An Analysis of Selected Labor Market Outcomes of College Dropouts in Germany – A Machine Learning Estimation Approach

Julia Heigle and Friedhelm Pfeiffer

Research Report

Acknowledgement: This report is part of the ZEW project "Analyses of Costs and Returns from Major Changes and Dropout from University". We gratefully acknowledge generous funding from the Federal Ministry of Education and Research (grant ID: 01PX16018A). We furthermore thank Francesco Berlingieri, Sarah Coyne and Holger Stichnoth and participants of the LERN annual scientific meeting in Nuremberg in April 2019 for helpful comments. The views expressed in this study are those of the authors, and do not necessarily reflect the views of the Ministry. All errors are our own.

Mannheim, 18th of June 2019

ZEW – Leibniz Centre for European Economic Research (ZEW)
Contact: PD Dr. Friedhelm Pfeiffer
Research Area Labour Markets and Human Resources
L 7, 1, 68161 Mannheim, Germany
Post Address: P.O.Box 10 34 43, 68034 Mannheim, Germany
Tel. +49 (0) 621/12 35 150 · Fax +49 (0) 621/12 35 225
E-mail friedhelm.pfeiffer@zew.de · wwww.zew.de

Content

Figures

Tables

1 Introduction

In the last two decades, there has been an increasing number of students entering the tertiary education system in Germany[1]. More specifically, there is a significant expansion of enrollment in universities. However, not all students graduate successfully. An increasing number of students (an absolute terms) leave the universities without a degree. [2] In this study, we investigate the labor market effects induced by dropping out of college without a degree. In 2016, a share of 78.3% of employed individuals with a college entrance qualification, and aged between 25 and 65, has been enrolled in college once in their lifetime. Amongst these individuals, 20.5% had dropped out of college without obtaining a formal degree.[3]

In 2014, the German Federal Ministry of Education and Research launched (in addition to existing initiatives for the prevention of college dropout, and for an increase in match-quality between the field of study and individual expectations) an initiative which aims to motivate college dropouts to attend vocational training. In addition, several projects were implemented in 2015 aimed at informing firms about the potential of college dropouts who are considered to be a target group for the recruitment of skilled labor.[4] Potential advantages of hiring college dropouts are emphasized, such as a lower likelihood of dropping out of vocational training, higher skills and experience. The ministry gives policy advice targeted at small and medium sized firms, regarding how to attract college dropouts. Calculations based on the SOEP data yield that in 2016, a 16.6% share of employed college dropouts aged between 25 and 65 do not obtain a vocational degree following the dropout. For employed individuals with a college entrance qualification who have never been to college, this share is 4.6%. For this paper, what is of particular interest is a detailed comparison between the

[1] See *Autorengruppe Bildungsberichterstattung (2018)*.
[2] See *Heublein et al. (2014)* for an analysis of the evolution of dropout rates.
[3] Source: own calculations, based on the German Socio-Economic Panel (SOEP).
[4] See https://www.bmbf.de/de/neue-chancen-fuer-studienabbrecher-1070.html for further details.

labor market prospects of college dropouts, and those individuals who possess a college entrance qualification but no college experience.

The focus of the (longitudinal) study described in the following is the investigation of the main effects of college dropout on employment, wages, and occupational position, using a data set consisting of employed individuals aged between 25 and 65 in 2016 who possess college entrance qualification. In this paper, selected labor market outcomes are compared exploiting a conditional-on-observables identification strategy. Whereas some studies to date have examined the effect of college dropout on occupational position and unemployment experience in Germany, relatively little is known about the effects of college dropout on wages. To the best of our knowledge, there is currently no study available that analyzes the wage differential between college dropouts and individuals with college entrance qualification but no college experience in Germany. A rich set of covariates is used for the analysis, including information on an individual's socio-economic background, academic achievement, personality traits, the type and the quality of college entrance qualification, and personnel and household characteristics.

The results indicate that college dropouts aged between 25 and 65 do, in expectation, not experience significant losses in terms of hourly wages. Furthermore, in terms of expectations, college dropouts end up in occupations with higher occupational prestige scores relative to individuals with a college entrance qualification but no college experience. There seem to be no significant differences in employment status between the two groups. A further descriptive analysis shows that college dropouts are more likely to end up in smaller firms.

The rest of the paper is structured as follows. Section 2 reviews the relevant literature on college dropout. Section 3 describes the data set used for the analysis. Section 4 discusses the integration of machine learning techniques into the causal inference framework. Section 5 introduces the method used for treatment effect estimation. In Section 6 treatment effect estimation results are presented for hourly wages and occupational prestige scores, and a multinomial logit model is estimated to investigate the relationship between employment and treatment group status. Section 7 concludes by critically discussing the empirical estimation strategy.

2 Related Literature

Whereas relatively little research investigates the labor market prospects of college dropouts, a comparably large amount of literature investigates the determinants of college dropout. Amongst others, *Aina et al. (2018)* provide a review of the theoretical and empirical literature on the determinants of college dropout. In their survey the determinants of college dropout are classified into four categories. The categories are given by the students' characteristics, abilities and behavior, the parental background and family networks, characteristics of the education system and institutions, as well as the labor market conditions. The most important determinants included in the first category, per the literature, are ethnicity, age at college enrollment, social interaction abilities and final high school grades. The most frequently analyzed determinants associated with parental background are the parents' respective level of education and occupation, both of which also represent good proxies for family income. Characteristics of the education system and institutions include, amongst others, availability of financial aid for students, the provision of student services and the application of admissions criteria. Regarding labor market conditions, a rise in unemployment rates is the most common focus of previous research.

Heublein et al. (2017) present the results of a representative survey of college dropouts in Germany for the year 2014, where the main focus is determinants of college dropout. The most prominent reasons for a college dropout given are intractable study requirements and a lack of prerequisites for the field of study, the inability to identify with the chosen subject of study, and the desire to focus more on practical activities. *Müller et al. (2013)* investigate the impact of pretertiary education pathways and social origin on dropout rates for Germany, using data from the National Educational Panel Study (NEPS). They find that individuals with a direct pathway have significantly lower dropout rates than individuals that first attained a vocational qualification before they enrolled in college.

Only a few studies investigate the effect of college dropout on labor market outcomes. Three popular theories provide hypotheses about the potential effects of college dropout on labor market outcomes. *Human capital theory*

(*Becker, 1962*) assumes that individuals accumulate human capital through education, which then raises their productivity. According to human capital theory, individuals profit from college experience even if they drop out before attaining a formal college degree due to the accumulation of human capital. Signaling theory (*Spence, 1978; Stiglitz, 1975*), in contrast, assumes that firms recruit candidates for a job under uncertainty; as the productivity of the candidates is not observable. Therefore, firms screen candidates based on observable characteristics, the so-called signals, like educational attainment, work experience, and any spells of unemployment. Educational attainment serves as a signal for the productivity of the candidates. According to signaling theory, the event of dropping out can constitute either a positive or a negative signal. If it is assumed that enrolment in college itself is considered to be a positive signal for the firm (*Arrow, 1973*), then college dropouts might have better labor market prospects (even without successful graduation) relative to comparable individuals who have never attended college, and gained no post-secondary education. A negative signaling effect might arise if dropping out is perceived to signal failure, or a lack of ability (*Heckman and Rubinstein, 2001*).

According to the *credentialism theory* (*Collins, 1979*), college dropout is expected to have a negative effect on labor market outcomes since graduation certificates essentially determine the attainable occupational positions of individuals. The credentialism theory is a sociological theory, which tries to explain educational expansion by status competition between groups. Studies have shown that credentials are more important the higher the linkage between vocational education and the labor market (*Allmendinger, 1989)*.

Schnepf (2015) discusses four major labor market characteristics that are likely to enhance the labor market prospects of college dropouts, and conducts a cross-country comparison of the labor market prospects of college dropouts between several European countries. She concludes that labor markets with a low percentage of college graduates, a high share of upper secondary school graduates who pursue vocational training, low participation of employers in vocational training, and high degrees of flexibility provide the best prospects for college dropouts.

Empirical studies seem to find, in general, that college dropouts have worse labor market prospects than college graduates but better prospects than those

who have never enrolled in college. For instance, *Davies and Elias (2003)* analyze the effect of college dropout on unemployment risk for UK tertiary-level education dropouts for the years 1996/97 and 1998/99. They find that dropouts are twice as likely to be unemployed during the year following tertiary-level education withdrawal in comparison to college graduates. However, dropouts seem to end up in occupations that are linked to their field of study, and thus dropouts achieve similar earnings compared to tertiary-level education graduates. *Johnes and Taylor (1991)* find that tertiary-level education dropouts in the UK who entered university in 1979 or 1980 faced significantly longer spells of unemployment in the years 1986-1988 compared to tertiary-level education graduates. *Matković and Kogan (2014)* demonstrate that the longer an individual stays at college in Serbia before dropping out, the higher their occupational status.

Using the 2011 Programme for the International Assessment of Adult Competencies (PIAAC) survey, *Schnepf (2015)* shows that a higher share of college dropouts is in high level positions in the labor market compared to individuals with an upper secondary education degree, but no college experience. However, after propensity score matching this difference turns out to be insignificant. *Scholten and Tieben (2017)* do not find evidence for the hypothesis that college dropouts with a vocational degree are expected to obtain a higher occupational position in their first stable job when compared to college dropouts without a vocational qualification.

A labor market outcome that is of particular interest for this paper is wages. For the US, several studies have found evidence that students who attend postsecondary education, without earning a degree, benefit from their studies in terms of future earnings, in comparison to individuals without any postsecondary education (*Bailey et al., 2004; Grubb, 2002*). Thus far, however, there are only a limited number of studies for European countries that investigate the effect of college dropout on wages. For some European countries (particularly European countries with multiple education tracks, such as Germany or France), credentials might be of higher importance in comparison to the US (*Card, 1999)*. *Johnes and Taylor (1991)* find significant wage benefits for UK higher education graduates relative to comparable higher education dropouts. *Reisel (2013)* shows that Norwegian college dropouts obtain lower incomes than upper secondary education graduates without any college experience.

3 Data

3.1 Data Preparation

For the analysis of the effects of college dropout on labor market outcomes, we construct a data set using the 2016 wave of the German Socio-Economic Panel (SOEP). The SOEP is a representative longitudinal study of private households in Germany that started in 1984. Each year around 30,000 individuals in nearly 11,000 households are interviewed. A sample is extracted that contains individuals aged between 25 and 65 who possess a college entrance qualification (either a so-called “Fachhochschulreife” or the “Abitur”).

The analysis considers three treatment groups in total: individuals that have successfully completed a college degree, individuals that enrolled in college but dropped out before having attained a formal degree, and individuals that in principle fulfill the formal requirements to attend college, but have not enrolled thus far.

College dropouts are identified using two sources of information in the SOEP. The first source is the biographical information stored in the file PBIOSPE. The spell types are restricted to spells with the content “School/College”. In order to distinguish school pupils from college students, only spells that start at age 18 and/or end earliest at age 21 are considered. The second source is the individual information on the first entry into tertiary education/ first exit from tertiary education obtained via the personal questionnaires[5].

For each labor market outcome of interest (employment status, wages, occupational prestige score) we construct a separate data set. Whereas the data sets for the labor market outcomes ‘employment status’ and ‘occupational prestige score’ require only the observability of the outcome variable (and the observability of several control variables), the data set for the labor market outcome wages is constructed based on the following procedure. The sample contains individuals who are employed at the time of interview in 2016, and who have not yet participated in an apprenticeship nor attended college. Furthermore,

[5] The information is stored in the file BIOEDU.

observations with missing or zero values in gross or net hourly wages are discarded. To exclude extreme values in the monthly working hours and the hourly gross wages, the smallest and the largest 1% of the data values are trimmed. First, trimming is conducted on the working hours and thereafter on the hourly gross wages[6].

The resulting sample (which is called the basic sample in the following) comprises 3,592 individuals[7]. An 82.1% share of the individuals with a college entrance qualification have been enrolled in college at least once. The dropout rate, which we define as the number of college dropouts (487) divided by the number of individuals that have attended college at least once in their lifetime (487 + 2,460), is 16.5%. Table 3-1 describes in detail the composition of the basic sample, and furthermore shows the composition of the estimation sample that will be used for the treatment effect analysis. The estimation sample is further restricted as a high number of control variables are added to the data set, and information on the control variables is not available for every individual in the basic sample.

One can see that the composition within the estimation sample changes slightly. The estimation sample comprises of a slightly higher share of individuals without college experience (21.7%) and a higher share of college dropouts (16.1%), which is due to the fact that individuals in these two groups have a higher response rate compared to college graduates. The estimated dropout rate of 20.5% therefore also differs from the estimate of the basic sample.

[6] The hourly wages are computed using the information on actual/agreed upon weekly working hours and the gross/net income of the last month stemming from the individual questionnaires of the SOEP survey. The hourly wage is thus the monthly (gross/net) income divided by the working hours multiplied by a factor of 4.33. The working hours correspond to the actual working hours if available in the data and to the agreed upon working hours in all other cases.

[7] Some individuals (N = 725) in the sample report a college degree but have not reported a college entrance qualification. We assume in these cases that the information on the college degree is valid and count these individuals as college graduates.

Table 3-1: Sample Composition within Basic and Estimation Sample

	Basic Sample		Estimation Sample	
	obs.	percentage	obs.	percentage
no college experience	645	17.96%	349	21.66%
college dropouts	487	13.56%	259	16.08%
college graduates	2,460	68.49%	1,003	62.26%
total	3,592	100%	1,611	100%

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, currently employed individuals between 25 and 65 years; own calculations. The basic sample is constructed imposing only some minor restrictions on the wages whereas the estimation sample puts restrictions on the observability of several control variables.

3.2 Control Variables

For the estimation of the effects of college dropout an informative set of control variables is needed. In the following we will present the control variables contained in the estimation sample. The data set includes information on the gender, age, migration background, and the socio-economic background of an individual. Variables characterizing the socio-economic background of individuals in the sample include the highest level of educational attainment of the parents, the age of the mother at the individual's time of birth, and the social status of the father, as measured by the International Socio-Economic Index of Occupational Status. Variables characterizing the current family background of the individual include the number of children in the household below age 8, the number of children in the household between ages 8 and 15, and the marital status.

The data set additionally includes information on academic achievement, given by the individual's grades in German, mathematics, and the first foreign language on the last report card. Furthermore, measures for the Big Five are derived. The Big Five include the following personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. A score is constructed for each of the five personality traits. Data on the Big Five was first collected in the individual questionnaires in 2005, followed by the years 2009 and 2013. Individuals are asked to respond to 15 statements; expressing their agreement with the respective statement by selecting a number on a Likert scale ranging

from 1 to 7 (where 7 corresponds to “does apply” and 1 to “does not apply”). Three questionnaire items were constructed for each trait[8]. The score for each personality trait (ranging from 3 to 21) is derived as the sum of the three individual scores. Thus, a higher score for a specific personality trait indicates that an individual is well-characterized by the respective personality trait. Information is available for different survey years. The final Big Five personality trait scores correspond to a simple average of the derived scores over all survey years.

The data set additionally takes into account the birth cohort of individuals, the federal state in which the last school was attended, as well as the type of the highest school degree obtained (“Fachhochschulreife” or “Abitur”). The birth cohort of individuals allows us to take into account time-specific factors that drive the decision to go to college, and time-specific differences in educational systems. Furthermore, birth cohorts allow us, to some extent, to control for differences in labor market entry conditions which affect labor market outcomes.

The results of the *INSM-Bildungsmonitor (2004-2016)* are used to partition the federal states according to the quality of their educational systems. The *INSM-Bildungsmonitor* is a study conducted on a yearly basis since 2004, which aims to evaluate the federal state-level education system within Germany by taking into account various performance indicators (93 indicators)[9]. The results of this study are used to construct a variable that indicates whether an individual obtained the highest school degree in a federal state which has consistently been in the top 5 of the ranking of the *Bildungsmonitor* since 2004, or in the lowest 5. Four federal states (Saxony, Thuringia, Baden-Wuerttemberg, Bavaria) consistently achieved high positions in the *Bildungsmonitor* ranking, whereas two federal states (North Rhine-Westphalia, Brandenburg) have a consistently low position in the rankings.

[8] Table 8-1 in the Appendix lists all the statements by personality trait.

[9] See https://www.insm-bildungsmonitor.de/ for more details on the *Bildungsmonitor*.

3.3 Descriptive Statistics

In the following, descriptive statistics are presented in order to gain first insights into the relationships between the control variables and the treatment group status.[10] The data set contains 349 individuals with college entrance qualification and no college experience, 259 college dropouts, and 1,003 college graduates. The group of individuals who have never been enrolled in college is termed the baseline group in the following. The type of college entrance qualification ("Abitur" or "Fachhochschulreife") varies substantially across treatment groups. An 84.2% share of the college graduates are endowed with the "Abitur", whereas the share is only 69.5% for college dropouts. For individuals without college experience the share is 62.5%. The remaining college graduates, college dropouts, and baseline group members are endowed with the so-called "Fachhochschulreife".

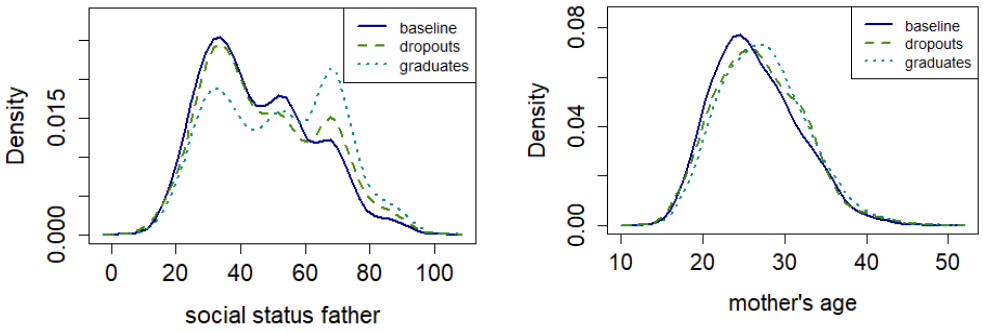
The fraction of individuals with a migration background also crucially depends on the respective treatment group under consideration. Only 9.7% of college graduates have a migration background, whereas the fraction is 15.4% for college dropouts. Regarding the socio-economic background of the individuals, we find that college graduates empirically tend to have a higher probability of having a father with a high socio-economic status according to the kernel density plots depicted in Figure 3-1. The distribution for college dropouts and individuals from the baseline group tend to be quite similar.

Table 8-5 in the Appendix depicts the results of the two-sided t-tests for the equality of means between treatment groups for several variables. The results indicate significant differences in the average social status of the father between the baseline group and the group of college graduates, and between the group of college dropouts and the group of college graduates. No significant difference in the average social status of the father is indicated between the baseline group and the group of college dropouts. The age of the mother at birth of the individual seems to be almost equally distributed for college graduates and dropouts (compare Figure 3-1). However, the kernel density graph

[10] Further summary statistics (overall and by treatment group) for the data set can be found in the Appendix in Table 8-3 and Table 8-4 as well as a description of variables in Table 8-2.

shows that the mothers of individuals in the baseline group tend to have given birth at a lower age compared to the other groups. According to the t-tests for the equality of means in Table 8-5 in the Appendix, there are significant differences in the means of the age of the mothers at birth of their child between the group of college graduates and the remaining two groups.

Figure 3-1: Socio-economic Background by Treatment Group



Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The baseline groups consists of individuals with a college entrance qualification who have never enrolled in college. The figure depicts the kernel density estimates for the distribution of the variables by treatment group. The left hand side gives the kernel density estimate for the social status of the father measured on a scale between 16 and 90, the right hand side gives the kernel density estimate for the mother's age at birth.

Table 3-2 depicts the types of the highest school degree attained by the individuals' parents by treatment group. It turns out that 22.3% of the mothers of college graduates attended schools which provide the college entrance qualification ("Gymnasium" or "Fachoberschule"), whereas only 17.4% of the mothers of college dropouts attended these schools. For the baseline group, the fraction is even lower at 12.9%. Similar results can be found for the highest school degree of the father. 35.7% of the fathers of college graduates attended either "Gymnasium" or "Fachoberschule". For college dropouts and the baseline group, however, the fractions are 29.0% and 22.1%, respectively. The results of the Pearson's chi-squared test leads to the rejection of the null hypothesis that treatment group status and parental education are statistically independent.

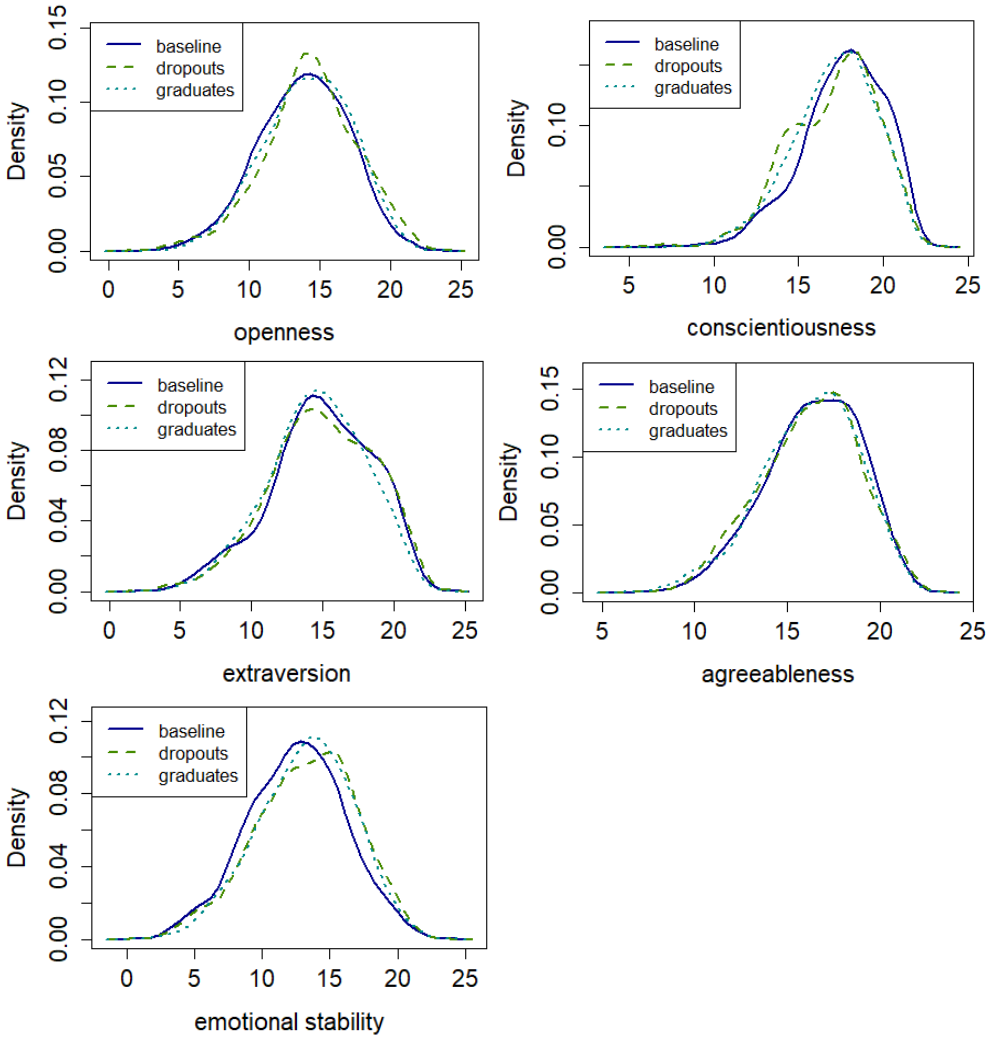
Table 3-2: Highest Parental School Degrees by Treatment Group

	Highest school degree of mother			Highest school degree of father		
	Baseline group	College dropouts	College Graduates	Baseline group	College dropouts	College Graduates
No school degree	2.01%	1.93%	1.30%	1.15%	3.47%	1.10%
Hauptschule	51.29%	45.46%	44.27%	50.72%	41.70%	41.48%
Realschule	33.81%	35.14%	32.10%	26.07%	25.87%	21.73%
Fachoberschule	0.29%	0%	0.60%	0.29%	0%	0.40%
Gymnasium	12.61%	17.37%	21.73%	21.78%	28.96%	35.29%
	$\chi^2(8) = 18.28$, p = 0.019			$\chi^2(8) = 33.13$, p = 0.000		

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The baseline group includes individuals with college entrance qualification who have never been enrolled at college. The last row depicts the results of Pearson's χ^2-test.

The investigation of kernel density estimates for the Big Five personality traits in Figure 3-2 yields that there do not seem to be substantial group differences in agreeableness. Table 8-5 in the Appendix also indicates no significant differences in average agreeableness scores between the treatment groups. Smaller differences can be detected for the remaining personality traits. For example, college dropouts and individuals of the baseline group tend to be slightly more extraverted than college graduates. However, according to Table 8-5 in the Appendix, the differences in average extraversion scores do not turn out to be significant. The baseline group seems to be more conscientious compared to college dropouts and graduates, who in turn tend to be more emotionally stable. The two-sided t-tests in Table 8-6 in the Appendix appear to confirm these findings.

Figure 3-2: Big Five Personality Traits by Treatment Group



Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The baseline group consists of individuals with a college entrance qualification who have never been enrolled in college. The figure depicts the kernel density estimates for the distribution of the various personality traits by treatment group.

Regarding the individual's school performance (see Table 3-3), we find that college graduates are endowed with the highest share of individuals who obtained grade 1 ("very good") or grade 2 ("good") on the last report card, independent of the subject under consideration. A comparison between college dropouts

and the baseline group yields that the baseline group has a higher share of individuals with grade 2 or better in the subjects German and in the first foreign language. However, 47.5% of college dropouts obtained very good or good grades in mathematics on the last report card, whereas only 41.6% of baseline group members did. The results of the Pearson's chi-squared test leads to a rejection of the null hypothesis that school grades on the last report card are statistically independent from treatment group status.

Table 3-3: School Grades on Last Report Card by Treatment Group

	German grade on last report card			Math grade on last report card			First foreign language grade on last report card		
	baseline	dropouts	graduates	baseline	dropouts	graduates	baseline	dropouts	graduates
1	8.9%	11.6%	16.5%	9.2%	13.1%	25.9%	9.5%	9.3%	18.2%
2	45.0%	40.2%	47.5%	32.4%	34.4%	35.3%	39.0%	34.8%	38.9%
3	40.7%	38.2%	28.8%	34.1%	27.0%	23.0%	38.4%	33.6%	29.8%
> 4	5.4%	10.0%	7.3%	24.4%	25.5%	15.8%	13.2%	22.4%	13.2%
	$\chi^2(8)$ = 34.59, p = 0.000			$\chi^2(10)$ = 77.20, p = 0.000			$\chi^2(10)$ = 46.60, p = 0.000		

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The baseline group contains individuals with a college entrance qualification who have never been enrolled in college. Grade 1 corresponds to "very good" and grade 4 to "sufficient". The German school grading system allows for grades until grade 6 "insufficient". The last row depicts the results of Pearson's χ^2-test.

3.4 Employment Path of College Dropouts

First, we describe differences in the acquirement of vocational degrees between the group of college dropouts and the baseline group. We find that only 16.6% of the college dropouts in the sample do not possess a vocational degree of any kind by the time the data was collected. An 83.4% share are endowed with a vocational degree at the interview date. Only 4.6% of individuals with a college entrance qualification who have never been enrolled in college have not completed a vocational degree, thus 95.4% do have a vocational degree. Table 3-4 shows in detail the shares of individuals for the various types of vocational degrees obtained by the group of college dropouts and the baseline group.

Table 3-4: Comparison of Highest Vocational Degrees Attained

	Highest vocational degree	
	Baseline group	College dropouts
no vocational degree	4.58%	16.60%
apprenticeship	56.45%	33.98%
vocational school	20.63%	20.08%
health care school	0.29%	1.16%
technical school	11.17%	18.53%
civil service training	2.87%	5.79%
other degree	4.01%	3.86%
	$\chi^2(7) = 49.35$, p = 0.000	

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The baseline group includes individuals with college entrance qualification who have never been enrolled at college. The last row depicts the results of Pearson's χ^2-test.

Major differences can be seen in the shares of individuals who completed an apprenticeship and the share of individuals that attended technical school. A greater share of college dropouts (18.5%) attended a technical school compared to individuals without college experience (11.2%). Technical Schools require either a completed vocational degree and/or work experience and qualify individuals for higher professional responsibility and management positions. Moreover, more than half of the individuals without college experience complete an apprenticeship (56.5%) whereas the share of college dropouts (34.0%) is lower. This finding can, presumably, be explained by the fact that college dropouts more often tend to obtain a degree from technical schools, which in many cases requires a completed apprenticeship. Pearson's chi-squared test indicates a statistical dependence between the attained vocational degree and the treatment group status.

Next, differences in labor market outcomes, like employment, firm size, wages, and occupational status, are examined, in particular between the group of college dropouts and the baseline group.

Table 3-5 shows the proportions of individuals not employed, partially or marginally employed, and full-time employed. Individuals are classified as not employed if they are not working, on maternity leave, or in a phased retirement

scheme with actual working hours currently zero. The group of college graduates has the smallest share of non-employed individuals, at 11.1%, and the smallest share of individuals who are part-time or marginally employed at 26.1%. The share of non-employed individuals for college dropouts (14.5%) and baseline group members (13.6%) is comparable.

Table 3-5: Employment Status by Treatment Group

	not employed	part-time/ marginally employed	full-time employed
baseline	13.62%	41.04%	45.34%
college dropouts	14.54%	30.61%	54.85%
college graduates	11.07%	26.06%	62.87%

$\chi^2(4) = 58.78, p = 0.000$

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. The category "not employed" comprises individuals who are not working, in maternity leave or in a phased retirement scheme with actual working hours being currently zero. Sample sizes are given by N=536 for the baseline group, N=392 for the group of college dropouts and N=1,815 for the group of college graduates.

The Pearson's chi-squared test, which tests the null hypothesis that treatment group status and employment status are statistically independent, yields to a rejection of the null hypothesis. Therefore, the relationship between treatment group status and employment status is investigated in greater detail by means of a multinomial logit model in Section 6.1.

Next, we will focus on differences in firm-size categories between the group of college dropouts and the baseline group. Firms are divided into 3 distinct categories based on size. Category 1 contains individuals employed at a firm with less than 200 employees as of 2016. Category 2 contains individuals employed by firms with 200 employees or more, but less than 2,000, and category 3 those with 2,000 or more employees. Self-employed individuals without co-workers are excluded from the analysis. Table 3-6 displays the shares of individuals within a certain firm-size category by treatment group. Empirically, we find that a smaller share of college dropouts (24.6%) end up at firms with more than 2,000 employees when compared to individuals that have never been enrolled

in college (27.8%). College graduates represent the largest share (37.2%) working in large firms. The majority of college dropouts (48.8%) can be found at small firms with less than 200 employees.

Table 3-6 shows in addition the results of Pearson's chi-squared test, which tests the null hypothesis that the variables firm-size category and treatment status are statistically independent. As the respective p-value is smaller than all usual significance levels, there is evidence that the two variables are not independent. Nevertheless, a more detailed analysis taking into account a set of control variables is not conducted due to the insufficient sample size for each partition of firm-size category and treatment group.

Table 3-6: Firm Size Categories by Treatment Status

	small	medium	large
baseline	47.00%	25.18%	27.82%
college dropouts	48.82%	26.60%	24.58%
college graduates	43.02%	19.79%	37.18%

$\chi^2(4) = 27.43, p = 0.00002$

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Sample sizes are given by N=417 for the baseline group, N=297 for the group of college dropouts and N=1,455 for the group of college graduates.

The empirical findings in Table 3-6 can be explained by signaling theory and might be explained by the fact that dropping out of college constitutes a negative signal (signal of failure, lack of ability) received by the labor market (*Heckman and Rubinstein (2001)*. Larger firms can presumably choose among a larger pool of applicants, as they are able to pay higher wages. Competition is particularly high for an individual who dropped out of college and competes against otherwise comparable individuals with a continuous educational/professional path. Due to this, it may be that college dropouts are ceteris paribus less likely to be employed in larger firms relative to baseline group members.

Table 3-7 shows the mean hourly gross and net wages by treatment group[11]. As differences between college dropouts and individuals without college experience are of major interest, a two-sided t-test for the equality of the means between the two groups is conducted. The p-value of the conducted test for the hourly gross wage indicates no significant difference, as the null hypothesis of the equality of the means cannot be rejected at usual significance levels. Nevertheless, the effect is very close to being significant at a 10% level. Consequently, it is possible that a larger sample size would lead to the detection of a significant effect. However, there is evidence for substantial differences in average hourly net wages. College dropouts tend to earn, on average, €0.63 more per hour relative to baseline group members. Possible reasons for the observed positive wage differential will be discussed in detail in Section 6.2.3.

Table 3-7: Mean Hourly Gross/Net Wage by Treatment Group

		mean	std. dev.	min	max
GROSS WAGE	no college experience	16.45	5.54	5.08	32.99
	college dropouts	17.18	5.71	5.54	32.33
		$\Delta = 0.73, \text{d.f.} = 606, \text{t} = 1.59, \text{p} = 0.11$			
	college graduates	22.13	6.74	4.81	35.41
NET WAGE	no college experience	10.94	3.24	3.70	21.72
	college dropouts	11.57	3.71	3.97	23.09
		$\Delta = 0.63, \text{d.f.} = 606, \text{t} = 2.23, \text{p} = 0.03$			
	college graduates	14.82	4.67	3.21	30.26

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Δ reflects the difference in mean hourly gross/net wages between college dropouts and individuals from the baseline group. Statistics of a two-sample t-test for the equality of means are given by: d.f. – degrees of freedom, t – t-value, p – p-value. The results are based on the basic sample (see Table 3-1).

[11] The same table for the weekly working hours by treatment group can be found in the Appendix in Table 8-6.

In addition, the autonomy level of individuals in their job is investigated. Three classes of autonomy levels are considered: low autonomy, medium autonomy and high autonomy. The low autonomy class contains manual workers, and workers in the production or service sector for which tasks require only a minimum level of qualification. Class 2 contains individuals with jobs that require the completion of the middle track of secondary education ("Mittlere Reife" or "Realschule"). The high autonomy class comprises jobs that require a degree from a college of applied sciences, or a college. Table 3-8 shows the degree of autonomy by treatment group. A higher share of college dropouts (27.5%) in the sample seem to be in a job with a high degree of autonomy compared to individuals who have never attended college (19.5%). The category with the lowest degree of autonomy contains 18.6% of the individuals without college experience, whereas only 13.9% of college dropouts are classified into this category. Self-employed individuals are either classified into category 2 or category 3, depending on their number of employees. According to Pearson's chi-squared test, there is evidence for a significant statistical dependence between autonomy level and treatment group status.

Table 3-8: Autonomy Level by Treatment Group

	low	intermediate	high
no college experience	18.55%	61.99%	19.46%
college dropouts	13.92%	58.54%	27.53%
college graduates	2.79%	24.50%	72.72%

$\chi^2(4) = 547.08, p = 0.0000$

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. D corresponds to the treatment level indicator. Sample sizes are given by N=442 for the baseline group, N=316 for the group of college dropouts and N=1,543 for the group of college graduates.

The effect of dropping out of college on wages and occupational prestige status (which is highly correlated to the occupational autonomy level) is investigated in more detail in Section 6.2 and Section 6.3, using a sophisticated machine learning procedure that allows us to control for observable individual characteristics.

The expression *effect of college dropout* refers throughout the entire paper to the effect of the pursued (educational) life path of college dropouts after secondary school before entering the labor market. It will be tested whether the observed differences in wages and occupational status are still existent after controlling for observable differences between individuals. The approach allows in addition to draw conclusions about the significance of the estimated effects.

4 Causal Inference with Machine Learning Techniques

The program evaluation literature is currently developing a roadmap for how machine learning techniques can be adequately adopted in order to infer a causal parameter of interest. Some promising approaches have been introduced which combine existing treatment effect estimation procedures and machine learning techniques. In the following, a review is given on the treatment effect estimation literature in the context of a conditional-on-observables identification strategy. The review summarizes possible estimation strategies for the assessment of the effect of college dropout on labor market outcomes. *Athey (2018)* gives a general overview on the contributions of machine learning methods to the economic literature.

The *classical (parametric) estimation* procedure in statistics is the following: a model is set up in the initial step, i.e. the econometrician assumes both that he knows all of the relevant control variables, and that they are all present in the data set. The number of relevant controls should be small relative to the sample size. Thus, the researcher needs to make a decision, choosing controls based on either economic theory or intuition, and in addition assumes a functional form via which the controls enter the model. Following the model specification step, data is collected to estimate the model and conduct causal inference in regard to the parameter of interest. The focus of the classical procedure is thus on causal inference. Classical treatment effect estimation procedures include, among others, regression imputation, propensity score weighting and doubly robust estimation procedures[12].

For instance, early *regression imputation* estimators rely heavily on specified models for the counterfactual outcomes, as missing potential outcomes are imputed by this method. The potential outcomes in the study presented in this report correspond to the wages/occupational prestige scores that can be obtained by an individual with specific characteristics given one of the three treatments under consideration, i.e. choosing to attend college and graduating from college, choosing to attend college, dropping out from college and thereafter

[12] *Imbens and Rubin (2015)* provide a summary of the (classical) treatment effect estimation approaches in conditional-on-observables settings.

attending vocational training or directly entering the labor market and choosing to pursue vocational training, or directly entering the labor market after school. Typically, researchers conduct an ad-hoc selection of control variables which they deem important.

Propensity score weighting procedures target a balance of the weighted distribution of covariates between treatment and control groups instead of relying on model-based imputations. Without machine learning techniques researchers typically select ex-ante the control variables they deem to be important for propensity score estimation. The propensity score is estimated, for instance, by means of a standard (parametric) logistic model using the set of selected variables. Finally, a consistent estimate of the average treatment effect can be computed (as long as the propensity score is correctly specified) by inverse probability weighting or blocking methods. The propensity score in our study describes the probability of belonging to a certain treatment group for an individual with given characteristics. Taking into account the knowledge that we have about the individuals up to the point at which they attained the college entrance qualification, we estimate the probability of pursuing each of the three possible educational life paths.

To weaken the requirements on correct model specification *doubly robust estimation* procedures have been introduced. Doubly robust estimation combines regression imputation and inverse probability weighting. Consistent estimates of average treatment effects can be obtained if either the propensity score or the conditional mean potential outcomes are correctly specified, or both. Nevertheless, researchers still select a set of control variables and define the functional forms based on a priori reasoning or based on trial and error.

Several extensions to the classical treatment effect estimation procedures have been proposed in the literature that allow for a more flexible estimation by imposing less parametric restrictions. These *semi-parametric estimation* methods include nonparametric kernel estimators and series estimators[13]. Although semi-parametric estimators allow the construction of more flexible models, they still meet severe limitations (particularly in high-dimensional settings, in

[13] *Imbens (2003)* reviews the literature on semi-parametric treatment effect estimation.

which the number of covariates exceeds the number of observations)[14]. Among other issues, they may suffer from the curse of dimensionality. The more covariates are considered in the nonparametric estimation, the lower the rate at which the bias vanishes. Therefore, nonparametric estimation still requires a manageable pre-selected set of variables. Similarly, series estimators (which are based on multiple generated variables, such as polynomials and splines) are not applicable in high-dimensional settings. In these settings it is necessary for the researcher to make a choice about the polynomials that are considered. Covariates are, in many cases, selected based on iterative searches.

At this point novel machine learning techniques come into play. A major advantage of *machine learning techniques* is that they make model specification redundant and allow for model selection instead. Machine learning tools provide a statistical data-driven way to select covariates, instead of performing iterative searches by hand. The data-driven automated covariate selection is based on statistical rules defined by the user.

Another benefit of machine learning tools is their applicability in high-dimensional settings in which the number of potential control variables exceeds the number of observations (which is not the case in our application). Data-driven selection of the most informative control variables makes a pre-selection of control variables unnecessary, and therefore requires less input on the part of the researcher in terms of her/his beliefs.

Thus, machine learning estimation can extend the semi-parametric estimation literature in the sense that nonparametric kernel estimators can be substituted by modern nonparametric machine learning estimators, such as Random Forest or Regression Trees, which perform a data-driven selection of the most informative control variables and are therefore applicable in high-dimensional settings. Furthermore, machine learning tools provide a useful way to extend series estimation, such as the LASSO enable a data-driven selection of polynomials and interaction terms.

[14] Data sets might either be inherently high-dimensional (big data) or artificially high-dimensional. Artificially high-dimensional data sets contain many generated regressors e.g. interaction terms or polynomials to permit more flexible models.

The main challenge of the algorithms is to find a (prediction) model that is flexible enough to capture the main signal of the data, but is not overfitted to the data in the sense that it is unable to generalize to independent data sets that were not used for estimation. The trade-off between accuracy and generalization capability is often taken into account by choosing so-called tuning-parameters in the machine learning prediction models in a way that the out-of-sample forecast performance of the prediction model is maximized[15].

To sum up, the *machine learning procedure* can be described as follows: it starts with the data collection step and does not require a model specification. Next, sophisticated machine learning tools provide a data-driven way to detect strong predictors for the outcome of interest, by looking for statistically informative patterns in the data. The algorithms select a specific model by imposing dimension reduction through the choice of a tuning parameter. The focus of machine learning techniques is on prediction rather than causal inference. Therefore, there are some challenges associated with the integration of machine learning tools into the traditional treatment effect estimation as perfect model selection through algorithms is a highly unrealistic assumption.

In the next section we will describe a machine learning procedure for the estimation of average treatment effects that can deal with small model selection mistakes. Machine learning tools are used to find high-quality approximations for the propensity score and the conditional mean potential outcomes. The final treatment effect estimates are robust to small model selection mistakes, and are based on a doubly robust estimator.

[15] *Friedman et al. (2001)* give an intuitive introduction into machine learning and the bias-variance trade-off.

5 Method

5.1 Double Machine Learning

To estimate the effect of different educational life paths on labor market outcomes, we apply a so-called double machine learning procedure for multivalued treatments introduced by *Farrell (2015)*. It is a *double* machine learning procedure in the sense that two models are estimated via machine learning tools. Thus, the final average treatment effect estimates are based on propensity score and conditional mean potential outcome estimates.

The multivalued treatments correspond in our case to the different educational life paths of the individuals. The treatment groups are individuals with a college entrance qualification who have never been enrolled in college, college dropouts, and college graduates. In the analysis, the (three) potential outcomes correspond to the potential wages or the potential occupational prestige scores for an individual with given characteristics that he or she could have obtained if they had received a certain treatment level. In the following, the multivalued treatment effect framework is formally introduced.

Let D denote the multivalued treatment variable which can take on $\tau + 1$ distinct values, i.e. $D \in \{0, 1, \dots, \tau\}$.[16] In our empirical application $\tau = 2$. The probability of obtaining a specific treatment level for given individual characteristics is described by the generalized propensity score, which is defined as $p_t(x) = Pr(D = t|X = x)$ where $t \in \{0, 1, \dots, \tau\}$ and X denotes the set of control variables. The $\tau + 1$ potential outcomes are denoted as $Y(t)$. Thus, the observed outcome can be computed according to $Y = \sum_{t=0}^{\tau} \mathbb{1}(D = t)Y(t)$. In our analysis the observed outcomes correspond to either wages or occupational prestige scores.

Sufficient conditions for the identification of the treatment effects are the mean independence assumption, which states that $E[Y(t)|D, X] = E[Y(t)|X]$. The assumption implies that, for each treatment group, we can set up a model for

[16] Notation: capital letters without index i denote random variables, capital letters with index i denote realizations of the random variables.

the potential outcomes using only the set of covariates X which sufficiently describe the variation in potential outcomes within treatment groups. We say 'sufficient' in the sense that there are no omitted covariates which could impact both the potential outcomes and the likelihood of being treated simultaneously. Another identification assumption is the overlap assumption, which requires $0 < \epsilon < \Pr(D = t|X = x)$ for some $\epsilon > 0$ and $\forall t \in D, \forall x \in \chi$ where χ denotes the realization set of the control variables. The overlap assumption requires there to be a non-zero probability of belonging to the various treatment groups for all possible realizations of the set of covariates X.

As already mentioned above, the machine learning procedure presented by *Farrell (2015)* is based on a doubly robust estimator for treatment effects which is sometimes also termed the augmented inverse probability weighting estimator[17]. The procedure has the appealing property that average treatment effect estimates remain consistent even when either the model for the propensity score or the model for the potential outcomes is parametrically misspecified, but not both. *Farrell (2015)* shows that this robustness property extends to model selection errors. Thus, if high-quality approximations for the propensity score and the conditional mean potential outcomes are used as plug-in estimates for the true functions, the average treatment effect can be consistently estimated under certain regularity conditions that will be discussed later. The conditional mean potential outcomes are conditional in the sense that potential outcomes are imputed using a set of covariates which have explanatory power for the potential outcomes.

Farrell (2015) suggests exploiting a score function for the estimation of the (unconditional) potential outcomes $\mu_t = E[Y(t)]$, as given by equation (1)

$$\psi_t\big(W; \mu_t, p_t(x), \mu_t(x)\big) = \frac{\mathbb{1}(D = t)(Y - \mu_t(x))}{p_t(x)} + \mu_t(x) - \mu_t \,, \qquad (1)$$

where $t \in (0,1, \dots, \tau)$, $W = (D, X, Y)$ and $\mu_t(x) = E[Y(t)|X = x, D = t]$. It can easily be checked that the score function is zero in expectation, i.e. that the

[17] *Glynn et al. (2010)* conduct an interesting Monte Carlo simulation in order to evaluate the performance of the augmented-inverse probability weighting (AIPW) estimator, relative to several competitors. They find that the AIPW estimator outperforms its competitors if either the propensity score model or the potential outcome model is misspecified.

moment restriction $E\left[\psi_t\left(W;\mu_t,p_t(x),\mu_t(x)\right)\right]=0$ holds. Therefore, the method of moments can be applied in order to obtain an estimate for μ_t. The score function (1) was derived by *Hahn (1998)* and combines (as already mentioned) inverse probability weighting and regression imputation.

Inverse probability weighting relies on the assumption that all of the important covariates which determine the individual probability of belonging to a certain treatment group are known, and can be controlled for. In our case we assume that, conditional on the socio-economic background, the gender, the migration background, the grades attained on the last report card, the quality of the educational system where the college entrance qualification was attained, the type of college entrance qualification attained, the personality traits and the birth cohort of the individual, the treatment status can be considered as good as randomly assigned.

This implies that within a group of individuals with identical characteristics (in terms of the covariates described above), all individual have a comparable generalized propensity score. Within this group treatment individuals can be considered to be as good as randomly assigned. If group members are most likely to be college graduates, for instance, college graduates are oversampled in some sense in this group. In order to eliminate biases in the estimation of the average treatment effect for the overall population, we have to give more weight to the underrepresented group members and less weight to the overrepresented group members. Inverse probability weighting creates a so-called pseudo-population in which the treatment is independent from the confounding variables. If the propensity score is correctly specified, a consistent estimate for the average treatment effect can be obtained by inverse probability weighting.

Regression imputation, in contrast, assumes that all relevant control variables that affect potential outcomes are known and can be controlled for. A model for the potential outcomes of the individuals is estimated for each treatment group separately. Conditional on these control variables, there are no further important variables explaining a significant share of the variation in potential outcomes. In our setting, this implies the assumption that, conditional on the gender, migration background, family background, the quality of secondary education received, age, and the personality traits of an individual, the potential

outcomes can be imputed with high degree of accuracy. The imputed potential outcomes indicate what an individual with the same characteristics would have earned *at the same age* if they had "chosen" a different educational path after secondary education.

As the score function, which combines the two described estimation strategies, is semi-parametrically efficient, the score automatically has another appealing property called *Neyman orthogonality*. Neyman orthogonality implies that the moment restriction is robust to small model selection mistakes. *Chernozuhukov et al. (2017)* claim that every semi-parametrically efficient score must also be Neyman orthogonal. Consequently, the restriction $E\big[\psi_t\big(W; \mu_t, p_t(x), \mu_t(x)\big)\big] = 0$ still holds approximately, even if only high-quality approximations for $p_t(x)$ and $\mu_t(x)$ are plugged in.

The key components of the double machine learning procedure of *Farrell (2015)* are the robustness of the score function with respect to small model selection mistakes (Neyman orthogonality) and the application of sophisticated machine learning tools which yield high-quality approximations for the generalized propensity score and the conditional mean potential outcomes.

The estimation procedure for the average treatment effect of interest can be summarized into four steps. In step 1, the conditional mean potential outcomes $\mu_t(x)$ are estimated by machine learning tools, in step 2 the generalized propensity score $p_t(x)$ is estimated by machine learning tools and in step 3 the predictions $\hat{\mu}_t(x)$ and $\hat{p}_t(x)$ are used as plug-in versions for their true functions and the moment restriction $E\big[\psi_t\big(W; \mu_t, p_t(x), \mu_t(x)\big)\big] = 0$ is replaced by its empirical counterpart in order to get an estimate for the (unconditional) mean potential outcome μ_t according to equation (2). The sample is assumed to contain N individuals in total.

$$\hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\mathbb{1}(D_i = t)(Y_i - \hat{\mu}_t(X_i))}{\hat{p}_t(X_i)} + \hat{\mu}_t(X_i)\right\} \tag{2}$$

In the final step 4, a set of pairwise treatment effects measuring the average effect of treatment m relative to treatment k can be computed based on equation (3) by using the estimated mean potential outcomes $\hat{\mu}_t$ as plug-in versions.

$$\Delta_{mk} \equiv \mu_m - \mu_k = E[Y(m) - Y(k)], \qquad \forall m, k \in \{0,1, \dots, \tau\} \tag{3}$$

Farrell (2015) proves the resulting estimator $\widehat{\Delta}_{mk}$ is both $\sqrt{N}$-consistent and asymptotically normal, under the condition that $\mu_t(x)$ and $p_t(x)$ are consistently estimated by machine learning tools yielding high-quality approximations to the true functions, and under the condition that the product of the convergence rates of the two estimators reaches an order of $N^{-1/2}$.

Farrell (2015) derives the asymptotic result presented in equation (4) where $\Sigma = \nabla_{\Delta_{mk}}(\mu)' V_\mu \nabla_{\Delta_{mk}}(\mu)$ and $\nabla_{\Delta_{mk}}$ denotes the gradient of the function $\Delta_{mk} = \mu_m - \mu_k$ with respect to $(\mu_0, \mu_1, \dots, \mu_\tau)$.

$$\sqrt{N}\left(\widehat{\Delta}_{mk} - \Delta_{mk}\right) \to N(0, \Sigma) \tag{4}$$

The derived asymptotic result is then used to estimate the standard errors corresponding to the treatment effect estimates in the following way. Let $\mu = (\mu_0, \mu_1, \dots, \mu_\tau)'$ denote the $(\tau + 1)$ vector of mean potential outcomes. First, a $[(\tau + 1), (\tau + 1)]$ variance-covariance matrix V_μ is estimated for the mean potential outcomes. The entry $[t, t']$ of the estimated variance-covariance matrix $\hat{V}_\mu$ is computed by equation (5) where $E_N[\cdot] = \frac{1}{N}\sum_{i=1}^{N}(\cdot)$.

$$\begin{aligned} \hat{V}_\mu[t, t'] &\equiv \hat{V}_\mu^W(t) + \hat{V}_\mu^B(t, t') \\ \hat{V}_\mu^W(t) &= E_N\left[\frac{\mathbb{1}(D_i = t)\left(Y_i - \hat{\mu}_t(X_i)\right)^2}{\hat{p}_t(X_i)^2}\right], \quad \text{for } t = t' \\ \hat{V}_\mu^B(t, t') &= E_N[(\hat{\mu}_t(X_i) - \hat{\mu}_t)(\hat{\mu}_{t'}(X_i) - \hat{\mu}_{t'})] \end{aligned} \tag{5}$$

Next, uniformly valid standard errors which will be used to conduct inference can be computed according to equation (6).

$$se\left(\widehat{\Delta}_{mk}\right) = \sqrt{\nabla_{\Delta_{mk}}(\hat{\mu})' \frac{\hat{V}_\mu}{N} \nabla_{\Delta_{mk}}(\hat{\mu})} \tag{6}$$

5.2 Choice of Machine Learning Algorithms

In order to select the algorithms used for the estimation of generalized propensity score and conditional mean potential outcomes, measures for the in-sample and out-of-sample fit of the prediction models are derived in order to select the algorithm with the greatest forecast performance.

Four potential machine learning estimators are considered for the estimation of the generalized propensity score: penalized ordered logit model (with/without generated regressors), Classification Tree and Random Forest. Six machine learning procedures are considered for the estimation of the potential outcomes: Regression Tree, Random Forest, Lasso (with/without generated regressors) and Post-Lasso (with/without generated regressors). The *chosen* algorithms will be described in greater detail in Section 5.3. Algorithms with generated regressors contain all of the interaction terms between the control variables in our data set, listed in Table 8-2 in the Appendix, as well as second and third order polynomials of the continuous variables.

The goal is to estimate prediction models which have some explanatory power for the treatment status and the potential outcomes, but do not overfit to the data, in the sense that the prediction model has poor generalizability. Therefore, a prediction model which has an extremely high in-sample fit is not necessarily also the preferred prediction model. Instead, the out-of-sample performance of the prediction model represents a better indicator of the quality of a prediction model. Due to this, we develop a procedure that allows us to estimate out-of-sample fit measures which indicate how well a prediction model can generalize to other data sets.

The share of correct predictions (SOP) is considered in order to evaluate the forecast performance of the different estimators for the propensity score estimation. Nevertheless, it should be kept in mind that it is in principle possible that a prediction model never predicts a certain treatment status, but is still be able to return good approximations to the generalized propensity score. This is due to the fact that individuals with certain (pre-treatment) characteristics might, for instance, simply be highly unlikely to become college dropouts. By considering the share of correct predictions we prefer, however, the prediction model which has a higher explanatory power for the treatment status.

The forecast performance for the prediction of potential outcomes is evaluated in terms of the mean-squared error. The *in-sample fit measures* are derived using the same data for the estimation and the evaluation of the prediction model. The *out-of-sample fit measures* are derived by 5-fold cross-validation.

The data was split randomly into five different folds of roughly equal size. Four folds are used to train the model, one fold is left out to form a test data set. The

test data set is used to compute the share of correct predictions, or the mean-squared error. The procedure was repeated five times and the out-of-sample measures (the cross-validated share of correct predictions or the cross-validated mean-squared errors) are computed as an average over the different shares of correct predictions, or mean-squared errors, obtained from the respective test samples. The cross-validation procedure was repeated 50 times (in order to take into account the finite sample size of our data), and an average was formed to obtain the final out-of-sample fit measures.

5.3 Machine Learning Estimation

Generalized Propensity Score Estimation

For the estimation of the generalized propensity score we use a penalized ordered logit model. Let D denote the random ordered response variable that in our case takes on values in the set $D \in \{0,1,2\}$. $D = 0$ indicates baseline group membership. We exploit a cumulative link model based on a logistic link function (see *Powers et al. (2008)* for more details). Consequently, the model is based on the cumulative probabilities as given by equation (7).

$$\delta_{ij} \equiv \Pr(D_i \leq j) = F\left(\mu_j + X_i'\beta\right) = \frac{\exp(\mu_j + X_i'\beta)}{1 + \exp(\mu_j + X_i'\beta)} \qquad (\mathbf{7})$$

Based on the cumulative probabilities δ_{ij} we can derive an expression for the individual probabilities of belonging to a specific treatment group j by taking the difference $\Pr(D_i = j|X_i) = \Pr(D_i \leq j\,|X_i) - \Pr(D_i \leq j-1|X_i)$. Introducing the dummy variable d_{ij} that is equal to 1 if an individual belongs to treatment group j, and zero in all other cases, allows us to write down the likelihood function in equation (8) (where $J = 2$ in our setting). The cumulative link model can be estimated by maximum likelihood. Maximum likelihood maximizes the probability of observing the given sample by choosing the threshold parameters μ_j and the coefficients β accordingly.

$$L = \prod_{i=1}^{N} \prod_{j=0}^{J} \Pr(Y_i = j|X_i)^{d_{ij}} \qquad (\mathbf{8})$$

A positive sign of the coefficient β_k indicates that an increase in the variable X_k increases the probability of belonging to the baseline group for which $j = 0$, and decreases the probability of belonging to the highest class for which $j = J$.

The penalized ordered logit model differs from the classical version, as a penalty term is added to the classical optimization problem which penalizes the magnitude of the coefficients. As a consequence some coefficients are shrunk to zero. The objective function for the penalized ordered logit model is given by equation (9)

$$\arg\min_{\beta} -\frac{1}{N}\log L + \lambda \sum_{s=1}^{p} |\beta_s|, \qquad (\mathbf{9})$$

where p denotes the number of potential control variables and λ corresponds to the penalization parameter, which is chosen by 5-fold cross-validation. The R package *ordinalNet* is used to estimate the penalized ordered logit model.

Potential Outcome Estimation

For the estimation of the potential outcomes we have chosen a Lasso estimator. The Lasso estimator minimizes the objective function (10)

$$\arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(Y_i(t) - \sum_{s=1}^{p} X_{is}\beta_s \right)^2 + \lambda \sum_{s=1}^{p} |\beta_s| \right\}, \qquad (\mathbf{10})$$

where $\beta = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ denotes the set of coefficients corresponding to the p potential regressors $(X_{i1}, \dots, X_{ip})'$ and $X_{i1} = 1$ if the model contains an intercept term. The penalty parameter λ is chosen via 5-fold cross-validation. The Lasso estimator penalizes the magnitude of the coefficients by adding a penalty term to the standard mean squared error objective function. Due to the penalty term some coefficients are shrunk to zero and thus the Lasso estimator can be interpreted as a variable selection tool. The higher the value of λ the more coefficients are set to zero. For $\lambda = 0$ the standard ordinary least-squares estimator is obtained.

6 Results

In the first part of the study described in this paper we apply the double machine learning procedure, as proposed by *Farrell (2015),* in order to estimate the effects of college dropout on the labor market outcomes wages and occupational status. In total, 22 control variables (including the second order polynomial for age) are used for the prediction of the potential outcomes and 21 control variables for the approximation of the generalized propensity score. Table 8-2 in the Appendix indicates which variables have been exclusively used for the estimation of the potential outcomes or exclusively for the estimation of the treatment status. Variables characterizing the socio-economic background of the individuals (parental education, social status of father, age of mother at birth of individual) are assumed to have no direct effect on labor market outcomes, other than through their effect on the educational path, and are thus only used for propensity score estimation. Variables that do not correspond to pre-treatment variables (marital status, number of children) can only be used for the outcome estimation. The propensity score estimation considers only the birth cohorts described in Table 8-2 in the Appendix instead of the individual age.

In Section 6.1 we investigate the effect of college dropout on employment by means of a multinomial logistic model. In Section 6.2 we focus on the effect of college dropout on hourly wages. In Subsection 6.2.1 we will discuss the choice of machine learning algorithms for our analysis, in Subsection 6.2.2 we will describe the chosen machine learning algorithms in more detail and finally in Subsection 6.2.3 the treatment effect estimation results for the hourly wages are presented. Section 6.3 presents the estimation results for the occupational prestige status.

6.1 Employment

In the following differences in employment status between college graduates, college dropouts and baseline group members are explored. The employment status represents an important indicator for the labor market prospects of individuals. A multinomial logistic model is estimated using the employment status as dependent variable. The probability of having a specific employment status

is estimated based on a set of control variables that give information on the socio-economic background, the personality traits, the academic achievement and the current family background of the individuals.

Table 8-7 in the Appendix shows the estimation results for the multinomial logistic model. According to the estimation results, there are no significant differences in the ratio between the probability of being part-time or marginally employed and the probability of being not employed between college dropouts and baseline group members after controlling for certain characteristics. Moreover, there are no significant differences in the ratio between the probability of being full-time employed and the probability of being not employed between the two groups. College graduates, however, have a significantly higher probability of being full-time employed relative to being not employed compared to baseline group members.

The result does not mean that such differences at the individual level never exist. There may be heterogeneity in the effect. The results suggest that, on average, when summing up all individual effects there is no significant effect.

6.2 Hourly Gross and Net Wages

6.2.1 Evaluation of the Machine Learning Algorithms

In the next part of the study the double machine learning procedure is applied to estimate the effect of college dropout on hourly wages. Several machine learning estimators are used to obtain high-quality approximation for the generalized propensity score and the conditional mean potential outcomes. In the following we want to examine in detail the algorithms that achieve the highest out-of-sample forecast performance, and thus presumably generalize best to other data sets. Only differences up to the second digit are considered for the comparison of the out-of-sample measures.

Table 6-1 summarizes the computed in-sample and out-of-sample fit measures for the prediction of the treatment status. The upper half of the table depicts the computed in-sample fit measures, as given by the share of correct predictions for algorithms considered in the analysis. The lower part of the table, in contrast, shows the results for the cross-validated share of correct predictions which serve as out-of-sample fit measures. Unsurprisingly, the share of correct

predictions decreases if different data sets are used for estimation and evaluation of the forecast model. Typically, the cross-validated share of correct predictions is lower which indicates that the algorithms are slightly overspecialized to the estimation sample.

Table 6-1: In-Sample and Out-Of-Sample Fit for Treatment Status Predictions

	Penalized Ordered Logit	Classification Tree	Random Forest
SOP	0.64	0.65	0.63
SOP_0	0.21	0.21	0.18
SOP_1	0	0.04	0.01
SOP_2	0.96	0.96	0.95
SOP^{CV}	0.64	0.64	0.63
SOP_0^{CV}	0.21	0.17	0.18
SOP_1^{CV}	0.00	0.00	0.01
SOP_2^{CV}	0.95	0.95	0.95

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. SOP – share of correct predictions, CV – cross-validated, 2 – college graduates, 1 – college dropouts, 0 – baseline group. The upper part of the table shows the results for the in-sample fit measures obtained by different machine learning algorithms, the lower part of the table to the out-of-sample fit measures obtained by a cross-validation procedure.

Table 6-1 additionally shows that all of the algorithms correctly predict the treatment status in expectation in at least 63% of the cases. However, a more detailed look at the share of correct predictions within the various treatment groups demonstrates that the algorithms classify (based on the estimated probabilities) graduates correctly in almost all cases. Nevertheless, it should be kept in mind that the group of college graduates constitutes the group with the highest number of observations, and therefore in some settings the algorithms are close to a naïve classifier which always predicts the treatment status that occurs most frequently in the data set. If an algorithm is a naïve classifier this also implies that the respective prediction model has no explanatory power. Thus, the

control variables in the data set cannot sufficiently describe the educational "choice" of an individual.

In the following we will focus on the out-of-sample fit measures. Table 6-1 shows that all applied algorithms correctly predict in expectation at least 17% of the cases in which individuals did not attend college. We select the penalized ordered logit model for the estimation of the generalized propensity score, as it obtains the highest cross-validated share of correct predictions for all treatment groups. The penalized ordered logit model correctly predicts in expectation 21% of cases for baseline group members, and in 95% of the cases for those with graduate status. We also consider a penalized ordered logit model with generated regressors, as we assume that a more flexible model might obtain an even higher prediction performance. However, we omitted the results for the out-of-sample fit measures in Table 6-1 as the high number of potential control variables makes it impossible for the algorithm to yield the results of the cross-validation procedure within a reasonable amount of time.

However, it should be noted that the penalized ordered logit model generally fails to predict the college dropout status. On the one hand this might be due to the small sample size of college dropouts, which may prevent the algorithms from detecting informative patterns in the data. On the other hand, this may instead reflect the fact that choosing to enroll in college after secondary education and thereafter dropping out from college is simply a highly unrealistic event. Therefore, the penalized ordered logit model can yield nonetheless high-quality approximations for the generalized propensity score even though it rarely predicts the college dropout status, and can lead to consistent average treatment effect estimates.

Next, we will focus on the choice of algorithms for the prediction of the potential outcomes. Table 6-2 and Table 6-3 present the cross-validated mean-squared errors for the potential outcomes corresponding to the different treatment groups. Again, various algorithms for the (log) hourly gross and net wages are considered, including the Random Forest, Regression Tree, Lasso (with/without generated regressors) and the Post-Lasso (with/without generated regressors).

Generally, the Lasso estimator with and without generated regressors yields the best out-of-sample prediction performance for log hourly gross and net wages.

Due to this reason we will use the Lasso estimators to approximate the conditional mean potential outcomes. The Lasso estimator is run separately for each treatment group in order to estimate the potential outcomes $Y(0), Y(1)$ and $Y(2)$.

Table 6-2: Out-of-Sample Fit for Log Hourly Gross Wages

	Random Forest	Regression Tree	Lasso	Lasso*	Post-Lasso	Post-Lasso*
MSE_0^{CV}	0.12	0.12	0.11	0.11	0.11	0.12
MSE_1^{CV}	0.13	0.12	0.12	0.12	0.12	0.13
MSE_2^{CV}	0.11	0.12	0.11	0.11	0.11	0.11

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. MSE – mean-squared error, CV – cross-validated, 2 – college graduates, 1 – college dropouts, 0 – baseline group. The * indicates that generated regressors were considered by the machine learning algorithm. The table shows the out-of-sample fit measures resulting from a cross-validation procedure.

Table 6-3: Out-of-Sample Fit for Log Hourly Net Wages

	Random Forest	Regression Tree	Lasso	Lasso*	Post-Lasso	Post-Lasso*
MSE_0^{CV}	0.09	0.09	0.08	0.08	0.08	0.08
MSE_1^{CV}	0.10	0.10	0.10	0.10	0.10	0.10
MSE_2^{CV}	0.10	0.11	0.10	0.10	0.10	0.10

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. MSE – mean-squared error, CV – cross-validated, 2 – college graduates, 1 – college dropouts, 0 – baseline group. The * indicates that generated regressors were considered by the machine learning algorithm. The table shows the out-of-sample fit measures resulting from a cross-validation procedure.

6.2.2 Description of the Machine Learning Algorithms

This section discusses the detected correlations (which will be exploited by algorithms for prediction) between the treatment group status and the control variables. The estimation results of the penalized ordered logit model *without* any generated regressors are given in Table 8-8 in the Appendix. It can be seen that the coefficients of two variables are shrunk to zero. The cohort dummy for

the birth cohort between 1962 and 1971 is shrunk to zero. Therefore, this group serves as reference group. Furthermore, the extraversion index is shrunk to zero. The variable therefore seems to have no substantial prediction power in addition to the predictive power of the selected variables. In total 19 prediction variables were selected by the penalized ordered logit model.

Positive coefficients indicate that the respective variables increase the probability of being in the baseline group, i.e. of having no college experience and decrease the probability of being a college graduate. For negative coefficients the reverse is true.

Having attained the "Abitur", for instance, seems to increase the probability of being a college graduate compared to individuals that have attained "Fachhochschulreife". Male individuals seem to be more likely to attain a college degree. Worse grades on the last school report card tend to increase the probability of belonging to the baseline group (independent of the subject considered). Moreover, individuals whose parents attained a higher school degree and whose mothers gave birth at an older age seem to have a higher probability of being a college graduate. Interestingly, having a migration background seems to increase the probability of being a college graduate for individuals with a college entrance qualification. Openness and emotional stability tend to increase the probability of being a college graduate whereas more conscientious and agreeable individuals seem to be more likely to have no college experience. To the best of our knowledge, there are currently no studies available that discuss the effect of personality traits on educational choice. Some literature investigates the effect of personality traits on college major choice (e.g. *Humburg; 2012*) finds that choice of college major significantly depends on personality traits. Another branch of the literature examines the effect of the Big Five on academic achievement. *Hakimi et al. (2011)*, for instance, find that conscientiousness is the most important predictor for academic achievement amongst the Big Five and is positively associated with academic achievement. In addition, some cohort effects can be identified. Individuals born before 1962 seem to have a lower probability of being a college graduate in comparison to the reference group whereas individuals born between 1972 and 1991 seem to have a higher probability. Finally, having attained the highest school degree in a federal state with a high-/low-quality educational system tends to increase/decrease the probability for college graduation.

For the penalized ordered logit model *with* generated regressors 30 predictors were selected. It is refrained from a detailed presentation of the results. The main findings of the penalized ordered logit model are again that higher educational levels of the parents are generally associated with a higher propensity of college graduation and having attained the "Abitur" tends to increase the propensity of college graduation.

Next, the detected correlation between wages and control variables will be discussed. The estimation results for the Lasso *without* any generated regressors are given in Table 8-9 (net wages) and in Table 8-10 (gross wages) in the Appendix. The Lasso algorithm selects 7 and 5 predictors for the outcomes (log) hourly net wage and (log) hourly gross wage of baseline group members. For the outcomes of college dropouts 3 predictors are chosen for the log hourly net wages and only 1 predictor for the log hourly gross wages. For the outcomes of college graduates 16 non-zero coefficients can be detected for the log hourly net wages and 11 for the log hourly gross wages respectively.

Two common strong predictors for the log hourly net wages are gender and the age of the individual. Males tend to have higher net wages within all three treatment groups, and older individuals also tend to have higher net wages; which presumably can mainly be explained by the fact age is highly correlated with work experience.

For the baseline group, further predictors for the log hourly net wages, in addition to age and gender, are mainly given by personality traits. For the group of college dropouts, marital status is important in addition to age and gender. Married individuals tend to have higher net wages. For the group of college graduates, personality traits are also additional strong predictors, as well as family background characteristics. In addition, possessing a migration background appears to decrease net wages, and the type of college entrance qualification attained also seems to matter. College graduates that have attained the "Abitur" tend to have higher net wages. Interestingly, the reported math grades seem to be positively correlated with the net wages of college graduates, whereas for reported language grades there seems to be a negative correlation. This finding might be explained by the fact that individuals with higher mathematical skills are more likely to choose a degree from the disciplines of Science, Technology,

Engineering and Mathematics, due to subject-specific interests, and are therefore also more likely to end up in better paid occupations. *Schiefele et al. (1993)* summarize the results of 21 reports which investigate the relationship between academic achievement and subject-specific interest. They assume that academic achievement can essentially be determined by three factors; ability, general motivation, and subject-specific interest. In all of the studies there is evidence of a strong correlation between subject-specific interest and academic achievement.

The estimation results for the log hourly gross wages are presented in the Appendix in Table 8-10, and yield similar results to the estimation results for the log hourly net wages. The estimation results for the Lasso estimation *with* generated regressors for net and gross wages are omitted. The strongest predictor for the estimation of the wages of baseline group members, college dropouts, and college graduates is an interaction term between gender and marital status (which implies that the interaction term explains most of the variation in wages). For each of the three groups, the Lasso estimator indicates that married male individuals tend to earn relatively higher hourly gross and net wages. The number of selected predictors (including generated regressors) for the outcomes of baseline group members, college dropouts, and college graduates are given by 8, 24 and 12 respectively for the log hourly net wages, and by 9, 14 and 20 for the log hourly gross wages.

6.2.3 Treatment Effect Estimation Results

First, we focus on the net hourly wages for which we observed a statistically significant positive average wage differential between college dropouts and the baseline group. Table 6-4 shows the treatment effect estimation results for the hourly net wages[18]. The expected gain in hourly net wages from college graduation relative to not having experienced any college education ($\widehat{\Delta}_{20}$) is approximately 24% for an individual randomly drawn from the population of individuals

[18] Trimming is conducted before the final treatment effect estimates are computed. Observations with probabilities below 3%, 11% and/or 25% for being in the baseline, college dropout or college graduate group respectively are discarded in order to have sufficient overlap in the generalized propensity score. Figure 8-1 in the Appendix depicts the generalized propensity score overlap plot before trimming was conducted.

endowed with a college entrance qualification. In contrast, the expected gain from college graduation relative to college dropout is a bit higher at 25% ($\widehat{\Delta}_{21}$). The two described effects are statistically significant at the usual significance levels. The models with and without generated regressors yield the same point estimates, and come to the same inferential conclusion.

Of further interest is the expected effect of college dropout on hourly net wages when comparing college dropouts to individuals who have never attended college. The estimation results yield that a college dropout leads in expectation to a loss in hourly net wages. However, the effect is not significantly different from zero. Consequently, our results indicate that there are no overall losses (on average) in terms of the hourly gross wages for employed individuals in 2016 for the group of college dropouts relative to the baseline group. Note that the expected effects are estimated taking into account individuals who are in different stages in their lives.

The results show that the unconditional positive hourly net wage differential between college dropouts and baseline group members was significant at usual significance levels (compare Table 3-7), but vanishes as soon as we condition on some control variables.

Table 6-4: Treatment Effect Estimation Results for Hourly Net Wages

		$\widehat{\Delta}_{20}$	$\widehat{\Delta}_{21}$	$\widehat{\Delta}_{10}$
POL/Lasso	estimate	0.23***	0.21***	0.02
	standard error	0.02	0.02	0.03
	t-value	10.80	8.70	0.70
POL*/Lasso*	estimate	0.23***	0.22***	0.01
	standard error	0.02	0.02	0.03
	t-value	11.06	9.61	0.40

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. POL – penalized ordered logit, * – indicates that a model with generated regressors was used. *** - significance at 1% level.

There are several possible reasons why controlling for covariates eliminates the observed positive wage differential in the data. First of all, however, let us discuss the reasons why we observe a statistically significant unconditional wage differential. For instance, it might be the case that the group of college dropouts is positively selected, i.e. college dropouts are endowed with higher academic achievements in school, and specific personality traits which are appreciated by employers. Thus, the potential wages of college dropouts are underestimated by a simple average over the wages of baseline group members. Controlling for academic achievement and personality traits might thus reduce this positive wage differential.

Another reason is that college dropouts and individuals without college experience presumably differ with respect to so-called intermediate outcomes. Intermediate outcomes include the occupations in which individuals are employed, and job characteristics such as occupational autonomy. These differences could represent driving forces promoting a positive wage differential between the two groups if college dropouts generally selected into better paid occupations and their jobs are characterized by higher levels of autonomy. College dropouts who accumulate subject-specific human capital at college presumably want to profit from their human capital, for example, by doing an apprenticeship in a field related to the field of study or by starting immediately a job in a related occupation. If those individuals who chose a field of study generally associated with better-paid occupations (with and without having a college degree) are those most likely to drop out of college, and their apprenticeship and occupational choice are correlated with the chosen field of study, this results in a mechanism driving a positive wage differential. For instance, if computer science students have a higher probability of dropping out of college, and thus represent a substantial group within the group of college dropouts, and are in addition quite likely to end up in an industry sector and/or occupation that is related to computer science, the industry sectors and occupations of college dropouts and baseline groups are structurally different.

Table 6-5 displays the fractions of baseline group members and college dropouts that are assigned to occupational classes according to the International Standard Classification of Occupations (ISCO-08). In particular, major differences can be seen in two occupational classes. A substantially lower share of

college dropouts are classified as clerical support workers (12.5%) when compared to the share of baseline group members (23.1%). Instead, a higher share of college dropouts are classified as professionals (23.8%), i.e. classified into academic occupations, whereas only 9.3% of baseline group members are classified as professionals. Therefore, there might be evidence that the subject of study and the associated skills and knowledge of college dropouts do impact their future professional opportunities and choices.

Table 6-5: ISCO-08 by Treatment Group

	baseline group	college dropouts	college graduates
elementary occupations	3.29%	1.21%	0.21%
clerical support workers	23.05%	12.50%	4.79%
service and sales workers	8.68%	6.85%	2.19%
skilled agricultural, forestry and fishery workers	0.60%	0.81%	0.21%
craft and related trade workers	2.99%	3.23%	0.62%
plant and machine operators and assemblers	0.60%	0.40%	0.42%
technicians and associate professionals	47.31%	45.97%	19.25%
professionals	9.28%	23.79%	65.66%
legislators, senior officials, managers	3.59%	4.84%	6.45%
armed forces occupations	0.60%	0.40%	0.21%

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Note that the information on the ISCO-08 is not given for all individuals. For the baseline group members the information is given for 334 out of 349 individuals, for college dropouts for 248 out of 259 individuals and for college graduates for 916 out of 1,003 individuals.

Table 6-6 shows in detail the chosen occupations of college dropouts if they are classified as professionals according to the ISCO-08. It turns out that 18.6% of college dropouts work as computer system designers, or computer professionals. Although we do not have data on the field of study of college dropouts, it is not unlikely that these individuals studied computer science. It could also be the case that these individuals are more likely to work in small innovative start-up

firms. Another example are individuals who are employed as social work professionals, whose subject of study was presumably related to the social sciences. In total, 13.6% of college dropouts end up as social work professionals.

Table 6-6: Professional Occupations of Baseline and College Dropout Group

occupation	baseline group	college dropouts
computer system designer, computer professional	16.13%	18.64%
civil engineer	0%	3.39%
architect, engineer	0%	3.39%
pharmacist	6.45%	3.39%
secondary education teaching professional	6.45%	6.78%
primary education teaching professional	6.45%	3.39%
special education teaching professional	0%	3.39%
other teaching professional	3.23%	3.39%
other professional	6.45%	1.69%
personnel, carrer professional	6.45%	1.69%
business professional	25.81%	10.17%
archivist, curator	0%	1.69%
psychologist	0%	1.69%
social work professional	12.90%	13.56%
author, journalist, other writer	0%	8.47%
public service administrative professional	9.68%	15.25%

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Types of professional occupations in which 31 baseline group members and 59 college dropouts are employed.

Intermediate outcomes, like the size of the firm at which individuals are employed, represent driving forces that reduce the observed wage differential if, for instance, college dropouts are less likely to find jobs in larger firms due to a negative signaling effect induced by the dropout.

One further driving force which narrows the positive observed wage differential between college dropouts and baseline group members is given by the fact that

we generally expect college dropouts to have less work experience than baseline group members at a given age. Calculations based on the SOEP data yield an average full-time work experience of 12.5 years for baseline group members, and 11.6 years for college dropouts. We find this in spite of the fact the kernel density graph in Figure 8-2 in the Appendix shows baseline group members tend to be younger. For instance, some college dropouts begin an apprenticeship after dropping out, but attain their professional degree later in comparison to baseline group members who immediately started vocational training upon completion of secondary education (unless their acquired skills or knowledge from college are taken into account in the vocational training). Therefore, they enter the labor market at a later point in time.

Next, it is important to discuss reasons why a treatment effect estimation procedure might eliminate the observed positive (unconditional) wage differential between college dropouts and baseline group members. Controlling for academic achievement and personality traits allows to account for potential positive selection effects. Generally speaking, as individuals within the group of college dropouts and the group of individuals with college entrance qualification but no college experience might simultaneously be structurally different in terms of covariates that affect wages, and their propensity to attend college, we apply the double machine learning procedure described in the previous chapter in order to account for these differences, and to allow for a causal interpretation of the effect of college dropout on wages. For instance, it is important to ensure that there is sufficient overlap in the age distributions (as age correlates with work experience) between the two groups. Controlling for age in the estimation of potential outcomes might reduce the observed positive (unconditional) wage differential if baseline group members tend to be younger (which seems to be the case according to the kernel density plots depicted in Figure 8-2). The respective treatments are interpreted in a rather broad sense and comprise the whole educational life path of the individuals and the associated gained knowledge and skills.

An important question still to be answered is what our final treatment effect estimates actually capture. The final treatment effect estimates capture direct wage effects (induced as wages reflect the expected productivity of individuals) as well as indirect effects on wages that occur through so-called intermediate

outcomes, which are the result of individual (educational) life paths and the associated skills and knowledge acquired. *Inverse probability weighting* does not account for the fact that individuals with specific fields of study have a higher probability of dropping out from college, unless we are willing to assume that controlling for academic achievement and personality traits sufficiently accounts for the selection into fields of study. If the selected subject does have an effect on wages through the occupations that college dropouts choose, the treatment effect captures the effect of differences in occupations between college dropouts and baseline group members as well. *Regression imputation*, in contrast, imputes the potential outcome of the individuals assuming that they have followed a completely different educational and professional life path. Therefore, the final treatment effect estimates also capture the effect of differences in life paths. Thus, as we cannot sufficiently control for the selection into occupations, autonomy levels and firm sizes, the effect of intermediate outcomes on wages are captured by our final treatment effect estimates. Finally, treatment effect estimates contain the effect of differences in work experience.

Similar results are obtained for the treatment effects for hourly gross wages. The results are given in Table 6-7. Again, there seems to be no significant effect of a college dropout on hourly gross wages.

Table 6-7: Treatment Effect Estimation Results for Hourly Gross Wages

		$\widehat{\Delta}_{20}$	$\widehat{\Delta}_{21}$	$\widehat{\Delta}_{10}$
POL/Lasso	estimate	0.23***	0.19***	0.03
	standard error	0.02	0.02	0.02
	t-value	12.06	8.80	1.36
POL*/Lasso*	estimate	0.22***	0.19***	0.03
	standard error	0.02	0.02	0.02
	t-value	12.28	9.27	1.11

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. POL – penalized ordered logit, * – indicates that a model with generated regressors was used. *** - significance at 1% level.

6.3 Occupational Prestige Scores

In this section we investigate a further channel through which wages could be affected. Differences in individuals' workplace autonomy, due to differences in educational path pursued, might also affect the wage differential between college dropouts and baseline group members.

In order to apply the machine learning procedure, we replace the three autonomy level categories (introduced earlier) by a continuous variable which is strongly correlated with the categories. The continuous variable used for the machine learning estimation is the *Treiman Standard International Occupation Prestige Score (SIOPS)* that is constructed based on the ISCO-88. The SIOPS ranges from 6 to 78, where higher values of the SIOP indicate a higher level of occupational autonomy.

For the estimation of the generalized propensity score we again use the penalized ordered logit model[19]. Table 8-11 in the Appendix shows the out-of-sample fit measures derived for the predictions of the occupational prestige score. There is no clear ranking between the Lasso estimator and the Regression Tree. However, the out-of-sample prediction performance is comparable for both. We choose the Lasso estimator for further analysis.

The estimation results of the Lasso are given in Table 8-13 in the Appendix. The Lasso algorithm does not detect any strong predictor for the occupational prestige scores of baseline group members. For college dropouts, however, the algorithm identifies two important predictor variables. College dropouts with higher emotional stability scores tend to be in occupations with higher occupational prestige scores. Moreover, college dropouts who attained the college entrance qualification in a federal state with a low-quality educational system tend to be in occupations with lower occupational prestige scores. For college graduates, the Lasso algorithm identifies several predictors for the occupational prestige scores. The most important predictors are the dummy variable for gender, and a dummy variable for the type of college entrance qualification. Male

[19] Table 8-12 in the Appendix shows the estimation results for the penalized ordered logit model for the occupational prestige score data set.

college graduates tend to end up in occupations with higher occupational prestige scores, and college graduates who are endowed with the "Abitur" and not the "Fachhochschulreife" also tend to end up in occupations with higher occupational prestige scores.

Table 6-8 shows the treatment effect estimation results. $\widehat{\Delta}_{20}$ denotes the expected difference in the occupational prestige score between college graduates and baseline group members for an individual randomly drawn from the population of individuals with college entrance qualification. The value 12.69 indicates that college graduates end up, on average, in jobs with occupational prestige scores 12.69 points higher compared to baseline group members. More importantly, we find evidence that college dropouts are more likely to end up in jobs with higher occupational prestige scores compared to baseline group members all other things equal ($\widehat{\Delta}_{10}$=3.53). The result is highly significant at the 1% level.

Table 6-8: Treatment Effect Estimation Results for the Occupational Prestige Score

		$\widehat{\Delta}_{20}$	$\widehat{\Delta}_{21}$	$\widehat{\Delta}_{10}$
POL/Lasso	estimate	10.97***	8.90***	2.07***
	standard error	0.55	0.58	0.71
	t-value	19.88	12.28	3.35

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. POL – penalized ordered logit. *** - significance at 1% level.

Consequently, it seems to be the case that college dropouts self-select into jobs with a relatively high degree of autonomy. Thus, college dropouts might profit as they can achieve positions with a higher level of autonomy than they could have otherwise achieved by directly completing an apprenticeship. This might be due to the fact that either college dropouts accumulated human capital at college, which makes them more productive relative to individuals without college experience if one believes in human capital theory, or that college enrolment itself is considered to be a positive signal conveying information about the productivity of college dropouts, if one believes in signaling theory. This channel

may also explain why we observe a positive (unconditional) wage differential between college dropouts and baseline group members.

7 Discussion

To sum up, the following chapter discusses possible limitations to our estimation procedure, as well as posing open questions for future research.

One data limitation is given by the fact that the SOEP provides a common indicator for the duration of college and school attendance, without distinguishing between the two. Therefore, in some cases clearly differentiating between pupils and students might not be possible. The majority of individuals who drop out, however, can be clearly identified as college dropouts, given we possess the knowledge that these individuals have appeared in the tertiary education system at least once in their lifetime, *or* have been to college after the period in which they acquired the college entrance qualification but before the individual (potentially) took up vocational training. There are some individuals who have only a single period during which they have been to school and/or college, which lasts until at least age 21. In some cases, these individuals might be misclassified as college dropouts if it indeed took them until age 21 to attain the college entrance qualification. In addition, an indicator for the duration of college and school attendance corresponding to technical school attendance might lead to a misclassification of individuals as college dropouts. Therefore, alternative data sources should be used to allow for a sharper identification of college dropouts.

Moreover, classification algorithms are generally biased towards the majority class in imbalanced data sets. Imbalanced data sets are data sets in which the classes are not represented equally. In such cases, the probability associated with the majority class tends to be overestimated, and consequently, the probabilities of the less common classes underestimated. The effect on the final treatment effect estimates is not clear a priori. However, we consider it important to account for data imbalance in future work e.g. by means of resampling techniques, boosting, or bagging methods, in order to be even more confident about the findings of this paper. Another possible extension for the generalized propensity score estimation is to replace the ordinal logit model, which is restrictive in the sense that it does not allow for separate structural mechanisms across treatment group categories, by a sequential logit model

that first models the decision to attend college and thereafter the college dropout "decision".

Although, the applied double machine learning procedure provides several advantages, as discussed in the paper, in comparison to previous attempts to identify the effect of the pursued educational path of college dropouts on their labor market prospects, there are still threats to identification in a conditional-on-observables setting. The standard ability bias problem might lead to a misleading imputation of counterfactual outcomes and might bias inverse probability weighting estimation.

There are also a variety of open questions left for future research. As the SOEP data contains no information on the fields of study for college dropouts nor any reliable measure of the duration of any college attendance spells, here we had to refrain from conducting any detailed analysis with respect to duration and/or field of study. Thus an interesting research question is whether or not more time spent at college increases the labor market prospects of college dropouts, in order to check to some extent the validity of human capital theory. Alternative data sources would be necessary to gain information on fields of study and study durations.

Finally, future research should focus on the effects of college dropout on labor market prospects over different individual life cycles. Immediately after the college dropout one could assume that dropouts experience comparably lower returns from college education, relative to individuals without college experience, than they do years later. Generally speaking, it is of interest to examine the heterogeneity in returns to college education for college dropouts. In addition to life cycle effects, heterogeneity in the returns to college education due to different field of study choices might also be of interest.

This paper found evidence that college dropouts between 25 and 65 years in 2016 do not experience significant losses in terms of hourly wages relative to individuals without college experience who possess college entrance qualification. Furthermore, college dropouts are more likely to end up in occupations with higher prestige scores compared to individuals without any college experience, and do not seem to suffer a significantly higher probability of being unemployed or partially/marginally employed.

References

Aina, C., Baici, E., Casalone, G. & Pastore, F. (2018), 'The economics of university dropouts and delayed graduation: a survey'.

Allmendinger, J. (1989), 'Educational systems and labor market outcomes', *European sociological review* **5**(3), 231–250.

Arrow, K. J. et al. (1973), 'Higher education as a filter', *Journal of public economics* **2**(3), 193–216.

Athey, S. (2018), The impact of machine learning on economics, *in* 'The Economics of Artificial Intelligence: An Agenda', University of Chicago Press.

Bailey, T. R., Kienzl, G. S. & Marcotte, D. (2004), 'The return to a sub-baccalaureate education: The effects of schooling, credentials and program of study on economic outcomes' (Report no. ED-00-CO-0023). U.S. Department of Education.

Becker, G. S. (1962), 'Investment in human capital: A theoretical analysis', *Journal of Political Economy* **70**(5, Part 2), 9–49.

Autorengruppe Bildungsberichterstattung (2018), *Bildung in Deutschland 2018: ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung*, W. Bertelsmann Verlag.

Card, D. (1999), The causal effect of education on earnings, in 'Handbook of labor economics', Vol. 3, Elsevier, pp. 1801–1863.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. & Newey, W. (2017), 'Double/debiased/neyman machine learning of treatment effects', *American Economic Review* **107**(5), 261–65.

Collins, R. (1979), *The credential society: An historical sociology of education and stratification*, Academic Press.

Davies, R. & Elias, P. (2003), *Dropping out: A study of early leavers from higher education*, Citeseer.

Farrell, M. H. (2015), 'Robust inference on average treatment effects with possibly more covariates than observations', *Journal of Econometrics* **189**(1), 1–23.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York, NY, USA.

Gerlitz, J.-Y. & Schupp, J. (2005), 'Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP', *DIW Research Notes* **4**, 2005.

Glynn, A. N. & Quinn, K. M. (2010), 'An introduction to the augmented inverse propensity weighted estimator', *Political analysis* **18**(1), 36–56.

Grubb, W. N. (2002), 'Learning and earning in the middle, part i: National studies of pre-baccalaureate education', *Economics of Education Review* **21**(4), 299–321.

Hahn, J. (1998), 'On the role of the propensity score in efficient semiparametric estimation of average treatment effects', *Econometrica* pp. 315–331.

Hakimi, S., Hejazi, E. & Lavasani, M. G. (2011), 'The relationships between personality traits and students' academic achievement', *Procedia-Social and Behavioral Sciences* **29**, 836–845.

Heckman, J. J. & Rubinstein, Y. (2001), 'The importance of noncognitive skills: Lessons from the GED testing program', *American Economic Review* **91**(2), 145–149.

Heublein, U. (2014), *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012*, Dt. Zentrum für Hochsch.-und Wiss.-Forschung.

Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. & Woisch, A. (2017), 'Zwischen Studienerwartungen und Studienwirklichkeit', *Forum Hochschule* **1**(2017), 134–136.

Humburg, M. (2012), 'The effect of the big five personality traits on college major choice: Evidence from a dutch longitudinal youth cohort study.', Maastricht: *Research centre for education and the labour market*.

Imbens, G. (2003), Semiparametric estimation of average treatment effects under exogeneity: a review, *Technical report*, Department of Economics, University of California—Berkeley.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Johnes, J. & Taylor, J. (1991), 'Non-completion of a degree course and its effect on the subsequent experience of non-completers in the labour market', *Studies in Higher Education* **16**(1), 73–81.

Krapp, A., Schiefele, U. & Schreyer, I. (1993), 'Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung', *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie* **10**(2), 120–148.

Matković, T. & Kogan, I. (2014), 'Relative worth of a bachelor degree: Patterns of labour market integration among drop-outs and graduates in sequential and integrated tertiary education systems', *Acta Sociologica* **57**(2), 101–118.

Müller, S. & Schneider, T. (2013), 'Educational pathways and dropout from higher education in Germany', *Longitudinal and Life Course Studies* **4**(3), 218–241.

Powers, D. & Xie, Y. (2008), *Statistical methods for categorical data analysis*, Emerald Group Publishing.

Reisel, L. (2013), 'Is more always better? Early career returns to education in the United States and Norway', *Research in Social Stratification and Mobility* **31**, 49–68.

Schnepf, S. V. (2015), 'University dropouts and labor market success', *IZA World of Labor,* doi: 10.15185/izawol.182.

Scholten, M. M. & Tieben, N. (2017), 'Labour market outcomes of higher-education dropouts in Germany: how formal vocational qualifications shape education-to-work transitions and occupational status.', *Arbeitspapiere/Mannheimer Zentrum für Europäische Sozialforschung, Working papers* 168.

Spence, M. (1978), Job market signaling, in 'Uncertainty in Economics', Academic Press, 281-306.

Stiglitz, J. E. (1975), ‘The theory of “screening", education, and the distribution of income’, *The American Economic Review* **65**(3), 283–300.

8 Appendix

Table 8-1: Statements of the Big Five Personality Traits

	statement
openness	"I am original." (+)
	"I value artistic experiences." (+)
	"I have a lively imagination." (+)
conscientiousness	"I am a thorough worker." (+)
	"I tend to be lazy." (-)
	"I carry tasks out efficiently." (+)
extraversion	"I am communicative." (+)
	"I am sociable." (+)
	"I am reserved." (-)
agreeableness	"I am sometimes to coarse with others." (-)
	"I am able to forgive." (+)
	"I am friendly with others." (+)
neuroticism	"I worry a lot." (+)
	"I am somewhat nervous." (+)
	"I deal well with stress." (-)

Source: SOEP 2005, 2009, 2013. See *Gerlitz et al. (2005)* for a detailed description on the collection of the Big Five. (+) – indicates that the original variable must not be recoded, (-) – indicates that original variable is recoded in order to construct a score for the respective personality trait.

Table 8-2: Variable Description

Variable	Description
Treatment variable	
D	=0 if no college experience, =1 if college dropout, = 2 if college graduate
Socio-economic background	
motage[a]	age of mother at birth of individual
motedu[a]	highest attained school degree mother
	1 – no school degree, 2 – Hauptschule, 3 – Realschule,
	4 – Fachoberschule, 5 - Gymnasium
fathedu[a]	highest attained school degree father
	1 – no school degree, 2 – Hauptschule, 3 – Realschule,
	4 – Fachoberschule, 5 – Gymnasium
fisei[a]	social status of father (ISEI-classification) ϵ [16,90]
migback	=2 if direct, =1 if indirect, =0 if no migration background
Individual and household information	
sex	dummy, =1 if male
age[b]	age of individual
married[b]	dummy, =1 if married
child07[b]	number of children age 0-7 in household
child815[b]	number of children age 8-15 in household

Note: a - indicates that the respective variable was only used for propensity score estimation, b - indicates that it was only used for potential outcome estimation.

Table continued on next page.

Personality Traits	
agree	score between 3 and 21 (21 – very agreeable, 3 – not very agreeable)
emostab	score between 3 and 21 (21 – very stable, 3 – not very stable)
consc	score between 3 and 21 (21 – very conscientious, 3 – not very conscientious)
open	score between 3 and 21 (21 – very open, 3 – not very open)
extra	score between 3 and 21 (21 – very extraverted, 3 – not very extraverted)
School grades on last report card	
german	German grade between 1 and 6 (1 – very good, 6 – insufficient)
math	Math grade between 1 and 6 (1 – very good, 6 – insufficient)
foreign	Foreign language grade between 1 and 6 (1 – very good, 6 – insufficient)
Birth Cohort	
cohort1	dummy, =1 if born between 1951 and 1961
cohort2	dummy, =1 if born between 1962 and 1971
cohort3	dummy, =1 if born between 1972 and 1983
cohort4	dummy, =1 if born between 1984 and 1991
Federal State where school degree attained	
high	dummy, =1 if degree obtained in state with high quality educational system
low	dummy, =1 if degree obtained in state with low quality educational system
Type of highest school degree attained	
abi	Dummy, =1 if “Abitur” was attained, =0 if “Fachhochschulreife” was attained

Table 8-3: Summary Statistics

	Mean	Std. Dev.	Min	Max
Dependent Variable				
log hourly net wage	2.54	0.35	1.17	3.41
log hourly gross wage	2.94	0.38	1.57	3.57
Socio-economic background				
mother's age at birth	27.12	5.03	17	45
highest educational degree of mother	2.89	1.13	1	5
highest educational degree of father	3.17	1.31	1	5
social status of father	49.31	17.82	16	90
migration background	0.13	0.42	0	2
Individual and household information				
sex	0.40	0.49	0	1
age	45.17	9.00	25	65
marital status	0.69	0.46	0	1
number of children age 0-7 in household	0.42	0.72	0	4
number of children age 8-15 in household	0.65	0.91	0	4
Personality Traits				
agreeableness	16.24	2.54	8	21
emotional stability	13.08	3.50	3	21
conscientiousness	17.23	2.39	7	21

Table continued on next page.

	Mean	Std. Dev.	Min	Max
openness	14.11	3.09	4	21
extraversion	14.73	3.38	4	21
School grades on last report card				
German	2.34	0.82	1	5
mathematics	2.48	1.09	1	6
first foreign language	2.49	0.96	1	6
Birth Cohort				
cohort 1 (born between [1951,1961])	0.17	0.38	0	1
cohort 2 (born between [1962,1972])	0.39	0.49	0	1
cohort 3 (born between [1973,1983])	0.36	0.48	0	1
cohort 4 (born between [1984,1991])	0.08	0.28	0	1
Federal State where school degree attained				
high quality educational system	0.30	0.46	0	1
low quality educational system	0.25	0.43	0	1
Type of highest school degree attained				
Abitur	0.77	0.42	0	1
Treatment group composition (N = 1,611)				
no college experience	N = 349			
college dropout	N = 259			
college graduate	N = 1,003			

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations.

Table 8-4: Summary Statistics by Treatment Group

	baseline group		college dropouts		college graduates	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
motage	26.39	5.01	26.86	5.18	27.45	4.97
motedu	2.70	1.01	2.85	1.10	2.97	1.17
fathedu	2.91	1.19	3.09	1.31	3.27	1.34
fatss	45.11	16.18	47.09	17.48	51.34	18.15
sex	0.26	0.44	0.46	0.50	0.44	0.50
migback	0.10	0.37	0.21	0.53	0.12	0.41
consc	17.64	2.34	17.07	2.52	17.13	2.36
open	13.83	3.08	14.31	3.22	14.15	3.05
ext	14.91	3.48	14.96	3.55	14.61	3.29
agree	16.41	2.48	16.21	2.58	16.19	2.55
emostab	12.49	3.49	13.23	3.59	13.24	3.47
german	2.43	0.75	2.47	0.85	2.27	0.83
math	2.77	0.99	2.71	1.11	2.32	1.09
foreign	2.58	0.91	2.72	0.97	2.39	0.96
married	0.65	0.48	0.60	0.49	0.73	0.45
kids7	0.37	0.65	0.45	0.76	0.42	0.73
kids15	0.66	0.88	0.61	0.91	0.66	0.92
age	43.62	9.07	43.50	8.68	46.14	8.93
cohort1	0.14	0.35	0.12	0.33	0.19	0.40
cohort2	0.32	0.47	0.39	0.49	0.41	0.49
cohort3	0.45	0.50	0.36	0.48	0.32	0.47
cohort4	0.09	0.29	0.13	0.33	0.07	0.25
top	0.25	0.25	0.26	0.44	0.33	0.47
low	0.34	0.34	0.29	0.45	0.21	0.41
abi	0.62	0.62	0.69	0.69	0.84	0.36
N	349		259		1,003	

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations.

Table 8-5: Two-sided t-Tests for the Equality of Means

social status of father		
Δ_{21} = 4.25	Δ_{20} = 6.22	Δ_{10} = 1.98
t = 3.3821, p = 0.0007	t = 5.6678, p = 0.0000	t = 1.4381, p = 0.1509
age of mother at birth of individual		
Δ_{21} = 0.58	Δ_{20} = 1.06	Δ_{10} = 0.47
t = 1.67, p = 0.0945	t = 3.41, p = 0.0007	t = 1.13, p = 0.2589
openness		
Δ_{21} = -0.16	Δ_{20} = 0.32	Δ_{10} = 0.48
t = -0.75, p = 0.4544	t = 1.69, p = 0.0910	t = 1.87, p = 0.0616
conscientiousness		
Δ_{21} = 0.06	Δ_{20} = -0.52	Δ_{10} = -0.58
t = 0.38, p = 0.7018	t = -3.53, p = 0.0004	t = -2.93, p = 0.0035
extraversion		
Δ_{21} = -0.35	Δ_{20} = -0.31	Δ_{10} = 0.04
t = -1.51, p = 0.1313	t = -1.49, p = 0.1372	t = 0.15, p = 0.8800
agreeableness		
Δ_{21} = -0.02	Δ_{20} = -0.22	Δ_{10} = -0.20
t = -0.11, p = 0.9106	t = -1.39, p = 0.1651	t = -0.96, p = 0.3370
emotional stability		
Δ_{21} = 0.00	Δ_{20} = 0.75	Δ_{10} = 0.74
t = 0.02, p = 0.9839	t = 3.46, p = 0.0006	t = 2.56, p = 0.0108

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Δ_{21} – difference in averages between college graduates and college dropouts, Δ_{20} – difference in averages between college graduates and baseline group members, Δ_{10} – difference in averages between college dropouts and baseline group members.

Table 8-6: Mean Weekly Working Hours by Treatment Group

		mean	std. dev.	min	max
HOURS	no college experience	33.05	11.63	8	60
	college dropouts	35.79	10.60	8	60
		$\Delta = 2.74, \text{d.f.} = 606, \text{t} = 2.98, \text{p} = 0.003$			
	college graduates	37.72	10.29	8	60

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Δ reflects the difference in mean hourly gross/net wages between college dropouts and individuals from the baseline group. Statistics of a two-sample t-test for the equality of means are given by: d.f. – degrees of freedom, t – t-value, p – p-value. The results are based on the basic sample.

Table 8-7: Multinomial Logistic Regression Results for Employment Status

	Coef.	Std. Err.	z	P > |z|
not employed		(base outcome)		
partially/marginally employed				
d1	-0.2528	0.2274	-1.11	0.266
d2	-0.0144	0.1815	-0.08	0.937
motage	0.0317**	0.01418	2.24	0.025
fatss	-0.0000	0.0051	-0.00	0.996
motedu	0.0310	0.0733	0.42	0.672
fathedu	-0.0546	0.0740	-0.74	0.461
sex	-1.0128***	0.1918	-5.28	0.000
migback	-0.2342	0.1555	-1.51	0.132
consc	0.0097	0.0305	0.32	0.751
emostab	0.0779***	0.0208	3.74	0.000
open	-0.0504**	0.0242	-2.08	0.037
extra	-0.0264	0.0232	-1.14	0.255
agree	-0.0337	0.0291	-1.16	0.248
german	-0.0123	0.1026	-0.12	0.905
math	0.0555	0.0698	0.79	0.427
foreign	0.0036	0.0852	0.04	0.966
married	-0.0172	0.1843	-0.09	0.926
kids7	-0.4091***	0.1053	-3.88	0.000
kids15	-0.0749	0.0888	-0.84	0.399
age	0.7695***	0.0720	10.69	0.000
agesq	-0.0087***	0.0008	-11.26	0.000
_cons	-14.9585***	1.8517	-8.08	0.000

Table continued on next page.

full-time employed

d1	-0.1246	0.2276	-0.55	0.584
d2	0.6068***	0.1828	3.32	0.001
motage	0.0019	0.0138	0.13	0.893
fatss	-0.0065	0.0049	-1.32	0.188
Motedu	0.0603	0.0721	0.84	0.402
fathedu	0.0063	0.0722	0.09	0.930
sex	1.8885***	0.1673	11.29	0.000
migback	-0.1192	0.1486	-0.80	0.422
consc	0.1080***	0.0300	3.61	0.000
emostab	0.1286***	0.0204	6.31	0.000
open	-0.0446*	0.0236	-1.89	0.059
extra	-0.0216	0.0226	-0.96	0.339
agree	-0.0640**	0.0281	-2.28	0.023
german	0.0238	0.0984	0.24	0.809
math	0.0572	0.0677	0.85	0.398
foreign	-0.1402*	0.0829	-1.69	0.091
married	-0.6637***	0.1764	-3.76	0.000
kids7	-0.8923***	0.1052	-8.48	0.000
kids15	-0.4986***	0.0904	-5.51	0.000
age	0.6154***	0.0641	9.60	0.000
agesq	-0.0074***	0.0007	-10.72	0.000
_cons	-11.3660***	1.6794	-6.77	0.000

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. d1 is a dummy variable that is equal to 1 if the individual is a college dropout and 0 else, d2 is a dummy variable that is equal to 1 if the individual is a college graduate and 0 else. *** - significance at 1% level, ** - significance at 5% level, * - significance at 10% level.

Table 8-8: Estimation Results Penalized Ordered Logit Model

	coefficient		coefficient
motage	-0.0169	math	0.2695
motedu	-0.0986	foreign	0.0675
fathedu	-0.0053	cohort1	-0.1934
fatss	-0.0085	cohort2	0
sex	-0.4464	cohort3	0.5186
migback	-0.0753	cohort4	0.6141
consc	0.0452	top	-0.2935
open	-0.0194	low	0.2326
extra	0	abi	-0.8261
agree	0.0072	μ_0	-1.1397
emostab	-0.0181	μ_1	-0.2525
german	0.1329		

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. μ_0 and μ_1 represent estimated threshold parameters.

Table 8-9: Lasso Estimation Results for the Log Hourly Net Wages

	$Y(0)$	$Y(1)$	$Y(2)$
intercept	2.4594	2.3359	2.1002
sex	0.1378	0.1201	0.1603
migback	0	0	-0.0135
consc	-0.0061	0	0.0069
open	0	0	-0.0086
ext	0	0	0
agree	-0.0130	0	-0.0013
emostab	0.0025	0	0.0029
german	0	0	0.0201
math	0	0	-0.0124
foreign	0	0	0.0075
married	0	0.0014	0
kids7	0	0	0.0223
kids15	0	0	0.0162
age	0.0030	0.0001	0.0069
age^2	0	0	0
cohort1	0	0	0
cohort2	0.0222	0	0.0156
cohort3	0	0	0
cohort4	0	0	-0.0684
top	0	0	0
low	0.0084	0	0.0178
abi	0	0	0.0986

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Y(0) – potential outcome model for baseline group members, Y(1) – potential outcome model for college dropouts, Y(2) – potential outcome model for college graduates.

Table 8-10: Lasso Estimation Results for the Log Hourly Gross Wages

	$Y(0)$	$Y(1)$	$Y(2)$
intercept	2.8003	2.7249	2.8148
sex	0.1569	0.1291	0.1595
migback	0	0	0
consc	-0.0014	0	0.0014
open	0	0	-0.0066
ext	0	0	0
agree	-0.0151	0	-0.0008
emostab	0	0	0.0017
german	0	0	0.0062
math	0	0	-0.0191
foreign	0	0	0
married	0	0	0
kids7	0	0	0
kids15	0	0	0
age	0.0038	0	0.0043
age^2	0	0	0
cohort1	0	0	0
cohort2	0.0148	0	0
cohort3	0	0	0
cohort4	0	0	-0.0910
top	0	0	0
low	0	0	0.0305
abi	0	0	0.0567

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Y(0) – potential outcome model for baseline group members, Y(1) – potential outcome model for college dropouts, Y(2) – potential outcome model for college graduates.

Table 8-11: Out-of-Sample Fit for Occupational Prestige Scores

	Random Forest	Regression Tree	Lasso	Lasso*	Post-Lasso	Post-Lasso*
MSE_0^{CV}	97.65	89.96	90.02	90.55	90.26	92.44
MSE_1^{CV}	102.53	97.55	97.90	98.30	99.96	102.81
MSE_2^{CV}	111.68	112.87	110.71	110.71	111.13	112.37

Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. MSE – mean-squared error. The table shows the results for the out-of-sample fit measures obtained by a cross-validation procedure for different machine learning algorithms.

Table 8-12: Estimation Results Penalized Ordered Logit Model (Occupational Prestige Score Data Set)

	coefficient		coefficient
motage	-0.0225	math	0.2946
motedu	-0.0954	foreign	0.0750
fathedu	-0.0374	cohort1	-0.3689
fatss	-0.0104	cohort2	0
sex	-0.4961	cohort3	0.4570
migback	-0.0555	cohort4	0.6164
consc	0.0442	top	-0.3543
open	-0.0201	low	0.2831
extra	0.0038	abi	-0.9098
agree	-0.0060	μ_0	-0.7515
emostab	-0.0169	μ_1	0.0869
german	0.1699		

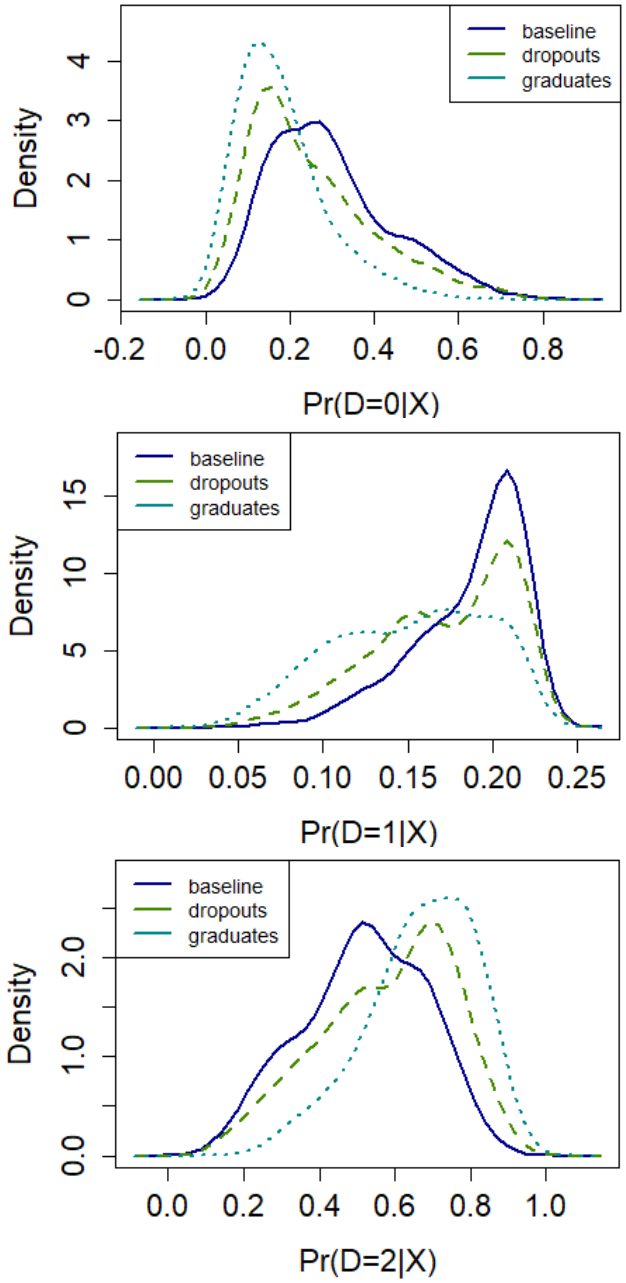
Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. μ_0 and μ_1 represent estimated threshold parameters. The dependent variable is given by the treatment group indicator.

Table 8-13: Lasso Estimation Results for the Occupational Prestige Scores

	$Y(0)$	$Y(1)$	$Y(2)$
intercept	45.3296	46.3925	57.7818
sex	0	0	1.8566
migback	0	0	0
consc	0	0	0.0756
open	0	0	0
ext	0	0	-0.1196
agree	0	0	0
emostab	0	0.1519	0
german	0	0	-0.0673
math	0	0	-0.6615
foreign	0	0	-0.0532
married	0	0	0.7649
kids7	0	0	0.0058
kids15	0	0	0.0205
age	0	0	0
age^2	0	0	-0.0001
cohort1	0	0	0
cohort2	0	0	0
cohort3	0	0	0.0245
cohort4	0	0	0
top	0	0	0
low	0	-0.4964	0.5818
abi	0	0	4.4793

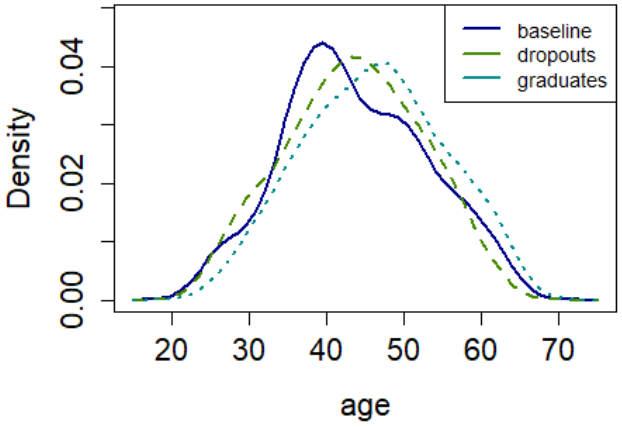
Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Y(0) – potential outcome model for baseline group members, Y(1) – potential outcome model for college dropouts, Y(2) – potential outcome model for college graduates.

Figure 8-1: Generalized Propensity Score Overlap



Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations. Kernel density estimates for the estimated generalized propensity score (based on a penalized ordered logit model) by treatment group.

Figure 8-2: Kernel Density Plot for Age



Source: Socio-Economic Panel (SOEP), data for the year 2016, version 33, own calculations.