Has the College Premium really Flattened?

Tomás Guanziroli
(PUC-Rio)

Ariadna Jou
(UCLA)

October, 2022

PRELIMINARY AND INCOMPLETE

Abstract

Recent studies show that the college premium has flattened in the last two decades in many Latin American countries. At the same time, there was a great expansion in the number of college graduates and college institutions in the region. This paper shows that the college premium in Brazil has not flattened, instead it is increasing. We do this by matching novel data with the names of around one million college graduates from 42 schools and 20 different cohorts with the Brazilian employer-employee matched dataset. First, the increase in the supply of college workers came from newer, lower ranked, and lower wage-premium universities. Second, the college premium has increased for workers from our constant sample of universities. Combining these two facts, we infer that there are more workers with a college degree but lower quality degrees, reflecting lower average wages. The findings in this paper are relevant for any study that uses college premium proxies in countries, or periods, with increasing access to lower-quality colleges. More precisely, we are concerned that the market equilibrium effects of college access have been overestimated by not accounting for these changes in quality.

I. Introduction

The statistics commonly known as the college wage premium has received great attention in the past fifty years.[1] Its charm comes from highlighting movements in wage inequality (Freeman, 1976; Katz and Murphy, 1992; Card and Lemieux, 2001; Goldin and Katz, 2008; Acemoglu and Autor, 2011) and for serving as reference for public and private investments in higher education (Rodriguez et al., 2016; Bank, 2019). While the college premium has increased in developed countries, it has flattened or even decreased in many Latin American countries during the last 20 years (Fernandez and Messina, 2018). At the same time, there was a great expansion in the number of college graduates and college institutions in the region. Although it is usual to interpret the effects of such expansion on the college wage premium through the lenses of market equilibrium, the characteristics of workers with a college degree have considerably changed over time and possibly affected this measure.

In this paper, we study whether a reduction in the average quality of college institutions and students is responsible for the flattening/decrease in the college premium in Brazil.[2] First, we show that students from older and more established institutions perform better in the end-of-degree standardized exams, which suggests that these are better schools. Second, we show that the college premium has increased when we hold constant this set of universities. Combining these two facts, we infer that a composition change is driving the flattening of the college premium. There are more workers with a college degree but lower quality degrees, which reflects lower average wages.

To perform this analysis, we create a long panel with college data matched with labor market data. First, we constructed novel data of one million students that graduated between 1990 and 2020 in 42 Brazilian universities.[3] The data were constructed from FOIL requests

[1] The college premium is the ratio between the average wages of all workers with a college degree and all workers with a high school degree.

[2] This is also referred to as the degraded tertiary hypothesis (Camacho et al., 2017)

[3] The universities in our sample are public and are representative of the best universities in the country.

(Freedom of Information Law) and include information on students' names, their institution, major, and year of graduation. To get information on students' wages, age, and other characteristics, we use the Brazilian employer-employee matched dataset (RAIS). We match both datasets by name from 2007 to 2018. We use a machine learning approach to decide which is the best match for student-workers with multiple matches. As a result, we have annual wage data on workers of different ages and different cohorts that graduated from a constant set of universities.

We first document an increase in the supply of college workers over the last decades. The number of people graduating from any college institution increased by three times between 2000 and 2018, from 400 thousand to 1.2 million people. The stock of college workers in the formal labor market increased as well, from 4 million to 12 million in the same period. These trends are not specific to Brazil. In fact, college enrollment in upper-middle-income countries—a group of 54 countries as defined by the World Bank—has increased by a similar magnitude during the same period as Brazil.[4]

Second, we present evidence that this growth in enrollment came from new and possibly worse-quality college institutions. For example, 80% of the students that graduated in 2018 were in a major founded after the year 2000. We show that newer universities and majors are often lower ranked, and their students perform worse in standardized exams. Furthermore, an increasing share of students graduates from distance education (INEP, 2020), which could presumably be of lower quality. Institution quality is not the only thing that changed in the period: new cohorts of students have a lower socio-economic background when compared to older cohorts (ANDIFES, 2019). We interpret this as preliminary evidence that the skill level of the median college graduate decreased in the last two decades.

The increase in labor supply and the changes in the median quality of workers can both play a role in the evolution of the college premium—the challenge is to decompose these

[4]UNESCO Institute for Statistics (uis.unesco.org). School enrollment, tertiary (% gross) - Brazil, Upper middle income. Data as of June 2022, accessed through https://data.worldbank.org/.

effects. We show that the college premium was flat between 1998 and 2005 and decreased after that.[5] Given an increase in the labor supply of college workers, there are two opposing explanations for the flattening and decline in the college premium trends: (i) labor demand for college workers remained fixed, and the decline in returns is defined by a movement along the labor demand; or (ii) skill-biased technological change increases the labor demand for college workers for any skill level, increasing returns to skill. However, the reduction in the median skill level of college workers has a stronger effect, reducing the college premium measure. Which explanation is more appropriate is an empirical question.

We identify the change in returns to skills by fixing one of the quality components: universities and majors. For this analysis, we use the RAIS dataset matched to our sample of college graduates. We observe each worker's labor market outcome for multiple years, which allows us to separately identify the year of graduation effects (cohorts) from year effects. Furthermore, workers graduate at different ages, such that we are not constrained by the age-cohort-year identification problem. Under the assumption that the average quality of universities in our sample is stable, we identify the change in returns to skill for high-skill workers.

Our main results show that the college premium restricted to workers from a set of high-quality universities has increased. We find that the college wage premium increased by 23% between 2007 and 2018—an average increase of 2% per year. When we consider the full sample of college workers—i.e., including changes in composition—we find that the college premium decreased by 12%.

A story in which all universities—including lower ranked institutions—have increasing college premium trends that aggregate into a decreasing overall college premium trend is consistent with our results. This conclusion depends on three facts: (i) The share of students graduating from lower ranked universities increased over time; (ii) At any given year, students

[5]We use the nationally representative household survey (PNAD). Similar trends are found using the universe of formal workers in Brazil (RAIS).

that graduated from higher ranked universities earn higher wages; and (iii) All universities have an increasing college premium trend. We have already mentioned that (i) holds in our data. We present evidence that (ii) holds by documenting a positive correlation between earnings and university ranking. For example, students from the top 10 universities receive 20% higher wages than students from universities ranked around the 100th place. Additional data is required to prove that (iii) holds, which should be the object of future research.

This paper contributes to a few strands of the literature. We are the first to our knowledge to show that changes in the quality composition of higher education institutions are responsible for the trends in the college premium in Brazil. Although this hypothesis has been brought up by several studies, the literature had not reached a consensus—probably due to data limitations.[6] Rodriguez et al. (2016) and Camacho et al. (2017) show that returns to college are heterogeneous in Chile and Colombia, respectively, with recently created programs having lower returns.[7] However, due to data limitations, they cannot explain the changes in the college premium trends. We contribute to the discussion by presenting evidence that uses a long panel with matched information on workers' wages and universities for different cohorts of workers.

Our results are also relevant to the growing literature that looks at the causes of the decline in earnings inequality in Latin America. Barros et al. (2010) argue that half of the decline in inequality was caused by an acceleration of educational progress. Ferreira et al. (2017), Alvarez et al. (2018) and Fernandez and Messina (2018) disagree with the latter statement and argue the decrease in earnings inequality is due to a compression of returns to firm and worker characteristics, such as experience and education We argue that decompositions of certain measures of inequality mistakenly attribute decreasing returns to

[6]This hypothesis is sometimes referred to as "degraded tertiary hypothesis". See Messina and Silva (2017) for a review on the topic.

[7]Camacho et al. (2017) argue that the effects come from self-selection.

education to a group that had increasing returns, possibly biasing the results.[8]

The results in this paper have strong implications regarding the public decision to invest in higher education. Previous studies found that the labor demand for college workers is relatively inelastic and that firms can easily substitute skilled for non-skilled workers (Katz and Murphy, 1992; Acemoglu, 2002; Ciccone and Peri, 2005;Haanwinckel, 2020). This implies that public investments that increase the supply of workers have the unintended effect of reducing relative wages for all workers with college. Our results show that returns to high quality education have continued to increase, consistent with skill-biased technological change and/or a more elastic labor demand. As a consequence, investments in higher education had increasing returns in the past, a trend that may continue in the future.

The paper proceeds as follows. Section II. presents the data sources and shows that there was a decrease in the college premium measure during the last two decades. In section III., we present preliminary evidence that the skill level of the median college graduate decreased over time. Section IV. presents the decomposition strategy and the main results of the paper. Section V. concludes.

II. Data Sources and Descriptive Statistics

We gathered information on college graduates from a selected sample of universities and match them to the Brazilian linked employer-employee dataset in order to compute the college premium.[9]

A. Linked Employer-Employee dataset (RAIS) and Household Survey (PNAD)

In our main analysis, we use earnings and schooling data from the *Relação Anual de*

[8]Common measures of inequality are the ratio of log earnings of the 90th and 50th or 10th percentile in the wage distribution.

[9]College premiums are usually computed using household surveys, which are representative at the national level. However, such surveys do not include information on worker's schools.

Informações Sociais (RAIS) from 2007 to 2018. RAIS is an administrative dataset that covers the universe of formal employees and firms in the private and public sectors, with the exception of domestic service workers. The data is administered by the Ministry of Labor and has restricted access. RAIS has information on individuals (CPF, full name, age, gender, race, schooling), on firms (CNPJ and sector), on establishments (county, zip code, and name), and on the employer-employee match (wages, occupation, tenure, dates of hiring and firing/separation).

To circumvent the fact that RAIS is not representative of the entire Brazilian population, we also use data from the *Pesquisa Nacional por Amostra de Domicílios* (PNAD). PNAD is the Brazilian household survey (PNAD), which was collected annually between 1970 and 2015. The survey was later replaced by PNAD-Continua, which is collected quarterly and provides a better representation of the Brazilian population. PNAD-Continua is available between 2012 and 2021.

B. College Graduates Sample

We gathered data on college graduates from public universities in Brazil through FOIL requests (Freedom of Information Law). For all universities, the data includes full name, university, major, year of admission, and year of graduation. Some universities provided more information, such as national identification number, gender, etc. Appendix Table A.1 lists the sample of universities that agreed with providing the basic information. We have information on 1.2 million students that graduated from 42 federal/state universities between 1990 and 2020.[10]

Below, we describe the procedure to match the college graduates' sample with the employer-employee dataset (RAIS) and clean the data. We first match the college graduate sample with the RAIS using student/worker's full names. This procedure leads to multiple matches.

[10]The variables and the number of cohorts available vary across universities.

Secondly, we use a machine learning algorithm to select the best match. Third, we impose some sample restrictions.

The raw college graduates sample includes information of 1,217,440 students that graduated from 42 universities. After removing special characters, we are able to match 74% of these students to one or more workers with the exact same full name in the RAIS dataset. As a result, 2,093,069 workers are matched to 906,420 students. Out of these students, 78% are matched to a single worker and 22% are matched to at most 20 workers with the exact same name. We drop students with very common names that are matched to more than 20 workers.

Among students with multiple matches, we select the best match using a machine learning procedure. One university provided us with students' identification number such that, for this university, we can match students and workers by both name and identification number. We proceed by matching students by name, and, for a training sample, we estimate a model that uses student and worker's characteristics to predict whether the match is correct—as defined by the match using the identification number.[11]

We estimate two different models—a logit regression, and a random forest model—using the training sample. Using the model's estimates, we calculate a score for each match. We define the correct match based on 3 rules: (i) The match has the highest score of all matches; (ii) The score of the match is sufficiently large (greater than 5%); and (iii) The score of the top match is sufficiently large relative to the second-best match (the ratio of scores is greater than 1.1).

Appendix Table A.3 compares the results of each model using two metrics: the positive predictive value (PPV or accuracy) and the true positive rate (TPR or efficiency). While

[11]We use the following characteristics that are specific to the student-worker match: five indicators for age at college admission (< 17, > 20, > 25, > 35, > 45 years old) where age is determined by the worker variable; difference between worker's earliest year on the data with year of college admission; an indicator for whether the individual had a full time job during college; an indicator for whether the schooling variable at RAIS reports an educational achievement inferior to college after the year of graduation; the number of worker observations at RAIS; and indicators for the maximum schooling from all worker's observations.

the random forest model has better PPV and TPR rates in the training sample (96.6% and 97.2%, respectively), these numbers do not produce better results in the test sample. Therefore, we decide to use the logit approach due to the consistency of training sample and test sample metrics (PPV of 87.1% and 87.2%). Appendix Table A.4 presents the sample sizes used for the in sample PPV and TPR calculations.

The machine learning algorithm finds a match for 806,893 students/workers. We further restrict the sample to students that graduated between 2000 and 2017. The final dataset includes 17,455,296 employment observations from 545,478 students/workers, that graduated from 42 universities. In the rest of the paper, we refer to students in this sample as the "college graduates' sample" and the universities as "sample universities". All the other universities in Brazil are referred to as "out-of-sample universities".

C. Higher Education Census and University Rankings

We use four additional data to complement the analysis. First, we use the Brazilian census of higher education to document the growth in enrollment by type of institution. The census is available for every year between 1995 and 2018 and covers all universities and majors in Brazil.[12] It includes information such as total enrollment, number of graduates, and each major's date of foundation. In 1995, there were 884 higher education institutions and around 7,000 majors in Brazil. There was strong growth in the sector, such that in 2019 there were 2,608 higher education institutions and more than 40,000 majors.

Secondly, we use national examination data from Exame Nacional de Desempenho dos Estudantes (ENADE), and two rankings of higher institutions that are published online (RUF, from Folha de Sao Paulo newspaper; and Web ranking) to rank universities. Using ENADE's data from 2014, we create a university score by aggregating scores from all students and majors from each university. In Section 3, we combine these data to describe the changes in the composition of college graduates over time in terms of school age and ranking.

[12]University identification numbers are only matched across years after the year 2000.

Using this data, we show that sample universities are older and better ranked. Table 1 presents the age distribution of sample universities and out-of-sample universities. The data comes from the RUF ranking and is limited to 194 universities. The table shows that most universities in our sample were founded more than 50 years ago (59.5%). In comparison, only 37.7% of out-of-sample universities were founded more than 50 years ago.[13]

Table 2 presents the RUF score and ranking, and the web ranking for the two samples. Universities in our sample have better scores and as a consequence are better ranked, according to the RUF ranking. The Web ranking includes more universities (1,285) and shows an even larger discrepancy between the sample universities and the out-of-sample universities. The median ranking in the sample is 37, and 663 out-of-sample. In summary, our sample includes many of the best and oldest universities in the country.

D. Age-adjusted College Premium

We define age-adjusted college premium as the weighted ratio of earnings for workers with a college degree and workers with a high school degree. Similar to Fernandez and Messina (2018), premiums are constructed using a fixed-weight average of every age subgroup, for workers of ages between 21 and 65 years old. The weights are equal to the mean employment share of each subgroup across all years. We present the weighted average by aggregating all groups.

Figure 1 presents the evolution of the college premium over time using the nationally representative household survey (PNAD, in Panel A) and the employer-employee matched dataset (RAIS, in Panel B). Panel A shows that, on average, college workers' wages were 123% higher than the wages of workers with only a high school degree (2.23 times higher). The college premium increased by around 20p.p. between 1997 and 2004, a 3p.p. annual growth rate. Between 2004 and 2015, the college premium decreased by 19p.p., or -2p.p. a

[13]This number is probably overestimated since the RUF ranking only includes 194 universities out of 2000 higher education institutions.

year. Fernandez and Messina (2018), describe a similar picture for Brazil and other countries in Latin America.

Panel B of Figuree 1 shows that in 1995, conditional on having formal employment, workers with a college degree earned 86% higher wages than workers with only a high school degree. There was a strong increase in the college premium between 1994 and 2002 when the difference in wages was 137%. This represents a 6p.p. annual growth. The college premium has a more moderate growth between 2002 and 2012 when it reaches the peak of a 150% difference in average wages (1p.p. per year). Finally, the college premium decreases to a 129% difference between 2012 and 2019 (-3p.p. per year).

In summary, the age-adjusted college premium had a ceiling during the last two decades. This is described both by the nationally representative household survey and by the census of formal employees, even though they have very distinct samples.

E. Changes in labor supply

During the same period, there was a large increase in the supply of college workers and in the supply of high school workers. Figure 2 shows that the number of formal workers with a college degree has substantially increased between 1995 and 2019 (RAIS). There were 3 million formal workers with a college degree in 1995 and 12 million in 2020. Using the census of higher education, we show that the number of people graduating from any college institution increased from 400 thousand to 1.2 million people per year. These trends are not specific to Brazil. In fact, college enrollment in upper-middle-income countries—a group of 54 countries as defined by the World Bank—has increased by a similar magnitude during the same period as in Brazil.

The number of workers with a high school degree has also increased but the quality of primary and secondary education remained low. Appendix Figure A.1 shows that there were 5 million formal job relations with positive wages in which workers had a high school degree

in 1995. In 2019, there were almost 30 million. This is a consequence of public policies such as compulsory school laws and also the increase in formal employment. However, public investments in education did not target the quality of education. In 2000, Brazil was ranked 30th out of 30 countries in the Programme for International Student Assessment (PISA). In 2018, Brazil was ranked 65th of 78 countries, with little evolution in scores.

We disagree with the standard economic analysis that uses these trends in college premium and labor supply to make inferences regarding returns to skill. Previous studies have associated a decreasing college premium with decreasing returns to skill. In this context, the labor supply curve for skilled workers shifted to the right while the labor demand curve remained fixed, resulting in decreasing returns to skill. In contrast, we argue that there were profound transformations in the higher education sector in Brazil. These transformations led to changes in the quality composition of college workers, such that trends in college premium cannot be interpreted as trends in returns to skills. In the rest of the paper, we make this point by comparing the college premium trends presented in panel B of Figure 1 with measures that account for changes in university composition.

III. Preliminary Evidence of Changes in Graduates' Skill Composition

We argue that the growth in the supply of college workers reduced the average skill of college graduates.

First, most of the growth in the supply of college graduates is due to the introduction of new majors and institutions. Figure 3 uses annual data from the higher education census and presents the number of students that graduated from a bachelor's program in each year. In 2000, around 400 thousand people graduated from a bachelor's program. The number of graduates was three times greater by 2018. Much of this growth comes from new majors

and new institutions. The figure shows that the number of graduates from majors that were created before the year 2000 is steady over time and possibly decreasing. Growth in number of graduates in the 2000's (2010's) comes from majors founded in the 2000's (2010's) and not by increases in class size from older majors.[14]

Secondly, new institutions are lower ranked, and their students perform worse on national exams. We regress the ENADE score and RUF ranking on dummy categories of institutions' age, as defined by the year of foundation of the university's first major. Table 3 presents the results, which should be interpreted as deviations from institutions founded before 1940. The table shows that students in universities founded recently have worse scores in the ENADE's exam in both specific knowledge (related to the major) and in general knowledge. Column 4 shows that RUD ranking is increasing on university's age.

In summary, older universities are better ranked but they used to represent a higher share of graduates. For example, Figure 4 shows that the share of graduates from Top 25 universities according to ENADE ranking has substantially decreased.

IV. The Adjusted College Premium Trends

In the previous section, we learned that a smaller share of students graduated from the older and higher ranked universities in the past years. In this section, we study the effects of such change in composition on the college premium. We do that by decomposing the college premium into two samples: (a) the college graduates' sample from FOIL requests matched to RAIS, and (b) all workers in the RAIS dataset with a college degree, excluding workers from the previous sample. Note that "sample a" holds the university composition constant over time, but "sample b" does not.

To compute the college premium, we regress log wages of individual i, of schooling s, age

[14]Appendix Figure A.2 shows similar trends categorized by the age of the oldest major in the higher education institution. The figure shows that half the growth in the number of college graduates comes from institutions that were founded between 1990 and 2000 and after the year 2000.

a, on year t on a set of dummies as shown by Equation 1:

$$ln(wage)_{i(s,a)t} = \delta_{t,s} + \alpha_{a,s} + \psi_c + \eta_{u,m} + \varepsilon_{it} \quad (1)$$

Where $\delta_{t,s}$ are year by schooling fixed effects, $\alpha_{a,s}$ are age by schooling fixed effects; ψ_c are cohort of graduation fixed effect; and $\eta_{u,m}$ are university by major fixed effects.[15] The previous two fixed effects also equal to zero for high school workers and for workers whom we do not have university data. We omit dummies for the first year in the data due to collinearity. The sample includes all workers in the RAIS data between ages 21 and 65 that were employed on December 31st, worked 40 hours a week, had positive wages in December and had either complete high school or college education. The resulting sample has 263 million observations.

As shown in Section 2, the unconditional college premium has a decreasing pattern. Figure 5 presents the estimates of $\delta_{t,s}$ from Equation 1, which represent the changes in log wages of each sample relative to the wages of high school workers, taking the year of 2007 as the benchmark. The dashed line represents the college premium for the sample of all workers in the RAIS dataset with a college degree, excluding workers from "sample a". The figure shows that the college premium remained constant between 2007 and 2011 and decreased by 14.5% between 2011 and 2018—a 1.8% annual decrease.

However, when fixing the sample to students that graduated from the same group of universities, we note that the college premium has actually increased. The black line in Figure 5 presents the college premium for the college graduate's sample from FOIL requests (Sample Universities). The figure shows that the college premium increased by 19% between 2007 and 2011—a 4% annual increase. The college premium increases at a slower pace between 2011 and 2018, by 5% in total or 0.6% annually.

[15] We are able to separately identify age, cohort, and year effects because we observe workers of different ages in the same cohort of graduation and over time. I.e., age, cohort, and year do not form a collinear relation.

In theory, all universities could have an increasing college premium that aggregates into a decreasing overall college premium. Table 4 presents the correlation between the university fixed effect from Equation 1 with university rankings (where #1 is the best). The table shows that a move in ten positions in the university ranking is associated with a wage difference of 2%. For example, students from universities ranked in the 1st place receive 20% higher wages than students from universities ranked in 100th place. The point is that all these universities could have increasing trends in the college premium but at different levels. However, the number of students graduating from lower ranked universities (and lower fixed effect) has increased. As a result, the overall college premium gives stronger weight to workers from lower ranked universities by the end of the period, showing a decreasing trend.

V. Conclusion

We presented evidence that the college premium in Brazil has increased, opposing previous results in the literature. The difference in results comes from the construction of a new dataset that identifies worker's university. We find that the college wage premium increased by 23% between 2007 and 2018 when holding constant the set of universities for which we have data and decreases by 12% in the overall sample. In addition, we showed that the supply of workers with college degree has significantly increased, but much of this increase came from newer, lower ranked and lower wage-premium universities.

The results are relevant for the estimation of labor demand elasticities and to calculate the importance of skill biased technological change. Future research should try to include measures of skill in such models in order to account for changes in skill composition of workers in the same schooling group. The results also inform individuals and policymakers regarding the decision to invest in higher education. That said, future research should focus on verifying if these trends are similar for all universities in Brazil and in other countries.

References

Acemoglu, D. (2002, March). Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature 40*(1), 7–72.

Acemoglu, D. and D. Autor (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4 of *Handbook of Labor Economics*, Chapter 12, pp. 1043–1171. Elsevier.

Alvarez, J., F. Benguria, N. Engbom, and C. Moser (2018, January). Firms and the Decline in Earnings Inequality in Brazil. *American Economic Journal: Macroeconomics 10*(1), 149–89.

Bank, W. (2019). *World Development Report 2019: The changing nature of work*. The World Bank Group.

Barros, R., M. de Carvalho, S. Franco, and R. Mendonça (2010). Markets, the state, and the dynamics of inequality in Brazil. In L. F. López-Calva and N. Lustig (Eds.), *Declining inequality in Latin America: A decade of progress?*, pp. 74–134. Brookings Institution Press.

Camacho, A., J. Messina, and J. P. Uribe (2017, February). The Expansion of Higher Education in Colombia: Bad Students or Bad Programs? Documentos CEDE 015352, Universidad de los Andes - CEDE.

Card, D. and T. Lemieux (2001). Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis. *The Quarterly Journal of Economics 116*(2), 705–746.

Ciccone, A. and G. Peri (2005). Long-Run Substitutability Between More and Less Educated Workers: Evidence from U.S. States, 1950-1990. *The Review of Economics and Statistics 87*(4), 652–663.

Fernandez, M. S. and J. Messina (2018). Skill premium, labor supply, and changes in the structure of wages in Latin America. *Journal of Development Economics 135*(C), 555–573.

Ferreira, F., S. Firpo, and J. Messina (2017). Ageing Poorly? Accounting for the Decline in Earnings Inequality in Brazil, 1995-2012. IZA Discussion Papers 10656, Institute of Labor Economics (IZA).

Freeman, R. B. (1976). *The Overeducated American*. San Diego: Academic Press.

Goldin, C. and L. Katz (2008). *The Race between Education and Technology*. Harvard University Press, Cambridge.

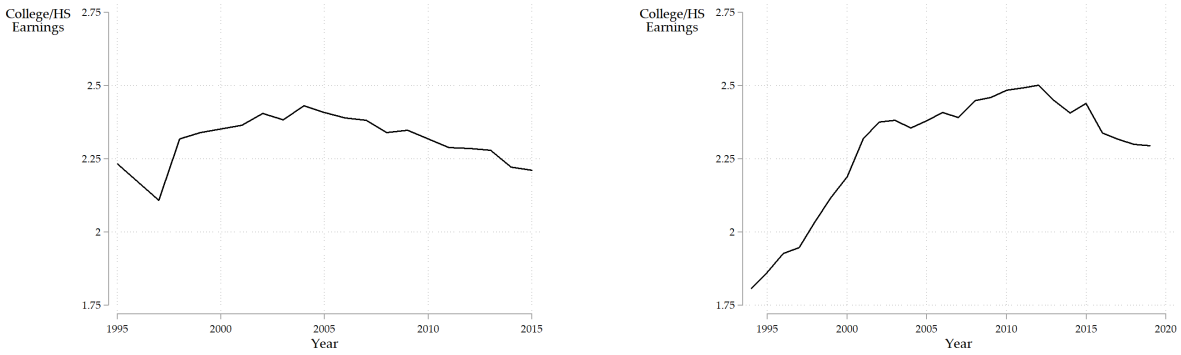
Haanwinckel, D. (2020). Supply, Demand, Institutions, and Firms: A Theory of Labor Market Sorting and the Wage Distribution. Unpublished.

Katz, L. and K. M. Murphy (1992). Changes in Relative Wages, 1963–1987: Supply and Demand Factors. *The Quarterly Journal of Economics 107*(1), 35–78.

Messina, J. and J. Silva (2017). *Wage inequality in Latin America: Understanding the past to prepare for the future*. The World Bank Group.

Rodriguez, J., S. Urzua, and L. Reyes (2016). Heterogeneous Economic Returns to Post-Secondary Degrees: Evidence from Chile. *Journal of Human Resources 51*(2), 416–460.

VI. Figures and Tables



((a)) PNAD ((b)) RAIS

Figure 1: College/High School wage premium

Note: The graph plots the ratio between wages of workers with college degree and workers with high school degree adjusted for age composition. The sample is restricted to workers between 21 and 65 years old. Panel A uses the Brazilian household survey (PNAD). Panel B uses the Brazilian matched employer-employee data (RAIS).

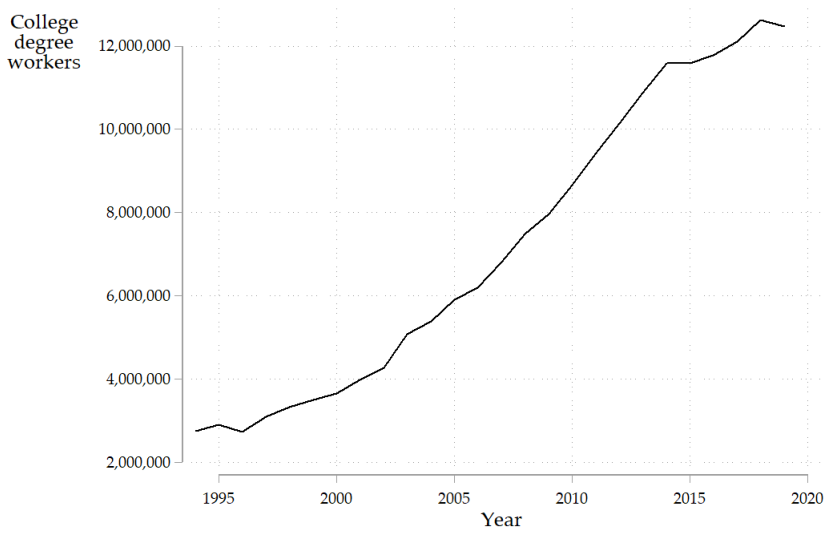


Figure 2: Formal workers with a college degree

Note: The figure plots the trends of the number of formal job relations with positive wage in which workers had a college degree. Data: RAIS 1994-2019.

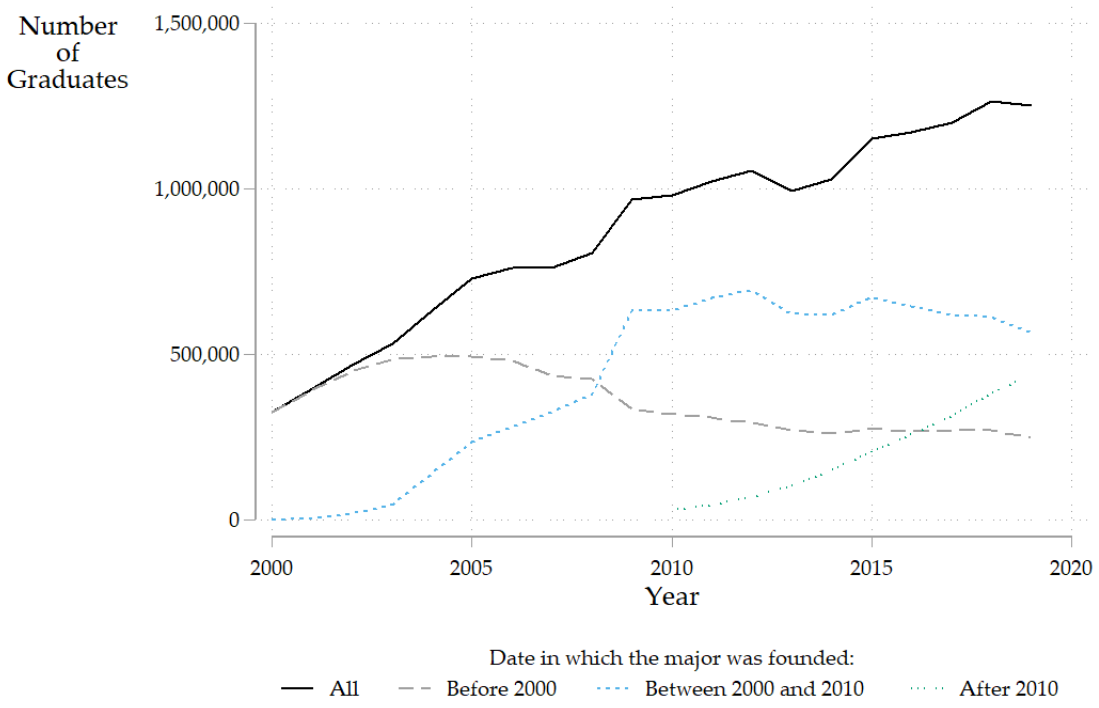


Figure 3: Number of graduates by major's date of foundation

Note: The figure presents the number of students that graduated in each year. Categories are defined by the year in which the major was founded. Source: Higher Education Census (2000-2018)

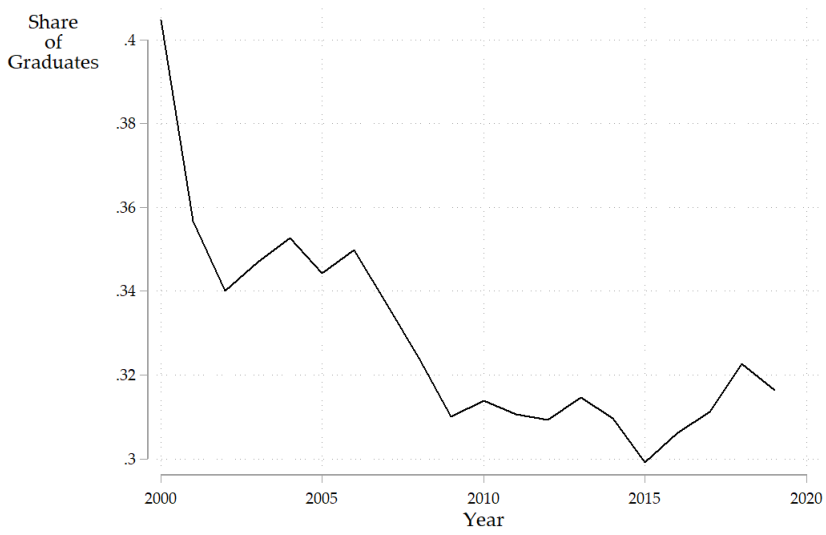


Figure 4: Share of graduates from Top 25 Universities according to ENADE ranking

Note: The number of graduates from each university comes from the higher education census.

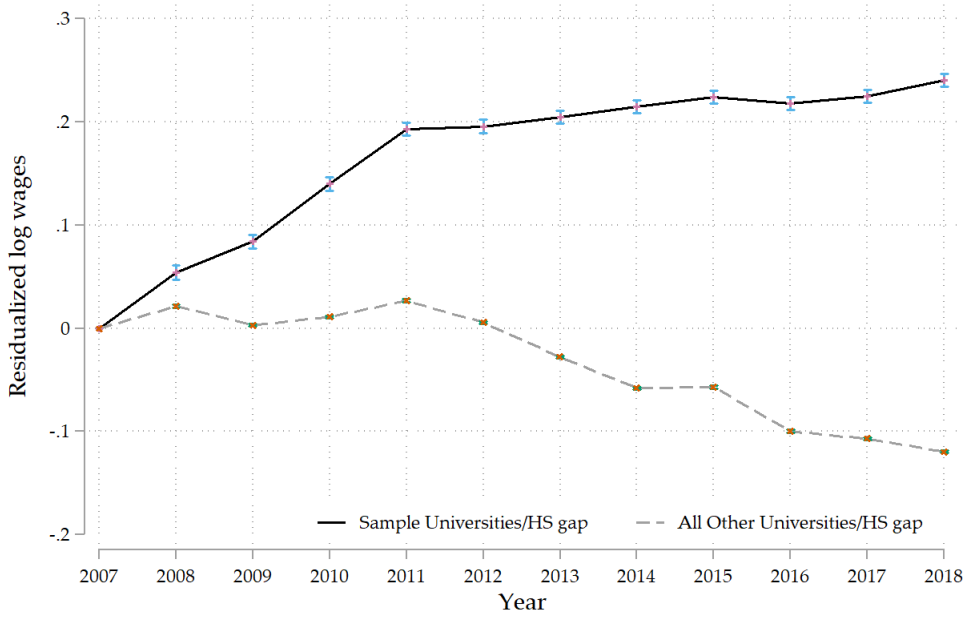


Figure 5: Trends in residualized log wages (college premium)

Note: The figure plots the evolution of the college premium, relative to 2007 values, for the college graduate's sample from FOIL requests (Sample Universities) and for the sample of all workers in the RAIS dataset with a college degree, excluding workers in the Sample Universities (All Other Universities). Both curves are relative to the same trends in the high school residualized wages. Estimates come from the estimation of Equation 1 over a sample of 263,181,905 observations. The figure plots 95% confidence intervals for each point estimate.

Table 1: University's age

University's Age	Sample	Out of Sample
$<$ 30 years	10.80%	31.20%
30 to 50 years	29.70%	33.10%
$>$ 50 years	59.50%	35.70%
N	37	157

Note: The first column refers to the universities from the college graduates' sample. The second column refers to all other universities in the RUF ranking.

Table 2: Institutions ranking and score

	Mean	S.D.	Median	Min	Max	N
RUF score						
Sample	69.6	21.7	42.1	4.8	98	37
Out of sample	43	21	74	4.2	97	157
RUF ranking						
Sample	48.3	48.7	33	3	197	37
Out of sample	110.2	52.6	111	1	196	157
Web ranking						
Sample	52.7	50.4	37	3	219	41
Out of sample	662.5	361	663.5	1	1285	1244

Note: RUF range between 0 and 100. The RUF ranking includes 194 universities and the Web ranking includes 1,285 universities.

Table 3: Association between university's age and quality

Dependent variable:	ENADE score			RUF	
	All	Specific knowledge	General knowledge	Ranking	Score
	(1)	(2)	(3)	(4)	(5)
Year in which the institution's first major was founded:					
$1940 < year < 1960$	-2.948***	-3.211**	-2.160*	46.12***	-19.85***
	(1.075)	(1.261)	(1.175)	(10.21)	(3.953)
$1960 < year < 1980$	-5.138***	-5.294***	-4.671***	72.66***	-31.13***
	(0.857)	(1.006)	(0.937)	(8.735)	(3.336)
$1980 < year < 2000$	-6.299***	-6.248***	-6.452***	97.31***	-37.90***
	(0.867)	(1.017)	(0.948)	(13.86)	(5.815)
$2000 < year < 2020$	-5.863***	-5.654***	-6.490***	112.2***	-43.10***
	(0.818)	(0.959)	(0.894)	(13.54)	(6.301)
Constant	49.32***	46.04***	59.13***	41.76***	74.76***
	(0.796)	(0.933)	(0.870)	(7.174)	(2.700)
Observations	1,469	1,469	1,469	197	171
R-squared	0.046	0.030	0.058	0.364	0.402

Note: Standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Table 4: Correlation between University ranking and wages

Dependent Variable:	University fixed effect	
	(1)	(2)
Ranking RUF	-0.0016**	
	(0.0008)	
Ranking Web		-0.0021**
		(0.0008)
Observations	33	36

Note: The dependent variable is the fixed effects from the estimation of Equation 1, with the exception that we include university fixed effects and do not include major by university fixed effects. Standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

A Appendix Figures and Tables

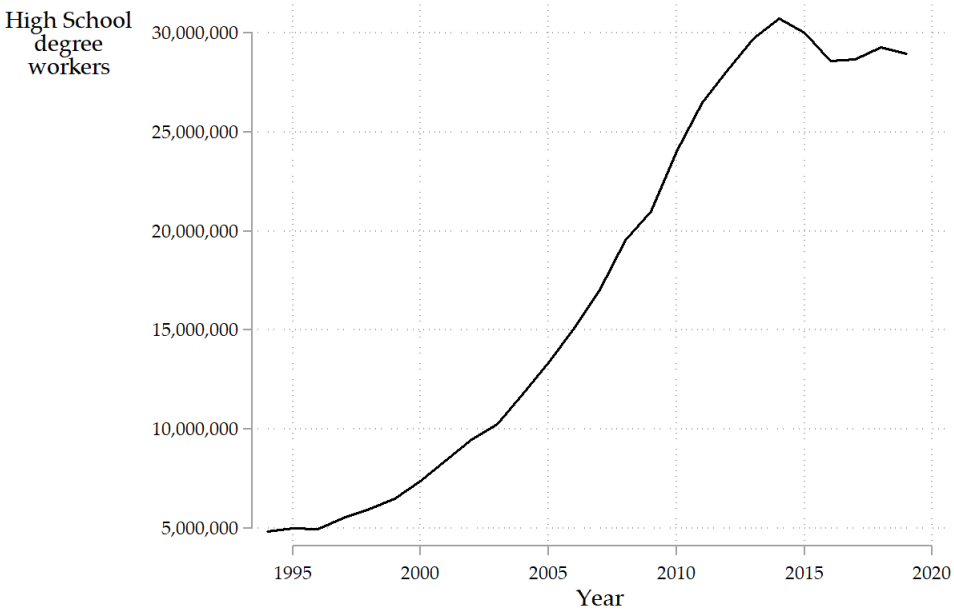


Figure A.1: Formal workers with a high school degree

Note: The figure plots the trends of the number of formal job relations with positive wage in which workers had a high school degree. Data: RAIS 1994-2019.

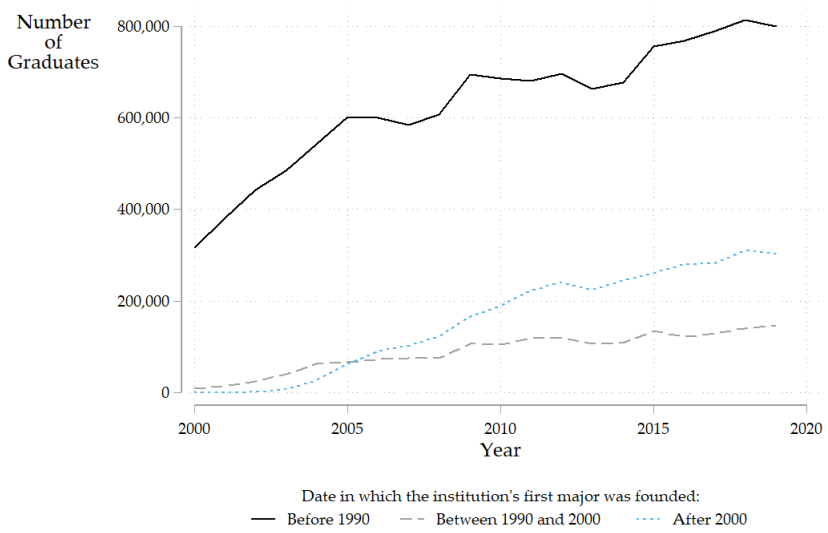


Figure A.2: Number of graduates by institution's date of foundation

Note: The figure presents the number of students that graduated in each year. Categories are defined by the year in which the institution's first major was founded. Source: Higher Education Census (2000-2018)

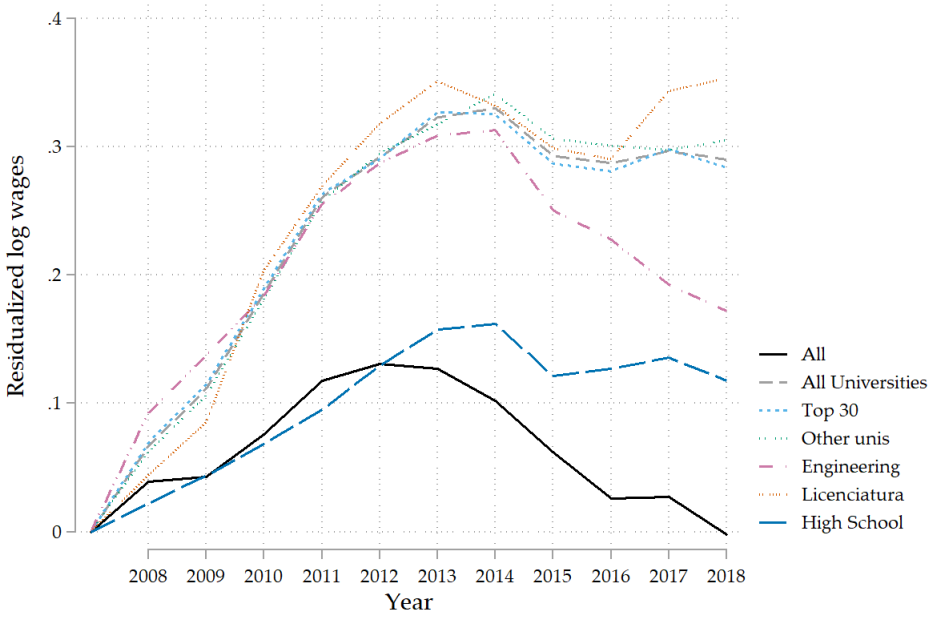


Figure A.3: Trends in residualized log wages

Note: The figure plots the evolution of earnings, relative to 2007 values, for different samples. Estimates come from the estimation of Equation 1 over a sample of 263,181,905 observations.

Table A.1: Schools with access to the data

Universidade/ Instituto	Acronym
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca	CEFET-RJ
Fundação Universidade do Amazonas	UFAM
Fundação Universidade Federal de Mato Grosso	UFMT
Fundação Universidade Federal de Ouro Preto	UFOP
Fundação Universidade Federal de Pelotas	UFPel
Fundação Universidade Federal de Rondônia	UNIR
Fundação Universidade Federal de Roraima	UFRR
Fundação Universidade Federal de São João Del Rei	FUNRei
Fundação Universidade Federal de Sergipe	UFS
Fundação Universidade Federal do ABC	UFABC
Fundação Universidade Federal do Acre	UFAC
Fundação Universidade Federal do Maranhão	UFMA
Fundação Universidade Federal do Tocantins	UFT
Fundação Universidade Federal do Vale do São Francisco	UNIVASF
Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro	IFRJ
Instituto Federal de Educação, Ciência e Tecnologia Fluminense	IFF
Instituto Militar de Engenharia	IME
Universidade de Brasilia	UNB
Universidade do Estado do Rio de Janeiro	UERJ
Universidade Estadual de Maringa	UEM

Table A.1: Schools with access to the data (cont.)

Universidade/ Instituto	Acronym
Universidade Estadual Paulista Júlio de Mesquita Filho	UNESP
Universidade Federal de Alagoas	UFAL
Universidade Federal de Alfenas	UNIFAL
Universidade Federal de Campina Grande	UFCG
Universidade Federal de Goiás	UFG
Universidade Federal de Itajubá	UNIFEI
Universidade Federal de Juiz de Fora	UFJF
Universidade Federal de Lavras	UFLA
Universidade Federal de Minas Gerais	UFMG
Universidade Federal de Santa Catarina	UFSC
Universidade Federal de Santa Maria	UFSM
Universidade Federal de Uberlândia	UFU
Universidade Federal de Vicosa	UFV
Universidade Federal do Ceará	UFC
Universidade Federal do Espírito Santo	UFES
Universidade Federal do Estado do Rio de Janeiro	UNIRIO
Universidade Federal do Pará	UFPA
Universidade Federal do Rio de Janeiro	UFRJ
Universidade Federal do Rio Grande do Norte	UFRN
Universidade Federal do Sul e Sudeste do Pará	UNIFESSPA
Universidade Federal Rural do Rio de Janeiro	UFRRJ
Universidade Tecnológica Federal do Paraná	UTFPR
Total	42

Table A.2: Sample size by university

University	Students	Matched to RAIS	% Matched
CEFET-RJ	30,446	23,029	75.60%
FUNRei	2,971	2,370	79.80%
IFF	16,387	10,682	65.20%
IFRJ	1,896	1,336	70.50%
IME	2,558	2,042	79.80%
UEM	53,160	40,605	76.40%
UERJ	72,562	56,330	77.60%
UFA	439	331	75.40%
UFAL	37,418	28,050	75.00%
UFABC	2,271	1,864	82.10%
UFAM	2,544	1,975	77.60%
UFC	65,155	51,206	78.60%
UFCG	32,647	23,678	72.50%
UFES	49,408	39,269	79.50%
UFG	46,335	36,233	78.20%
UFJF	31,459	23,471	74.60%
UFLA	13,231	9,480	71.60%
UFMA	1,851	1,511	81.60%
UFMG	99,656	75,839	76.10%
UFMT	45,547	35,156	77.20%
UFOP	27,024	21,423	79.30%
UFPA	47,025	35,051	74.50%
UFPel	24,189	16,107	66.60%
UFRJ	105,380	78,696	74.70%

Table A.2: Sample size by university (cont.)

University	Students	Matched to RAIS	% Matched
UFRN	81,761	50,166	61.40%
UFRR	325	259	79.70%
UFRRJ	13,395	10,003	74.70%
UFS	37,705	29,139	77.30%
UFSC	53,713	41,741	77.70%
UFSM	7,272	5,457	75.00%
UFT	1,537	1,101	71.60%
UFU	25,170	13,893	55.20%
UFV	26,383	18,231	69.10%
UNB	62,106	45,813	73.80%
UNESP	3,077	2,639	85.80%
UNIFAL	10,573	6,766	64.00%
UNIFEI	6,553	5,070	77.40%
UNIFESSPA	4,701	3,398	72.30%
UNIR	18,813	15,192	80.80%
UNIRIO	19,464	14,887	76.50%
UNIVASF	4,405	3,138	71.20%
UTFPR	28,928	23,793	82.20%
Total	1,217,440	906,420	74.50%

Table A.3: Comparing Matching Algorithms

Algorithm	Hyper Parameters		Algorithm Quality: Training sample (50%)		Algorithm Quality: Test sample (50%)	
	b1	b2	PPV	TPR	PPV	TPR
Logit	0.1	1.25	87.10%	92.00%	87.20%	92.20%
	0.05	1	86.70%	92.50%	87.10%	93.00%
Random Forest	0.1	1.25	96.60%	97.20%	88.80%	90.30%
	0.05	1	96.00%	98.70%	86.70%	92.70%

Note: Hyper parameters b1 and b2 are the threshold for whether the match's score is sufficiently large and the threshold for whether the ratio between the best and the second-best scores is sufficiently large, respectively. Positive Predictive Value (PPV or Accuracy): number of true positives over total number of positives. True Positive Rate (TPR or Efficiency): number of true positives over total number of correct cases.

Table A.4: Confusion Matrix—Out of Sample Predictions

Algorithm Prediction	True Status		Total
	False	Correct	
Not Matched	13,613	691	14,304
Matched	1,362	8,900	10,262
Total	14,975	9,591	24,566

Note: The table presents the sample sizes from the logit model with b1=0.05 and b2=1.1. Positive Predictive Value (PPV or Accuracy): number of true positives over total number of positives = 8900/(8900+1362) = 86.7%. True Positive Rate (TPR or Efficiency): number of true positives over total number of correct cases = 8900/(8900+691)= 92.8%.