while only 16% of recruitment posts belong to IT industry. Another observation is that on the supply side, job seekers' intentions are largely concentrated in a few industries, where intentions for 5 industries alone account for more than 90%. In comparison, the recruiting part as the demand side exhibits more diversified needs, where posts from the top 5 industries add to just under 50%.

Table 3: Comparison of intended industries ranking and recruiting industries ranking

Industry	Supply (Job seeker's intention)	Demand (Recruiting firm)
IT	42.65%	16.34%
Education	12.58%	10.20%
Real Estate/Construction	16.79%	9.48%
Finance	11.46%	4.48%
Consulting	9.91%	3.25%
Others	6.61%	56.25%
Total	100.00%	100.00%
N	17,810	15,901

We proceed to examine the cross tabulation of education mismatch and major mismatch. Since it is relatively vague as to what degree majors are needed in the education, real estate/construction and consulting industries, we focus on the IT and the Finance Industry in Table 4A and Table 4B, respectively. We define overeducation as workers obtaining 2 or more years of education relative to what their jobs require. There are several reasons why we choose 2 years as the threshold.[5]

First, regarding the supply side, the gaps between closest education levels in the education system range from 1 year (e.g. between some college and bachelor) to 3 or more years (e.g. between high school and some college). Then 2 years would seem to be a reasonable middle-point.

Second, regarding the demand side, industries could be indifferent between applicants with, say some college and bachelor degrees. Therefore, making the threshold 1 year would be meaningless. Presumably, industries may NOT be so indifferent between a bachelor and a master's degree (2 years gap), thus making the threshold 3 years would be too large to capture the overeducation of a master degree holder facing the company's demand of a bachelor degree.

Third, we calculate a worker's years of overeducation as the average years of overeducation of his/her multiple intentions in job seeking, which is essentially the years of education one have achieved minus the years of education required by the employer. This definition is consistent with literature (e.g., Green & Zhu, 2010). Note that using the average of multiple job intentions as the required education years is embedded in our dictionary-building process, i.e., building a dictionary indexed by city/industry with values being the years of education required. Admittedly, matching by industry and city could be less than perfect, compared to variables like the rank in work which is not available. In fact, the closest information to the work rank that job posts may have is just job titles and responsibility requirements, none of which appear to be standardized.

Finally, since this is a novel dataset, there exists no previous definition that can be readily applied to our data. This is a limitation with our definition. However, we have tried our best to find the most reasonable definition consistent with previous literature.

[5] Alternative definitions for overeducation would change the magnitudes of the results but not the significance.

Table 4A: Overeducation by Major-Mismatch, IT Industry

	Overeducated	Not Overeducated
Relevant academic degree	7.94%	14.39%
No relevant academic degree	37.02%	40.65%

Table 4B: Overeducation by Major-Mismatch, Finance Industry

	Overeducated	Not Overeducated
Relevant academic degree	38.22%	28.41%
No relevant academic degree	17.91%	15.46%

Table 4A shows that only 22% of job applicants for the IT industry hold relevant degrees while 45% are overeducated. Conditional on holding a relevant academic degree, there are fewer people with overeducation than without overeducation. As for those without relevant academic degree, the proportion of people with overeducation is similar to the proportion of those without overeducation. Overall, the odds ratio of overeducation to non-overeducation is much lower when the job applicant for an IT job holds a relevant rather than a non-relevant degree.

In contrast, Table 4B shows that two-thirds of job seekers in the Finance Industry have relevant academic degrees while 56% are overeducated. Moreover, the odds ratio of overeducation to non-overeducation is much higher when the job applicant holds a relevant rather than a non-relevant degree.

Table 5: Descriptive statistics of job seekers

Var. Name	Definition	Mean	Std. Dev.	Min	Max
Salary	Desired Salary (Yuan/Month)	10639.850	8180.561	3000	42500
Age	Age	31.169	6.096	18	63
Female	Female	0.514	0.500	0	1
Education:					
Eduy	Years of Education	15.145	1.314	9	21
Grad985	Project 985 University graduate	0.125	0.331	0	1
Grad211	Project 211 University but not Project 985 University graduate	0.115	0.319	0	1
Overedu	Overeducation dummy (Over-Educated≥2 Years)	0.500	0.497	0	1
Work Experience:					
Work_Exp	Years of work experience	10.033	6.283	0	46
Job_History#	Counts of previous jobs	3.255	1.712	1	20
Job-seeking Status:					
Leave	Left the Last Job and available immediately	0.555	0.497	0	1
Working	Still with current job	0.348	0.476	0	1
Satisfied	Satisfied with current job but looking for better	0.061	0.240	0	1
Graduate	Fresh graduate	0.027	0.163	0	1
Variables for Treatment Model:					
Self_esteem	Self-esteem	2.321	2.032	0	13
Self_efficacy	Self-efficacy	2.327	1.896	0	11
Description_length	Word count of self-description	140.322	102.842	1	822
Cert_num	Number of certifications	1.071	1.758	0	25
N		17,810			

Next, we conduct summary data analysis. We apply 99% winsorization to the salary to remove outliers. As shown in Table 5, the average age of job seekers on *zhaopin.com* is about 31. More than half of the population have left the last job and are available immediately, and the majority of the rest are being employed but looking for better opportunities. Women account for a slightly larger proportion (53%) than men (47%). An average job seeker on the platform has been in the labor force for 10 years and has held about 3 job positions in the past. On average, the sample

population received 15.145 years of education, with 1.830 years being over-educated relative to job requirements. We define overeducation as workers obtaining at least 2 years more education relative to those their jobs require. 50% of the sample population are overeducated. A quarter of the sample population obtained degrees from Key universities, of which half from National Project 985 universities, and another half from National Project 211 universities.[6]

Note that our sample is based on an online recruitment website, which is by no means similar to census population, because our sample population is younger, better-educated, and more often resort to online channel for job seeking activities. It comes as no surprise that their desired salaries are around 10,000 yuan/month, much higher than the national average of ordinary people. Because our key focus is on overeducation in the relatively higher-end labor market, we deem it most appropriate to study these well-educated population, where there is a higher probability that overeducation occurs, instead of studying the general population.

The dependent variable is *Ln(salary)*, the logarithm of salaries in yuan per month. The key regressor is *overedu*. There are different measures of overeducation in literature (Hartog, 2000), as already discussed in the introduction. Since our dataset is a novel dataset, there exists no previous measure that can be readily applied to our data. We match the supply side and the demand side, and define overeducation as the situation when a job seeker's years of education is 2 or more than the years of education required by the job, which is more objective than most definitions in literature. The justification of 2 years as the threshold is already given earlier in this section. The results with alternative definitions by using years of overeducation are presented in Table A2.

Based on previous literature (Quinn & Rubb, 2006; Green & Zhu, 2010, etc.), the control variables include *Female* (dummy variable indicating whether female or not), Eduy (years of education), *Work_Exp* (years of work experience) and its squared term (*Work_Exp^2*), *Job_History#* (number of previous jobs), *Grad985* (graduates from National 985 Universities), *Grad211* (graduates from National 211 Universities, but not National 985 Universities), *Relevant Degree* (whether hold an industry-specific academic degree, statistics given in Table 4A and 4B), and the interaction of *Overedu* and *Relevant Degree*. Moreover, to take advantage of the richness

[6] The Chinese government ranks domestic universities and classifies them as "Project 985 Universities" and "Project 211 Universities". As of 2018, there are 39 universities listed in "Project 985 Universities", as the first-tier, and 112 universities listed in "Project 211 Universities", as the second-tier.

of the information provided by our data, we control for enterprise types, marriage status, job seeking status, industry categories, city tiers, all as dummy variables, and finally, a constant term.[7] The summary statistics in given in Table 5.

For possible concerns of endogeneity of the key regressor, we introduce new variables to the treatment model of IPWRA and PSM method. The added variables in the treatment model are *Self_esteem, Self_efficacy, Description_length*, and *Cert_num*. The first three stand for self-esteem, self-efficacy, word counts of self-description in the resume. They are self-evaluation related variables widely adopted in psychology literature (Rosenberg, 1979; Capraro & Sippel, 2017; Zhang & Zheng, 2019). The fourth variable *Cert_num* stands for the number of certifications one lists in the resume. The four variables serve as instruments for unobservable abilities. The summary statistics of these variables are shown in Table 5. More justification and explanation will be given in Section 6.1 which outlines the IPWRA and PSM methods.

Table 6 shows the effect of overeducation on log salary. In the full sample, after controlling for years of education, etc., the dummy variable overeducation has a negative coefficient, meaning overeducation causes 5.1% less salary than that if the same job seeker was put in a position in which the education requirement matches his/her own level. Column (2) – (4) show the results for the IT industry. Interestingly, as compared to the full sample, the coefficient of the penalty by overeducation is not significant, whereas holding a relevant academic degree has a significantly positive effect in the IT industry. As for the Finance Industry, as shown in column (5)-(7), there is indeed a pay premium brought by overeducation, while there is no premium brought by relevant academic degree. Column (8) and (9) show results for the education industry and the real estate industry respectively, where the effect of overeducation is significantly negative, similar to that of the full sample. Therefore, column (8) and (9) can serve as reference groups.

Despite the theoretical interest and policy relevance, there has been few studies on the effect of industry-specific human capital, as opposed to general human capital. Notable exceptions include Carrington (1993) and Kletzer (1996), both of which highlight the importance of industry-

[7] The variable *age* is not controlled, because we define *work_experience* = *age* – *education_years* – 6, consistent with the existing literature. With a single cross-section, when the variables *work_experience* and *education_years* are already controlled, further including *age* as a control variable will result in multicollinearity. The implicit assumption is that people directly join the labor force after finishing schooling, and there are no gap years between jobs. There can be exceptions to the assumption. However, the results won't alter if we use variables (age and education_years) instead of variable (education_years and work_experience).

specific human capital using surveys of displaced workers. Our data offers a rare opportunity to shed new light on this important issue in a developing country context.

The contrast between the IT and the Finance industries is quite striking. First, in the Finance Industry, having a relevant major does NOT lead to higher salary. In other words, major mismatch is not a problem, and major is not what the finance industry values. Instead, education is what this industry values in the sense that employees who possess excess educational qualifications relative to those their jobs require are expected to get paid in accordance with their educational achievements, or even more (as shown by the non-negative coefficients of overeducation). In total, overeducation might lead to pay premium, and major mismatch does not mean pay penalty in the Finance Industry.

Second, in the IT industry, having a relevant major DOES lead to higher salary. In other words, major mismatch is associated with pay penalty, meaning the IT industry values relevant majors. On the other hand, overeducation has neither positive nor negative effect on pay level, i.e., employees who possess excess educational qualifications relative to those their jobs require are expected to get paid in accordance with their educational achievements. In total, overeducation is not associated with pay premium or pay penalty, and major mismatch DOES mean pay penalty in the IT industry. This finding is consistent with Kim *et al*. (2014) who find higher returns to industry-specific human capital than general or firm-specific human capital for IT-enabled business process outsourcing industry professionals in India, especially for junior-level professionals whose jobs are relatively more standardized.

Note that the results in this table does not explicitly account for the small minority of the undereducated. The results that includes the undereducated group are presented in Tables A1 in the Appendix. An alternative specification which is popular in overeducation studies using years of over-education and years of required education are presented in Table A2. It is clear that our results are robust to both variations in model specification.

Consistent with most existing studies, Table A2 shows that the effect of overeducated years is smaller than that for the required years, even for the Finance Industry. This is robust with or without allowing for undereducation. It is also worth noting that the returns to overeducated years are almost the same between Finance Industry and IT industry, while the returns to required years are significantly higher for the IT industry.

Table 6: OLS regression results

	All	IT Industry			Finance Industry			Education	Real Estate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	（8）	（9）
Overeducation	-0.051***	-0.011	-0.004	-0.003	0.076***	0.076***	0.051	-0.064***	-0.046**
	(0.008)	(0.013)	(0.013)	(0.015)	(0.026)	(0.026)	(0.039)	(0.025)	(0.022)
Relevant Degree			0.086***	0.089***		-0.006	-0.026		
			(0.014)	(0.018)		(0.023)	(0.035)		
Overedu X Relevant Degree				-0.006			0.036		
				(0.029)			(0.046)		
Eduy	0.134***	0.135***	0.135***	0.135***	0.110***	0.110***	0.111***	0.136***	0.135***
	(0.004)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)	(0.012)	(0.011)
Grad985	0.132***	0.117***	0.116***	0.116***	0.155***	0.156***	0.158***	0.144***	0.095***
	(0.012)	(0.017)	(0.017)	(0.017)	(0.029)	(0.029)	(0.029)	(0.033)	(0.029)
Grad211	0.096***	0.072***	0.074***	0.074***	0.093***	0.093***	0.094***	0.034	0.100***
	(0.012)	(0.018)	(0.017)	(0.017)	(0.032)	(0.033)	(0.033)	(0.033)	(0.028)
Work_Exp	0.056***	0.069***	0.069***	0.069***	0.065***	0.065***	0.065***	0.047***	0.056***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.006)	(0.005)
Work_Exp^2	-0.001***	-0.002***	-0.002***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Job_History#	0.037***	0.041***	0.042***	0.042***	0.039***	0.039***	0.040***	0.034***	0.042***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.009)	(0.009)	(0.009)	(0.008)	(0.007)
Female	-0.217***	-0.227***	-0.210***	-0.210***	-0.183***	-0.182***	-0.181***	-0.258***	-0.191***
	(0.007)	(0.012)	(0.012)	(0.012)	(0.023)	(0.023)	(0.023)	(0.024)	(0.019)
N	16,517	5,769	5,769	5,769	1,777	1,777	1,777	2,006	2,553
Adj. R-Sq	0.498	0.501	0.504	0.504	0.511	0.510	0.510	0.494	0.488

Note: the dependent variable is Ln(salary). All regressions control for enterprise types, marriage status, job seeking status, industry categories, city tiers, and a constant term. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%. The results in this table does not include the small minority of the undereducated. The results that includes the undereducated group are presented in Tables A1 in the Appendix. Alternative specifications using years of over-education and years of required education are presented in Table A2.

6. Extensions

6.1. IPWRA and PSM method

The problem with OLS estimation is that salary will, to some extent, reflect a job participant's ability, and high ability person may or may not be overeducated depending on the compatibility of her strength with academic requirements during schooling, the information of which we do not have. An individual can choose his/her education level, making overeducation self-selected in a sense. In an ideal world, we would design an experiment to test cause-and-effect and treatment-and-outcome relationships. We would randomly assign subjects to the treated or untreated groups, such that the treatment is independent of the outcome. To deal with the endogeneity, we apply the IPWRA (Inverse Probability Weighted Regression Adjustment) method, which could well solve the endogeneity problem brought by self-selection (Wooldridge 2007; Cattaneo, 2010; Walker & Zhu, 2018). The method of IPWRA is a combination of two methods: IPW (inverse probability weighting) and RA (regression adjustment).

In the RA method, the counterfactual outcomes, or the unobserved potential outcomes, are estimated, by fitting separate linear regression models with the observed data to the two treatment groups (overeducated and non-overeducated). Since overeducation is no longer a regressor in the regression, endogeneity issue is partly alleviated. More specifically, we have one regression line for the overeducated and a separate regression line for the non-overeducated. The difference between the expectations on the two lines is the estimate of the covariate-specific treatment effect for each subject in the data. Averages of these effects over all the subjects estimate the ATE (average treatment effect).

In the IPW method, the treatment assignment process is modelled, compared to the outcome process being modelled in the RA method. By using a probit (or logit) model to fit the treatment assignment, the predication Pr(overeducated) is obtained for each observation in the data, denoted as p_i. Then weight observations on the overeducated by $1/p_i$, so that weights will be large when the probability of being overeducated is small. Similarly, weight observations on the non-overeducated by $1/(1-p_i)$, so that weights will be large when the probability of being non-overeducated is small. By doing so, we will achieve two more balanced subsample (overeducated vs. non-overeducated) in the sense that given a set of covariates, the probability that there exist comparable pairs of both the treated and untreated, is higher.

The IPWRA method has both the advantages of PA method and IPW method, and solves the self-selection problem. Besides, the IPWRA estimator has the double-robust property, meaning that the estimates of the effects will be consistent if either the treatment model or the outcome model, but not both, are mis-specified (Wooldridge 2007; Cattaneo, 2010; Walker & Zhu, 2018).

Column (1) and (2) in Table 7 display the RA coefficients for the treated (overeducated) and untreated (non-overeducated) groups, respectively. Column (3) displays the coefficients for the probit treatment model. The added variables in the treatment model are *Self_esteem, Self_efficacy, Description_length*, and *Cert_num*. The first three stand for self-esteem, self-efficacy, word counts of self-description in the resume. They are self-evaluation related variables widely adopted in psychology literature (Rosenberg, 1979; Capraro & Sippel, 2017; Zhang & Zheng, 2019). The fourth variable *Cert_num* stands for the number of certifications one lists in the resume. The summary statistics of these variables are shown in Table 5. We use the four added variables as instruments to deal with the possible endogeneity of overeducation. Though it is hard to guarantee the absolute exogeneity of the instruments, as with many other instruments, we believe the adoption of these instruments in this context alleviate the endogeneity problem.

Table 7: IPRWA results

	RA coefficients (Dep. Var: Ln(salary)		Probit (Dep. Var: Overedu)
	Treated (overeducated)	**Untreated (non-overeducated)**	**Treatment model**
ATE	-0.084***		-
	(0.009)		-
Female	-0.205***	-0.236***	0.076***
	(0.010)	(0.010)	(0.020)
Eduy	0.187***	0.127***	
	(0.010)	(0.004)	
Grad985	0.134***	0.139***	-0.132***
	(0.017)	(0.016)	(0.030)
Grad211	0.100***	0.083***	-0.101***
	(0.016)	(0.016)	(0.031)
Work_Exp	0.055***	0.054***	
	(0.003)	(0.003)	
Work_Exp^2	-0.001***	-0.001***	
	(0.000)	(0.000)	
Job_History#	0.031***	0.044***	
	(0.004)	(0.004)	
Self_esteem			-0.019***
			(0.005)
Self_efficacy			0.004
			(0.006)
Description_length			-0.000
			(0.000)
Cert_num			0.025***
			(0.006)
N			16,517

Note: Control variables include enterprise types, marriage status, job seeking status, industry categories, city tiers, and a constant term. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

Table 8 shows comparison of OLS and IPWRA results for the coefficients of overeducation. The coefficients in the IPWRA all shift towards the negative direction compared to the OLS results, meaning OLS may underestimate the negative effect, if any, of overeducation. Specifically, for the IT industry, compared to insignificance of overeducation effects with OLS estimates, the IPWRA estimates yield negative effect of overeducation with a significance level of 1%. In the Finance Industry, the pay premium brought by overeducation with OLS results has vanished altogether.

Table 8: Comparison of OLS and IPWRA results

	All	By industry			
		IT	Finance	Education	Real estate
OLS	-0.051***	-0.011	0.076***	-0.064***	-0.046**
	(0.008)	(0.013)	(0.026)	(0.025)	(0.022)
IPWRA	-0.084***	-0.070***	-0.002	-0.093***	-0.098***
	(0.009)	(0.015)	(0.026)	(0.025)	(0.021)

Note: the table shows the coefficients of the variable *overeducation*. Controls in the outcome (Ln(salary)) equation in all columns are the same as in column (1) of Table 6. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

Following Wu and Wang (2018), we also run IPWRA fixing the education level. The reason behind is that both education and overeducation are simultaneously endogenous. It would be more convincing if there was only one endogenous variable. Since the majority population (just below 90%) in our sample are with bachelor or some college degree, it is appropriate to analyze respectively for the two subsamples. Now the education_year variable becomes a constant for the workers with the same education level, and thus is omitted from the RA model. The results shown in Table 8B are consistent with the full sample results from Table A3 in the Appendix. It further reveals that people with some college degree suffer more when over-educated than their counterparts with a bachelor degree. Table A4 uses the propensity score matching (PSM) as an additional robustness check, with the full sample and also subsample by education level. Again, the effects of overeducation are significantly negative, consistent with IPWRA and OLS results (a detailed explanation of this method can be found in Wooldridge (2010)).

6.2. Subsample from key universities

Earlier studies find that graduates from higher quality colleges are not only less likely to be overeducated in the first place, but also more likely to exit the situation over time (Robst 1997). Therefore, we further explore whether job seekers with degrees from Key Universities (i.e., Project 985 and 211 universities) will exhibit different patterns. Table 9 shows results for the coefficients of overeducation. We have two observations. First, the coefficients for the subsample mostly shift towards the positive direction compared to the results for the full sample, to the extent that the signs of the coefficients of overeducation for the IT industry change from insignificantly negative to significantly positive, and the sign for the Real Estate industry changes from significantly

negative to insignificantly positive. This reflects that overeducation may even be encouraged for the graduates from Key Universities. A possible explanation is that on average people from Key Universities possess higher learning and creativity ability, and their overqualification may bring innovation and productivity improvements to employers. Second, in accordance with the full sample results, holding relevant academic degrees is rewarded in the IT industry but not in the Finance Industry; and overeducation is more welcome by the Finance Industry than by the IT industry, presumably because signaling is more important to Finance Industry employers.

Table 9: OLS results for subsample from key universities

	All	IT industry			Finance industry			Educa-tion	Real Estate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	（8）	（9）
Over-education	-0.037**	0.053**	0.062**	0.069**	0.092**	0.092**	0.019	-0.134***	0.006
	(0.017)	(0.027)	(0.027)	(0.029)	(0.045)	(0.045)	(0.075)	(0.052)	(0.046)
Relevant degree			0.088***	0.100***		0.009	-0.049		
			(0.029)	(0.035)		(0.043)	(0.069)		
Overedu X rel. degree				-0.036			0.097		
				(0.059)			(0.086)		

Note: dependent variable is Ln(salary). Controls in all columns are the same as in column (1) of Table 6. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

6.3 . Heterogeneity effects across city tiers

Table 10 shows the proportion of overeducated job seekers and the effects of overeducation across tiers of cities.[8] The results show that both the proportion and the effect of overeducation is less severe in first-tier cities than the rest, and the second-tier cities perform not much differently from the cities of tiers below the third. Presumably this pattern reflects differences on the demand side of the labor market. In cities like Beijing and Shanghai, it is much easier to find a high-end job that requires high education and pays well. On the other hand, high rent and costs of living in cities like Beijing or Shanghai often force young graduates to return to their hometowns or to seek employment in the second and third tier cities, resulting in more prevalent overeducation and stronger pay penalty in these cities.

Table 10: Heterogeneous results with respect to city tiers

	City Tiers		
	First Tier	**Second Tier**	**Lower tiers**
Proportion of overeducation	0.391	0.668	0.702
Effect of overeducation	-0.036***	-0.055***	-0.050*
N	9,696	6,163	1,131

Note: Row "effect of overeducation" shows the coefficients of the variable *overeducation*. Controls in Row "effect of overeducation" are the same as in column (1) of Table 6 except that city-tier dummies are removed. ***, **, * denote significance levels of 1%、5% and 10%.

6.4 Trade-off between majors and tiers of prestige of institutions with pooled OLS

We proceed to conduct heterogeneity analysis according to broad majors and tiers of prestige of graduating institutions. Majors are broadly divided to three categories: LEM (law, economics/finance, management), STEM (science, technology, engineering, mathematics), and Others. Table 11 shows the pooled estimation results with interaction of broad majors and tiers of prestige of graduating institutions. The left panel uses nine categories by the interactions of broad majors (LEM, STEM, and others) and tiers of prestige (Project 985 university graduates, Project 211 university graduates, and others). The right panel uses six categories by the interactions of

[8] The first-tier cities refer to Beijing, Shanghai, Guangzhou, and Shenzhen, while the second-tier cities refer to 32 cities including Hangzhou, Chengdu, Jinan, etc., most of which are provincial capitals. The rest belong to tiers of third, fourth, and below, according to the ranking provided by China Business Network Co. Ltd in 2017.

broad majors (LEM, STEM, and others) and tiers of prestige (Key University graduates, and others).

The upper panel of Table 11 shows coefficients for the variable *overeducation*. Overall Key University graduates benefit from overeducation while others receive pay penalty brought by overeducation. In other words, Key University graduates earn more when they are in a position in which the education requirement is lower than their own level than if they were put in a position in which the education requirement matches his/her own level. On the contrary, graduates from less prestigious universities earn less if they were overeducated. The reason might be that over-educated Key University graduates possess abilities that make them outstanding among colleagues in a lower position, resulting in a higher reward. Besides, the results show that Key University graduates with other majors than LEM and STEM do not enjoy pay premium. Combined, the story being told is that in an era of higher education expansion, education from Key Universities, especially National Project 985 universities, is much more valuable than education from less prestigious institutions, and that the differences in effects between majors are not as strong as that between universities. Therefore, if a student faces trade-off between university tiers and broad majors when pursuing a higher degree, maybe higher than most jobs would require, the suggestion is going for a higher-ranked university, instead of a more popular major. A subtler advice is that within key universities, National Project 985 universities outperforms National Project 211 universities in LEM majors, but not STEM majors, in terms of overeducation effects.

Table 11: Overeducation by broad majors and university tiers

Variables	Pooled OLS (9 categories)		Pooled OLS (6 categories)	
Overeducation	LEM_985	0.065**	LEM_key	0.046**
		(0.028)		(0.019)
	LEM_211	0.028		
		(0.026)		
	LEM_other	-0.072***	LEM_other	-0.072***
		(0.011)		(0.011)
	STEM_985	0.065**	STEM_key	0.074***
		(0.027)		(0.020)
	STEM_211	0.085***		
		(0.030)		
	STEM_other	-0.072***	STEM_other	-0.072***
		(0.010)		(0.010)
	other_985	-0.012	other_key	-0.014
		(0.026)		(0.018)
	other_211	-0.017		
		(0.023)		
	other_other	-0.096***	other_other	-0.096***
		(0.011)		(0.011)
Other variables	female	-0.220***		-0.220***
		(0.007)		(0.007)
	eduy	0.136***		0.136***
		(0.004)		(0.004)
	work_exp	0.054***		0.054***
		(0.002)		(0.002)
	work_exp^2	-0.001***		-0.001***
		(0.000)		(0.000)
	job_history#	0.039***		0.039***
		(0.003)		(0.003)
N		16,990		16,990
Adj. R-sq		0.496		0.496

Note: dependent variable is Ln(salary). All regressions control for enterprise types, marriage status, job seeking status, industry categories, city tiers, and a constant term. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

6.5 Linear probability model with overeducation as the dependent variable

It is also interesting to see how different combinations of broad majors and university tiers affect the probability of overeducation itself. Table 12 shows the linear probability model (LPM) results with overeducation as the dependent variable.

Table 12: Linear Probability Model with overeducation as the dependent variable

Variables	**Pooled OLS (9 categories)**		**Pooled OLS (6 categories)**	
Overeducation	LEM_985	-0.063***	LEM_key	-0.047***
		(0.020)		(0.017)
	LEM_211	-0.031		
		(0.020)		
	LEM_other	0.023	LEM_other	0.023
		(0.014)		(0.014)
	STEM_985	-0.070***	STEM_key	-0.092***
		(0.019)		(0.017)
	STEM_211	-0.120***		
		(0.022)		
	STEM_other	-0.026*	STEM_other	-0.027*
		(0.015)		(0.015)
	other_985	-0.033	other_key	-0.011
		(0.023)		(0.019)
	other_211	0.012		
		(0.022)		
	other_other	0.068***	other_other	0.067***
		(0.016)		(0.016)
N		16,990		16,990
Adj. R-sq		0.149		0.149

Note: All regressions control for: female, years of education, and a constant term. Standard errors in parentheses. ***, **, * denote significance levels of 1%、5% and 10%.

Generally, Key University graduates are less likely to be overeducated for their jobs, i.e., they are more likely to find a job that matches their education qualification. Similarly, people with STEM majors, whether from Key University or not, are more likely to find a job that matches their education qualification. As for LEM majors, only graduates from Key Universities are more likely

to find a job that matches their education qualification. The most unfortunate group is those graduates from non-Key Universities studying non-LEM/STEM majors, who are most likely to be overeducated for their job.

7. Conclusions

Employing a novel dataset collected from an online recruitment platform (*zhaopin.com*) in China, we investigate the impact of overeducation on labor income. With the help of the word segmentation and dictionary building techniques, we extract the pre-match information of both the supply side and the demand side in the labor market. We find that about half of the job seekers in the online platform are overeducated by two or more years, and that overeducation lead to a 5.1% penalty in wage payments on average. Moreover, we examine the heterogeneous effects of overeducation on payment across industries, university types, majors, as well as city tiers. The results indicate that overeducation will be penalized in the IT industry, but not in the Finance Industry; job-seekers graduated from the Key universities or living in the first-tier cities are less likely to be overeducated and less likely to be punished when overeducated.

With a 10-fold expansion of the HE sector in China since 1999, there is now compelling evidence of over-education. The mismatch between the supply side and the demand side can cause significant imbalances that could have profound influences on both individuals and the nation at large. From the individual's perspective, pursuing degrees in more selective universities with "hot" majors, and trying to work in first-tier cities could monetarize their educational input to a maximum degree. From the government's perspective, more focus should be diverted to improving the quality of teaching and subject matching rather than further expansion of the HE sector.

We contribute to the literature by using a novel dataset yet reaching remarkably consistent results with previous literature. Our word segmentation and dictionary building techniques, applied on internet data, is easily adaptable to other countries where similar online recruitment platforms are available. This could largely enrich the empirical literature on overeducation over the globe in the future, given the expansion of the HE sector across countries. The timeliness of objective data reflecting both labor demand and labor supply could help us monitor the situation more frequently. There is limitation, of course. The data can only provide pre-match information. It could be more desirable if after-match scenario is tracked and documented, which could be a promising venue for future research.

References

Albrecht, J. and Vroman, S., 2002. A matching model with endogenous skill requirements. *International Economic Review 43*(1), 283-305.

Bauer, T.K., 2002. Educational mismatch and wages: a panel analysis. *Economics of Education Review 21*(3), 221-229.

Berg, I., 1970. Education for Jobs; The Great Training Robbery.

Capraro V, Sippel J., 2017. Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive processing* 18(4): 399–405.

Carrington, W., 1993. Wage Losses for Displaced Workers: Is It Really the Firm That Matters? *Journal of Human Resources* 28, 435-62.

Cattaneo, M. D. 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155: 138– 154.

Cohn, E. and Ng, Y.C., 2000. Incidence and wage effects of overschooling and underschooling in Hong Kong. *Economics of Education Review 19*(2), 159-168.

Duncan, G.J. and Hoffman, S.D., 1981. The incidence and wage effects of overeducation. *Economics of Education review 1*(1), 75-86.

Eckaus, R.S., 1964. Economic criteria for education and training. *The Review of Economics and Statistics* 46(2), 181-190.

Elias, P. and Purcell, K., 2004. Is mass higher education working? Evidence from the labour market experiences of recent graduates, *National Institute Economic Review,* pp. 60–74.

Frank, R.H.,1978. Why Women Earn Less: The Theory and Estimation of Differential Overqualification. *American Economic Review* 68(3), 360-373.

Freeman, R., 1976. The overeducated American.

Gao, K., Wang, Y. F., & Zheng, P. Q., 2017. Over-Education in China. *Chinese Studies* 6, 37-43.

Galasi, P., 2008. *The effect of educational mismatch on wages for 25 countries* (No. BWP-2008/8). Budapest Working Papers on the Labour Market.

Green, C., Kler, P. and Leeves, G., 2007. Immigrant overeducation: Evidence from recent arrivals to Australia. *Economics of Education Review 26*(4), 420-432.

Green, F. and Zhu, Y., 2010. Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. *Oxford Economic Papers 62*(4), 740-763.

Gustafsson, B., LI, S., and Sato, H., 2014. Data for studying earnings, the distribution of household income and poverty in China. *China Economic Review* 30, 419–431.

Hartog, J. and Oosterbeek, H., 1988. Education, allocation and earnings in the Netherlands: Overschooling? *Economics of Education Review 7*(2), 185-194.

Hartog, J., 2000. Over-education and earnings: where are we, where should we go? *Economics of Education Review*, *19*(2), 131-147.

Kang, L., Peng, F. and Zhu Y., 2019. Returns to higher education subjects and tiers in China: evidence from the China Family Panel Studies, *Studies in Higher Education* (online first).

Kiker, B.F., Santos, M.C. and De Oliveira, M.M., 1997. Overeducation and undereducation: evidence for Portugal. *Economics of Education Review*, *16*(2), 111-125.

Kim, K., Mithas, S., Whitaker, J.W. and Roy, P.K., 2014. Industry-Specific Human Capital and Wages: Evidence from the Business Process Outsourcing Industry, *Information Systems Research* 25(3), 618-638.

Kletzer, L.G., 1996. The Role of Sector-Specific Skills in Post-displacement earnings, *Industrial Relations* 35(4), 473-90.

Korpi, T. and Tåhlin, M., 2009. Educational mismatch, wages, and wage growth: Overeducation in Sweden, 1974–2000. *Labour Economics 16*(2), 183-193.

Lazear, E., 1977. Education: consumption or production? *Journal of Political Economy*, *85*(3), 569-597.

Leuven, E. and Oosterbeek, H., 2011. Overeducation and mismatch in the labor market. In *Handbook of the Economics of Education* 4, 283-326. Elsevier.

Li, Q., Li, S. and Wan, H., 2020. Top incomes in China: Data collection and the impact on income inequality, forthcoming, *China Economic Review.*

Luo, R. and Peng, M., 2010. Overeducation and its trend: A research based on CGSS. *Comparative Economic & Social Systems* 05, 173–179.

Maynard, D.C., Joseph, T.A. and Maynard, A.M., 2006. Underemployment, job attitudes, and turnover intentions. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *27*(4), 509-536.

Office for National Statistics (ONS), 2019. Overeducation and hourly wages in the UK labour market; 2006 to 2017. London.

Quinn, M.A. and Rubb, S., 2006. Mexico's labor market: The importance of education-occupation matching on wages and productivity in developing countries. *Economics of Education Review*, *25*(2), 147-156.

Robst, J., 1994. Measurement error and the returns to excess schooling. *Applied Economics Letters 1*(9), 142-144.

Robst, J., 1997. College quality and overeducation. *Economics of Education Review 14*(3), 221-228.

Robst, J., 2007. Education and job match: The relatedness of college major and work. *Economics of Education Review* 26, 397-407.

Robst, J., 2008. Overeducation and college major: expanding the definition of mismatch between schooling and jobs. *Manchester School* 76(4), 349-368.

Rosenberg M. Conceiving the self. 1979. Basic, New York. 1979.

Sicherman, N. and Galor, O., 1990. A theory of career mobility. *Journal of Political Economy 98*(1), 169-192.

Sloane, P.J., 2003. Much ado about nothing? What does the overeducation literature really tell us? in F. Buchel, A. de Grip, and A. Mertens (eds.) *Overeducation in Europe: Current Issues in Theory and Policy*, Edward Elgar, Cheltenham.

Spence, M., 1978. Job market signaling. In *Uncertainty in Economics* (pp. 281-306).

Thurow, L.C., 1975. *Generating inequality*. Basic books.

Verdugo, R.R. and Verdugo, N.T., 1989. The impact of surplus schooling on earnings: Some additional findings. *Journal of Human Resources* 24(4), 629-643.

Walker, I. and Zhu, Y., 2008. The college wage premium and the expansion of higher education in the UK, *Scandinavian Journal of Economics* 110, 695–709.

Walker, I., Zhu, Y., 2018. University selectivity and the relative returns to higher Education: Evidence from the UK. *Labour Economics* 53, 230-249.

Wooldridge, J., 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141, 1281-1301.

Wooldridge, J. M., 2010. Econometric analysis of cross section and panel data. MIT Press.

Wu, X. and Lai, D., 2010. On the incidence of over-education and its influencing factors: based on the analysis of Beijing city, *Research in Educational Development*, 30, 36-41.

Wu, N. and Wang Q., 2018. Wage penalty of overeducation: New micro-evidence from China, *China Economic Review* 50, 206-217.

Yin, L. 2016. Overeducation in the Chinese Labour Market. *Unpublished PhD thesis*. University of Sheffield.

Zhang, J. and Zhao, W., 2019. The unreported income and its impact on Gini coefficient in China, *Journal of Chinese Economic and Business Studies* 17:3, 245-259.

Zhang, X., Zheng, Y., 2019. Gender differences in self-view and desired salaries: A study on online recruitment website users in China. *PLoS ONE*, 14(1).

Appendix

Table A1: OLS regression results allowing for undereducation

	All	IT industry			Finance industry			Education	Real estate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	（8）	（9）
Overeducation	-0.060***	-0.024*	-0.018	-0.013	0.063**	0.063**	0.029	-0.067***	-0.071***
	(0.008)	(0.013)	(0.013)	(0.014)	(0.025)	(0.025)	(0.038)	(0.024)	(0.021)
Undereducation	0.111***	0.139***	0.140***	0.139***	0.048	0.049	0.052	0.197***	0.177**
	(0.025)	(0.044)	(0.045)	(0.045)	(0.121)	(0.121)	(0.121)	(0.073)	(0.074)
Relevant Degree			0.083***	0.091***		-0.013	-0.040		
			(0.014)	(0.017)		(0.023)	(0.033)		
Overedu X Relevant Degree				-0.022			0.050		
				(0.029)			(0.044)		
Eduy	0.136***	0.137***	0.137***	0.137***	0.110***	0.111***	0.111***	0.130***	0.136***
	(0.004)	(0.007)	(0.007)	(0.007)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)
Grad985	0.138***	0.117***	0.115***	0.115***	0.170***	0.171***	0.173***	0.159***	0.105***
	(0.011)	(0.017)	(0.017)	(0.017)	(0.028)	(0.029)	(0.029)	(0.031)	(0.027)
Grad211	0.097***	0.071***	0.072***	0.072***	0.079**	0.079**	0.081**	0.033	0.094***
	(0.011)	(0.017)	(0.017)	(0.017)	(0.032)	(0.032)	(0.032)	(0.031)	(0.027)
Work_Exp	0.054***	0.068***	0.067***	0.068***	0.062***	0.062***	0.062***	0.044***	0.056***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.006)	(0.005)
Work_Exp^2	-0.001***	-0.002***	-0.002***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Job_History#	0.037***	0.039***	0.040***	0.040***	0.039***	0.039***	0.039***	0.033***	0.039***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)
Female	-0.220***	-0.234***	-0.218***	-0.218***	-0.197***	-0.196***	-0.194***	-0.265***	-0.202***
	(0.007)	(0.012)	(0.012)	(0.012)	(0.022)	(0.023)	(0.023)	(0.022)	(0.018)
N	16,947	6,239	6,239	6,239	1,999	1,999	1,999	2,279	2,938
Adj. R-Sq	0.499	0.505	0.508	0.508	0.520	0.520	0.520	0.490	0.489

Note: the following variables have also been controlled for in all regressions above: enterprise types, marriage status, job seeking status, industry categories, city tiers, and a constant term. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

Table A2: OLS regression results with required years of education and years overeducated

	All	IT industry			Finance industry			Education	Real estate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	（8）	（9）
years_overedu	0.117***	0.123***	0.123***	0.124***	0.123***	0.123***	0.123***	0.103***	0.111***
	(0.003)	(0.006)	(0.006)	(0.006)	(0.011)	(0.011)	(0.011)	(0.009)	(0.008)
Years_edu required	0.159***	0.176***	0.173***	0.172***	0.143***	0.143***	0.143***	0.157***	0.181***
	(0.005)	(0.010)	(0.010)	(0.010)	(0.018)	(0.018)	(0.018)	(0.016)	(0.014)
Relevant degree			0.075***	0.174		-0.011	0.055		
			(0.014)	(0.124)		(0.023)	(0.262)		
Overedu X relevant degree				-0.099			-0.067		
				(0.124)			(0.262)		
grad985	0.136***	0.115***	0.113***	0.113***	0.168***	0.169***	0.169***	0.157***	0.103***
	(0.011)	(0.017)	(0.017)	(0.017)	(0.028)	(0.029)	(0.029)	(0.031)	(0.027)
grad211	0.096***	0.071***	0.072***	0.072***	0.079**	0.080**	0.080**	0.032	0.095***
	(0.011)	(0.017)	(0.017)	(0.017)	(0.032)	(0.032)	(0.032)	(0.031)	(0.027)
work_exp	0.054***	0.068***	0.068***	0.068***	0.062***	0.062***	0.062***	0.045***	0.056***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)	(0.007)	(0.006)	(0.005)
work_exp^2	-0.001***	-0.002***	-0.002***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
job_history#	0.037***	0.040***	0.041***	0.041***	0.039***	0.040***	0.040***	0.034***	0.040***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)
female	-0.221***	-0.232***	-0.218***	-0.218***	-0.199***	-0.198***	-0.197***	-0.266***	-0.203***
	(0.007)	(0.011)	(0.012)	(0.012)	(0.022)	(0.022)	(0.022)	(0.022)	(0.018)
N	16,947	6,239	6,239	6,239	1,999	1,999	1,999	2,279	2,938
adj. R-sq	0.501	0.508	0.510	0.510	0.519	0.519	0.519	0.491	0.493

Note: the following variables have also been controlled for in all regressions above: enterprise types, marriage status, job seeking status, industry categories, city tiers, and a constant term. Standard errors in parentheses. ***,**,* denote significance levels of 1%、5% and 10%.

Table A3: IPWRA results when education level is fixed

	All	By industry			
		IT	Finance	Education	Real estate
Bachelor	-0.060***	-0.056***	-0.067*	-0.052	-0.160***
	(0.013)	(0.019)	(0.035)	(0.033)	(0.043)
Some college	-0.224***	-0.297***	-0.293***	-0.387***	-0.280***
	(0.011)	(0.023)	(0.044)	(0.037)	(0.026)

Table A4: PSM results

	All	Bachelor	Some college
ATE	-0.056***	-0.053***	-0.095***
	(0.015)	(0.017)	(0.015)
ATET	-0.031**	-0.055***	-0.073***
	(0.015)	(0.021)	(0.018)
N	17,565	7,930	7,504