RUHR

ECONOMIC PAPERS


Matthias Westphal

Daniel A. Kamhöfer

Hendrik Schmitz


# Marginal College Wage Premiums under Selection into Employment


technische universität dortmund

rwi

#855

# Imprint

## Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

## Editors



## Editorial Office



## Ruhr Economic Papers #855

Responsible Editor: Wolfgang Leininger



ISSN 1864-4872 (online) – ISBN 978-3-86788-991-9

Ruhr Economic Papers #855

Matthias Westphal, Daniel A. Kamhöfer, and Hendrik Schmitz

# Marginal College Wage Premiums under Selection into Employment

technische universität dortmund

## Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/86788991
ISSN 1864-4872 (online)
ISBN 978-3-86788-991-9

Matthias Westphal, Daniel A. Kamhöfer, and Hendrik Schmitz[1]

# Marginal College Wage Premiums under Selection into Employment


## Abstract

*In this paper, we identify female long-term wage returns to college education using the educational expansion between 1960–1990 in West Germany as exogenous variation for college enrollment. We estimate marginal treatment effects to learn about the underlying behavioral structure of women who decide for or against going to college (e.g., whether there is selection into gains). We propose a simple partial identification technique using an adjusted version of the Lee bounds to account for women who select into employment due to having a college education, which we call college-induced selection into employment (CISE). We find that women are, on average, more than 17 percentage points more likely to be employed due to having a college education than without. Taking this CISE into account, we find wage returns of 6–12 percent per year of education completed (average treatment effects on the treated).*



*JEL-Code: C31, I26, J24*

*Keywords: Marginal treatment effect; partial identification; returns to higher education; female labor force participation*

*June 2020*



*1 Matthias Westphal, TU Dortmund, RWI and LSCR; Daniel A. Kamhöfer, University of Düsseldorf and IZA; Hendrik Schmitz, Paderborn University, RWI and LSCR.* *– Funding from the German Research Foundation (DFG, Grant number SCHM 3140/1-1) and from the Leibniz Science Campus Ruhr is gratefully acknowledged. We thank three anonymous referees and the editor for very helpful comments that considerably improved the paper. We also thank Pedro Carneiro, Damon Clark, Thomas Cornelissen, Yingying Dong, Christian Dustmann, Sylvi Rzepka, Andrea Weber, and participants at various workshops for valuable comments.* *– All correspondence to: Matthias Westphal, Department of Economics, TU Dortmund, 44221 Dortmund, Germany, e-mail: matthias.westphal@tu-dortmund.de*

# 1 Introduction

A recent empirical literature (Carneiro et al., 2011; Nybom, 2017) studies the effect of education on wages and goes beyond identifying local average treatment effects. By estimating marginal treatment effects (MTEs), the authors not only derive (arguably) more relevant treatment parameters, such as average treatment effects on the treated, but they also learn about economic behavior such as "selection into gains", meaning that individuals who will benefit most from education are more likely to self-select into it. This literature, as most of the previous causal literature in this field, restricts estimation to men (see Angrist and Krueger, 1992, 1995; Lemieux and Card, 2001), mostly due to data availability or because exploited exogenous variation only applies to men.

Estimating the labor market effects of education for women is beneficial, not least because women have mainly benefited from educational expansions in previous decades throughout the world. The rise in tertiary education has provided the strongest societal change in individual education trajectories in the previous decades. For example, college graduation rates in the U.S. increased from around 10 percent for those born 1920–1930 through 20 percent (birth cohorts 1940–1950) to 30–40 percent among the 1980 birth cohort (Goldin et al., 2006). Similar developments are observed in other developed countries such as Germany (see Figure A1). The rise in college education has been accompanied by the "reversal of the college gender gap" (e.g., Goldin et al., 2006) or, more generally, the gender education gap. While, historically, men have had higher tertiary education rates, women have caught up over time and even surpassed men in the U.S. Again, this is fairly similar in other countries. While full reversal was not accomplished in each and every country, the gender education gap has been mostly closed throughout the developed world (Parro, 2012; Pekkarinen, 2012; Riphahn and Schwientek, 2015).

In this paper we contribute to the literature by estimating causal effects of education on labor market outcomes for women using a representative German data set. Thus, to our knowledge, we are the first to estimate MTEs of education for women using the approach of Heckman and Vytlacil (2005) to learn about their underlying economic behavior. Of course, female wage returns to education have been studied before in structural models that allow to identify comparable magnitudes (see, e.g., Eckstein and Wolpin, 1989, Eckstein and Lifshitz, 2011, Blundell et al., 2016).[1] Arguably, the Heckman and Vytlacil (2005) framework makes weaker distributional assumptions than fully parameterized models but keeps the power to draw more externally valid conclusions than more reduced-form approaches (Heckman, 2010).

[1] And, more generally, the structural econometric literature on female labor market behavior, not necessarily related to education, is large; see Blundell and MaCurdy (1999) for a review.

In our analysis, we measure wage returns in 2011/2012 while college education occurred between 1961 and 1990, thus estimating long-term returns.[2] For identification, we rely on arguably exogenous variation induced by the German "educational expansion" that strongly increased the number of colleges and available spots between 1960 and 1990 and thus reduced college enrollment costs.

When estimating female wage returns to education, another selection problem (in addition to the well-known problem that individuals who select into education probably also earn more due to their innate abilities) arises: sample selection due to non-random selection into employment. Individuals with certain unobserved characteristics not only self-select into education but also into employment as a result of being college educated. Moreover, these characteristics can also affect wages. We call this sample selection problem *college-induced selection into employment* (CISE).

This problem cannot be fixed by an instrument for college education. Selection into the labor market has been considered a source of estimation bias since the seminal works of Gronau (1974) and Heckman (1979) but has recently predominantly not been addressed in the returns to education literature.[3] In some applications selection into the labor market might not be empirically relevant (as in Zimmerman, 2014), particularly if effects are only estimated for men. There, CISE might not matter too much, as male labor force participation is very high anyway and is not easily manipulated (see Blau and Kahn, 2017). Yet, if female college returns are considered, this might well become an issue, particularly in applications from Europe where female employment rates are still lower than in the U.S., leaving more room for sample selection problems.

Another reason the literature does not address the CISE problem could be due to difficulty in solving it in applied work. Sample selection models in the spirit of Heckman (1979) rely on instruments that affect employment but not wages. Strong and convincing instruments of this kind are hardly available in practice. Apart from strong distributional assumptions (which can, however, easily be relaxed), we assume this lack of appropriate instruments to be a major reason why sample selection models have come somewhat out of fashion in applied work. A notable exception are studies on trends in the gender wage gap over time where observed wages for women are routinely adjusted for changing labor force participation rates.[4] Still, this literature does not aim at identifying causal returns to education.

[2] We refer to college (education) and university (education) interchangeably. The average duration to receive this education was 6 years in the time under review, the counterfactual being usually a vocational education (see in what follows).

[3] In theoretical econometrics, non-random sample selection is still frequently being discussed; see, e.g., Arellano and Bonhomme (2017) and Fernandez-Val et al. (2018) for examples of recent work.

[4] See, e.g., Mulligan and Rubinstein (2008), Machado (2017), and the literature review by Blau and Kahn (2017) for the estimation for gender gaps, as well as Huber and Melly (2015) and Bar et al. (2015) for a discussion of the validity of those approaches.

In addition to estimating marginal college wage premiums for women, we contribute to the literature by proposing a simple partial identification solution explicitly addressing the sample selection problem in an instrumental variables (IV) setting. Estimating MTEs and distributions of potential outcomes help us to impose some, arguably mild, restrictions and to nevertheless receive meaningful bounds on wage effects. We start by transferring the natural bounds proposed by Lee (2009) from randomized controlled trials to our IV setting. We then complement these bounds by imposing shape restrictions along the unobserved distaste to study, which are used in the literature to derive more externally valid, policy-relevant effects (Mogstad et al., 2018). These restrictions might be, for instance, that MTEs are linear (as is also done in the literature in other settings; see Brinch et al., 2017) and there is no selection into losses. Our scenarios complement each other and, arguably, require weaker assumptions than the exclusion restrictions typically used in sample selection models. Moreover, our approach is simple and easily implemented yet is still informed by the underlying data and can deliver bounds narrow enough to draw meaningful conclusions.

The work of Dong (2019) and Lee (2009) are most closely related to ours. The seminal paper by Lee (2009) proposes a partial identification approach to bound the effect of selection into employment when a (job training) treatment is randomly assigned. A major difference in our study is that the treatment we analyze (college education) can hardly be considered random. Dong (2019) applies Lee's approach to a regression discontinuity framework to derive sharp bounds at the cutoff. We build on her work and adjust Lee's approach to our MTE framework. Although the use of partial identification to address selection in wage regressions is not completely new (see Blundell et al., 2007 and Lee, 2009), to our knowledge, it has not been used in the context of returns to education—or, more technically, to a setting with an endogenous variable that appears both in the selection and in the outcome equation.

We find that women are, on average, more than 17 percentage points more likely to be employed due to college education. The college-induced change in the propensity to work varies a bit along the unobserved distaste to study (and therefore all treatment parameters like the average treatment effect, ATE, the average treatment effect on the treated, ATT, or the local average treatment affect, LATE, differ to some extent), but generally the effect is substantial. This number makes sample selection a relevant issue in this application, affecting not only MTEs but also possibly all wage estimates. Without adjusting for selection, the ATE college wage premiums are estimated to be 10 percent per year of education completed. After accounting for potential selection, we can bound the effects between 6 and 12 percent. Thus, accounting for selection into employment may absorb some portion of the college wage premium. Nevertheless, college education still seems to be always pecuniary beneficial for women.

The paper proceeds as follows. Section 2 discusses the empirical approach, selection problems and our proposed solutions. Section 3 presents the policy environment and the data used, and Section 4 reports the baseline results. Section 5 provides an outline of an alternative partial identification strategy. Section 6 provides robustness checks, and Section 7 concludes.

# 2 Empirical strategy

In this section, we first outline a conventional estimation approach, as is standard in the returns to education literature, and discuss how selection into the labor market may affect the structural estimates in this setting. We then introduce the estimation of marginal treatment effects through potential outcomes and conclude by suggesting a bounding approach for the potential outcome with college education under selection into employment.

## 2.1 Baseline identification strategy

Let $D_i$ be a binary treatment indicator that takes a value of one if individual $i$ has college education and zero otherwise. The observed outcome $Y_i$ is the gross log hourly wage of individual $i$ in year 2011/2012 (cross-section). Each individual has two potential outcomes, $Y_i^0$ without college education and $Y_i^1$ with college education. Only one of the two is observed, inducing the fundamental evaluation problem, as the individual causal effect of college on wages, $Y_i^1 - Y_i^0$, is unobserved. We can then specify the following simple regression model:

$$Y_i \quad = \quad \alpha_1 + \beta_1 D_i + \boldsymbol{X}_i' \gamma_1 + \varepsilon_{i,1} \qquad \text{if } L_i = 1, \tag{1}$$

where $\boldsymbol{X}_i$ is a vector of individual and region specific controls, like birth cohort and county fixed effects and other pre-determined variables. $L_i$ is a binary variable for the observed employment status and a censoring indicator (or observation rule) in the wage regression equation: wages are only observed for those who are employed.[5]

As is well known, college education is endogenous. Therefore, parameters from an ordinary least squares (OLS) regression of Eq. (1) do not have causal interpretations. Assume that we have an instrument $Z$ which fulfills the exclusion restriction, is as good as random conditional on controls, and has predictive power for $D$. Then we might want to estimate the parameters in Eq. (1) by two-stage least squares (2SLS). Yet, this still does not yield estimates of causal effects of education on wages, as there is a second problem.

[5]We use the terms "employment" and "working for pay" interchangeably.

The parameters in Eq. (1) are only estimated for those who work. The employment status could be determined by the following model:

$$L_i \quad = \quad \alpha_2 + \beta_2 D_i + \boldsymbol{X}_i' \gamma_2 + \varepsilon_{i,2}. \tag{2}$$

Again, $D_i$ is endogenous and could be instrumented by $Z_i$. Yet, while the 2SLS coefficient $\widehat{\beta}_2$ is an estimate of the complier-specific causal effect of college education on working for pay (if the necessary assumptions for the instrument hold), this does not hold for $\widehat{\beta}_1$ in Eq. (1), as the effect of $D_i$ on $L_i$ differs across individuals, partly based on unobservables. Some individuals enter the regression sample in Eq. (1) irrespective of their realization of $D$, while others change their employment status due to $D$. This selection problem cannot be solved by the instrument $Z$ alone.

One solution to this problem would be to find an instrument $W$ that affects $L$ but not $Y$ and solve the problem using a Heckman-type selection estimator. That is, estimate the parameters in Eq. (2) and generate a term (e.g., an inverted Mills ratio) to be included in Eq. (1) as a control function, and then use an IV estimator in this augmented regression. This is proposed by Wooldridge (2010, chapter 19.6.2), although our model is slightly different in that the endogenous variable also enters the selection equation. Schwiebert (2015) explicitly derives the "Heckman model" with an endogenous variable that appears both in the selection and in the outcome equation. Yet, the assumption that this instrument $W$ can be excluded from Eq. (1) usually does not seem innocuous in the setting of employment and wages.

Thus, because finding sufficiently credible instruments is difficult, we abstain from this approach and instead suggest partial identification. Our solution benefits from us estimating MTEs instead of LATEs to generate more comprehensive and, arguably, more policy-relevant treatment effects. Estimating marginal college and non-college wages, as well as the corresponding employment probabilities needed to calculate the MTE, allows us to make credible assumptions to partially identify effects that take sample selection into account. In the next subsection we formalize the selection problem, while Section 2.3 briefly introduces the framework of MTEs and 2.4 proposes the main bounding solution.

## 2.2 College-induced selection into employment (CISE)

Like wages, employment has two potential outcomes, $L_i^1$ and $L_i^0$, with and without college education. In what follows, we suppress the individual indicator $i$. We start by ignoring the sample selection problem and only addressing the endogeneity of $D$ in Eq. (1). Assume for simplicity that $Z$ is a binary instrument.[6] Applying the IV approach to estimate the effect of college education on wages yields effects for those who go to college because of a change in the instrument (the compliers).

Table 1: Illustration of college-induced selection into employment (CISE) for the compliers to the instrument $Z$

| Realization of $Z$ ↓ | Type ① $L^1 = 1, L^0 = 1$ | ② $L^1 = 1, L^0 = 0$ | ③ $L^1 = 0, L^0 = 1$ | ④ $L^1 = 0, L^0 = 0$ |
|---|---|---|---|---|
| $Z = 1$ $\Rightarrow D = 1$ | $L = 1$ $Y = Y^1$ observed Ⓐ | $L = 1$ $Y = Y^1$ observed Ⓒ | $L = 0$ $Y = Y^1$ : – | $L = 0$ $Y = Y^1$ : – |
| $Z = 0$ $\Rightarrow D = 0$ | $L = 1$ $Y = Y^0$ observed Ⓑ | $L = 0$ $Y = Y^0$ : – Ⓔ | $L = 1$ $Y = Y^0$ observed Ⓓ | $L = 0$ $Y = Y^0$ : – |

*Notes:* Own illustration. This table categorizes compliers to the college instrument. The four columns show the four types of compliers that characterize how their employment status responds to the treatment $D$. The two lines exhibit two possible realizations of $Z$. Since we only consider compliers, the instrument $Z$ determines the treatment status $D$. Given the instrument assumptions hold, individuals randomly appear in either the upper or the lower part.

The subsample of compliers can be divided into four types depending on their employment status in response to college education. Each type can further be divided into two groups depending on whether the instrument is switched on or off. Table 1 illustrates the resulting eight cells of compliers.[7] The first type of college compliers (type 1, groups A and B in the table) works whether or not they have a college education; i.e., both potential outcomes $L^0$ and $L^1$ as well as the observed outcome $L$ take on a value of one, and wages of this type are observed. In the spirit of Machado (2017) we call these compliers the *always employed*. Women in group A are randomized into $D$ by the instrument $Z$, and their observed wage is $Y = Y^1$. Women in group B do not go to college because of the instrument, and their observed wage is $Y = Y^0$.

For type 2 compliers, we observe the wage only if $Z = 1$, as they work only if they have a college education (group C). For the counterpart of group C, type 2 with $Z = 0$, wages are not observed (group E in the table). The opposite holds for type 3, where wages are only observed in cases where $Z = 0$ (group D). No wages are observed for type 4, irrespective of the instrument.

[6] We will not need this later and our instrument is, in fact, continuous. Yet, assuming a binary instrument makes the exposition in this subsection much simpler.

[7] Note the similarity to the general setting of Heckman and Pinto (2018), who demonstrate identification results not with two different binary endogenous variables ($D$ and $L$) but with one ($D$) that can, however, indicate more than one (unordered) choice. In their language, the columns of the matrix in Table 1 are the response types or strata that are defined by counterfactual treatment choices (which are written in the cells).

For the researcher, the types are unobserved. Observed instrument value and employment status only allows the researcher to see that, for instance, an individual is either in group A or C (if $Z = 1$ and $L = 1$) or in group B or D (if $Z = 0$ and $L = 1$). Assume there were no type 2 or 3 individuals. The IV regression of Eq. (1) would then give us the causal effect of education on wages, conditional on being employed. Yet, if there are type-2 individuals, identification is not possible. We refer to the existence of type 2 individuals (i.e., the existence of individuals who only work conditional on having a college education) as college-induced selection into employment (CISE). Under CISE, the IV regression puts together groups A, B, and C and implicitly uses group B as a counterfactual outcome for type 2 without college education. Hence, as long as outcomes B and E (for type 1 and 2) differ, IV estimation is biased (this analogously applies to type 3). Due to the instrument $Z$, it is random whether individuals appear in the upper or lower panel of Table 1. Yet, the type of individuals, and thus whether wages are observed, is not random, and this issue is not solved by $Z$.

A simple IV estimation of Eq. (1) provides us with $E(Y^1 \mid L^1 = 1) - E(Y^0 \mid L^0 = 1)$ for compliers. This is equal to $E(Y^1 \mid \text{type 1 or 2}) - E(Y^0 \mid \text{type 1 or 3})$. Note that, with a binary instrument $Z$, these two parts can be identified separately, following the insights by Imbens and Rubin (1997).[8] The 2SLS estimation with a continuous instrument can only identify the difference. Nevertheless, since the conditioning set of the two expectations differs, the 2SLS estimand is not a meaningful number (i.e., may not be a causal effect).

Instead of $E(Y^1 \mid L^1 = 1) - E(Y^0 \mid L^0 = 1)$, we want to obtain $E(Y^1 \mid L^1 = 1, L^0 = 1) - E(Y^0 \mid L^1 = 1, L^0 = 1)$, the effect of education on wages for the *always employed*. Because the labor force participation rate is already fairly high for women without a college education (even though it is lower than for their college-educated peers) and because female unemployment is not a major problem, *always employed* constitute, arguably, the most interesting subpopulation of Table 1. This type of individuals is also studied in Lee (2009) and Machado (2017). To arrive at the effect for *always employed*, we start with

[8] For "$E(Y^1 \mid L^1 = 1)$ for compliers" start with

$$E(Y \mid Z = 1, D = 1, L = 1) = \frac{\phi_C}{\phi_C + \phi_{AT}} E(Y \mid Z = 1, D = 1, L = 1, C) + \frac{\phi_{AT}}{\phi_C + \phi_{AT}} E(Y \mid Z = 1, D = 1, L = 1, AT),$$

where $C$ stands for compliers, $AT$ stands for always-takers, and the $\phi$ are their respective shares. $E(Y \mid Z = 1, D = 1, L = 1)$ is observed in the data (more correctly, it can be estimated using the observed numbers $Y, D, Z, L$). With instrument exogeneity (first equality in the next equation) and monotonicity (second equality), $E(Y \mid Z = 1, D = 1, L = 1, AT) = E(Y \mid Z = 0, D = 1, L = 1, AT) = E(Y \mid Z = 0, D = 1, L = 1)$; this number is also observable. As the shares of types can also be estimated using the same assumptions, we can solve for $E(Y \mid Z = 1, D = 1, L = 1, C)$, which is $E(Y^1 \mid L^1 = 1)$ for compliers because $E(Y \mid Z = 1, D = 1, L = 1, C) = E(Y^1 \mid Z = 1, D = 1, L^1 = 1, C) = E(Y^1 \mid Z = 1, L^1 = 1, C) = E(Y^1 \mid L^1 = 1, C)$, where the second equality uses conditioning on compliers and the third, again, uses instrument exogeneity. Following the same logic, one can derive $E(Y^0 \mid L^0 = 1)$ for compliers, $E(Y \mid D = 0, Z = 0, L = 1, C)$, by purging out the respective outcome for never-takers from $E(Y|Z = 0, D = 0, L = 1)$.

separately estimating $E(Y^1 \mid L^1 = 1)$ and $E(Y^0 \mid L^0 = 1)$ and discuss how we deal with both numbers in turn.

$E(Y^0 \mid L^0 = 1)$ consists of type 2 and type 3 women. In what follows, we assume that no woman is of type 3 so that $E(Y^0 \mid L^0 = 1)$ directly yields $E(Y^0 \mid L^1 = 1, L^0 = 1)$. On the individual level, this assumption corresponds to $\Pr(L^1 \geq L^0) = 1$ (see Lee, 2009), meaning the incentive to work is monotonously increased by college education. This is comparable to the IV monotonicity assumption (for a general treatment on choice restrictions, consult Heckman and Pinto, 2018). We assume women who work without a college education but who would not work if they had a college education are insignificantly few and hence negligible. Certainly, this is not a trivial assumption. Yet, de Chaisemartin (2017) show that causal effects are still identified as long as the potentially existing type 3 women have the same average college returns as any arbitrary subset from the type 2 women, which is a weaker assumption. Nevertheless, we are not aware of findings from the literature that identify *sizeable* groups that systematically have negative causal effects of college education on employment (a sufficient condition for the existence of type 3 women). In Section 6.1 we discuss this assumption in more detail, provide evidence in favor of it and discuss quantitative implications of its violation.

We call $p^1 = \Pr(L^1 = 1)$ and $p^0 = \Pr(L^0 = 1)$. Then, without type 3 women, $p^0$ denotes the share of type 1 women in the population. The share of type 2 women is $\Pr(L^1 = 1) - \Pr(L^0 = 1) = p^1 - p^0$. The fraction of type 1 women among those who work with the treatment can be directly estimated by calculating $p_0/p_1$. In our data, 77.3 percent of the compliers in our sample are always working women which, arguably, justifies our focus on this group.

$E(Y^1 \mid L^1 = 1)$ consists of type 1 and type 2 individuals. We show in Section 2.4 how we follow Lee (2009) to make extreme assumptions on where the type 1 and 2 individuals are located in the distribution of $E(Y^1 \mid L^1 = 1)$ to generate upper and lower bounds of $E(Y^1 \mid L^1 = 1, L^0 = 1)$ and, ultimately, upper and lower bounds of $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$. Before we do this, we briefly sketch the MTE framework that we use. This framework helps us in two dimensions. First, it enables us to separately estimate $E(Y^1 \mid L^1 = 1)$ and $E(Y^0 \mid L^0 = 1)$ even with a continuous instrument. Note that we only assumed a binary instrument in this section for exposition and actually employ a continuous instrument. Second, while this section focused on the compliers to the instrument, the MTE framework allows us to derive effects for other groups too, such as the ATE and the ATT.

### 2.3 Estimating average college and non-college outcomes for marginal women

Here we briefly sketch out how conditional mean wages and employment probabilities with and without college education can be estimated for different groups of women who are at the margin of having a college education: $E(Y^1|\boldsymbol{X} = \boldsymbol{x}, U_D = u)$ and $E(Y^0|\boldsymbol{X} = \boldsymbol{x}, U_D = u)$, where $\boldsymbol{X}$ are observable variables and $U_D$ is an unobserved variable that measures the willingness to take the treatment (see below). Carneiro and Lee (2009) derive this potential outcome estimation approach and provide details. The difference between both potential outcomes is the MTE at point $U_D = u$, which is $MTE(U_D = u, \boldsymbol{X}) = E(Y^1 - Y^0|U_D = u, \boldsymbol{X})$. The basic framework of MTE builds on Björklund and Moffitt (1987) and Heckman and Vytlacil (2005, 2007). Carneiro et al. (2011) provide an in-depth application of MTE estimation and Cornelissen et al. (2016) a gentle introduction.

The starting point of the MTE approach is the generalized Roy model, consisting of the following:[9]

$$Y^1 = \boldsymbol{X}'\boldsymbol{\beta_1} + U_1 \quad \text{and} \quad Y^0 = \boldsymbol{X}'\boldsymbol{\beta_0} + U_0, \qquad \text{(potential outcome equations)}$$
$$Y = Y^0 + D(Y^1 - Y^0), \quad \text{and} \qquad \text{(switching equation)}$$
$$D = \mathbf{1}[D^* \geq 0], \quad \text{where } D^* = \boldsymbol{Z}'\boldsymbol{\delta} - V. \qquad \text{(selection equation)}$$

The potential outcomes are determined through observed characteristics $\boldsymbol{X}$ and unobserved factors $U$. The realized outcome $Y$ depends on the observed college decision $D$, which in turn depends on the latent desire for college education $D^*$. $D^*$ is a function of observed factors $\boldsymbol{Z}$ ($\boldsymbol{X}$ plus the instruments) and the unobserved component $V$. Following the conventional terminology, and without loss of generality, $V$ can be interpreted as the *unobserved costs* of college education, while $\boldsymbol{Z}'\boldsymbol{\delta}$ denotes the *observed benefits*. A woman decides to obtain a college education ($D = 1$) if her observed benefits $\boldsymbol{Z'\delta}$ exceed or equal her unobserved costs $V$. Her decision is denoted by the indicator function $\mathbf{1}[\cdot]$.

$V$ may correlate with ($U_1$,$U_0$), introducing selection into both the level of college education (the general endogeneity problem) and into its gains (heterogeneity in the returns). Using the same monotonous transformation on both sides (applying a cumulative distribution function), we can rewrite the treatment decision as $D = \mathbf{1}[D^* \geq 0] = \mathbf{1}[\boldsymbol{Z}'\boldsymbol{\delta} \geq V] = \mathbf{1}[\Phi(\boldsymbol{Z}'\boldsymbol{\delta}) \geq \Phi(V)] = \mathbf{1}[P(\boldsymbol{Z}) \geq U_D]$, where $P(\boldsymbol{Z}) \equiv P(D = 1|\boldsymbol{Z}) = \Phi(\boldsymbol{Z}'\boldsymbol{\delta})$ is the propensity score (i.e., the probability that a women enrolls in college given $\boldsymbol{Z}$). $U_D$ is uniformly distributed on the unit interval and measures an individual's relative rank within the distribution of $V$. As is standard in this literature, we refer to $U_D$ as the *distaste to education*. Note, however, that this is just a label and can be termed differently.

[9] We restrict the exposition to outcome $Y$, and it is equivalent for outcome $L$. Moreover, for the sake of exposition, we ignore the selection problem in this section.

Individuals who are indifferent between taking the treatment or not (i.e., those with $P(\boldsymbol{Z}) = U_D$) are shifted into the treatment by a marginal increase in $P(\boldsymbol{Z})$, exogenously induced by the instrument $Z$. These individuals identify treatment effects at different observed values of $P(\boldsymbol{Z})$. Because of the equality between $P(\boldsymbol{Z})$ and $U_D$, given the college decision and the propensity score, we can infer the individuals' otherwise unobserved ranks within $V$ (i.e., $U_D$).

Estimating the parameters in the Roy model requires the standard IV assumptions: conditional on $\boldsymbol{X}$, the instruments are independent of the unobserved factors; exclusion restriction; monotonicity; and relevance (see Vytlacil, 2002). As is common in the literature, we additionally assume additive separability between the observed and unobserved factors. Without this assumption, the slope of the estimated MTEs would potentially depend on which realization of $\boldsymbol{X}$ we condition on. Thus, assuming additive separability eases the interpretational and computational burden of the MTE approach. We now simplify the notation by keeping conditioning on $\boldsymbol{X} = \boldsymbol{x}$ implicit. In the estimation, the analogous step is to pre-process the outcome variable $Y$ and adjust for any differences in the observables $\boldsymbol{X}$. We do this by applying the semiparametric estimator suggested by Robinson (1988).

In contrast to the conventional "joint estimation approach" which is the standard case in the literature to estimate MTEs, we follow Carneiro and Lee (2009) and estimate both potential outcomes separately. To estimate the unobserved potential outcomes, we use the observed quantities

$$\begin{aligned} E(Y^1|U_D \leq p, D=1) &= E(Y|P(\boldsymbol{Z})=p, D=1) \quad \text{for college graduates, and} \\ E(Y^0|U_D > p, D=0) &= E(Y|P(\boldsymbol{Z})=p, D=0) \quad \text{for non-graduates.} \end{aligned}$$

This is the average outcome $Y$ among all those willing to take the treatment ($U_D \leq p$) and vice versa for $D = 0$. However, we want to have $E(Y^1|U_D = p)$, which we get by determining

$$E(Y^1|U_D = p) = E(Y^1|U_D \leq p, D=1) + p \times \frac{\partial E(Y^1|U_D \leq p, D=1)}{\partial p}, \tag{3}$$

where $p$ is any value on the unit interval. Please refer to Carneiro and Lee (2009) for a derivation of this formula. In Appendix B, we give a detailed exposition that provides the intuition for it.

Similarly, we can infer from $E(Y|P(\boldsymbol{Z}) = p, D = 0)$ and the selection equation that only individuals with larger values of $U_D$ than $p$ will not opt for college education. To get the mean wage of indifferent women without college education with a rank equal to $p$, we need to adjust $E(Y|P(\boldsymbol{Z}) = p, D = 0)$ as follows:

$$E(Y^0|U_D = p) = E(Y^0|U_D > p, D = 0) - (1-p) \times \frac{\partial E(Y^0|U_D > p, D = 0)}{\partial p}. \tag{4}$$

After the mentioned semiparametric pre-step that adjusts for observable differences, we estimate the two unknown components within Eq. (3) and (4) by local linear regressions that we evaluate at 100 values $p$ over the unit interval. Doing this gives us the level and the derivative of $E(Y|P(\boldsymbol{Z}) = p, D = 1)$ and $E(Y|P(\boldsymbol{Z}) = p, D = 0)$, respectively, at different values of $P(\boldsymbol{Z})$.

The difference between (3) and (4) yields the MTE. Combining this general exposition of MTE estimation with the problem outlined in the previous subsection, we call the MTE estimation that does not address CISE the *naïve marginal treatment effect* (NMTE): $NMTE(U_D) = E(Y^1 \mid U_D, L^1 = 1) - E(Y^0 \mid U_D, L^0 = 1)$.

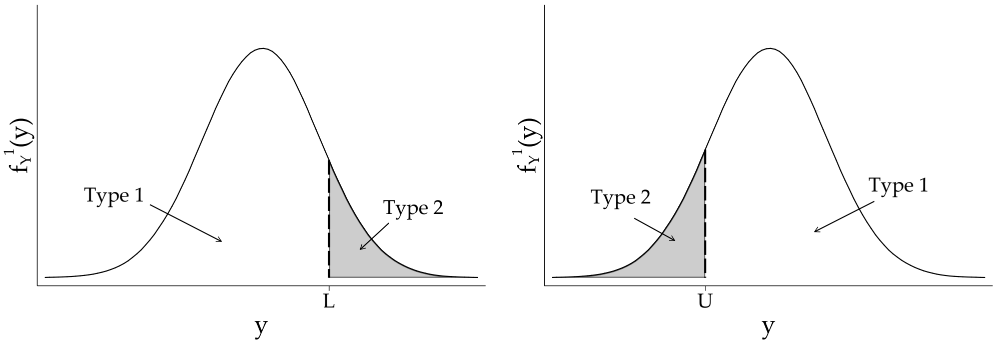
## 2.4 Lee bounds for MTEs

To estimate $E(Y^1 \mid L^1 = 1)$ with a continuous instrument, we need the MTE framework. Yet, to estimate a meaningful causal effect, we also need $E(Y^1 \mid L^1 = 1, L^0 = 1)$ so that we can calculate $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$. We know, however, that $E(Y^1 \mid L^1 = 1)$ consists of type 1 and type 2 women, with shares that can be estimated from the data. Figure 1 presents the general intuition of a bounding approach in the spirit of Lee (2009). We receive a lower bound of $E(Y^1 \mid L^1 = 1, L^0 = 1)$ if we assume that the wages of all type 2 women are located in the upper part of the distribution of $Y^1|L^1 = 1$. This assumption is illustrated in the left panel of Figure 1. Given an estimate of the distribution $f_{Y^1}$, and the shares $p_0$ and $p_1$, we can determine the threshold $L$ as the $\frac{p_0}{p_1}$ percentile of the distribution of $Y^1$ (i.e., $L = F_{Y^1}^{-1}\left[\frac{p_0}{p_1}\right]$, where $F_{Y^1}(\cdot)$ is the cumulative distribution function of $Y^1$).

Thus, the lower bound of $E(Y^1 \mid L^1 = 1, L^0 = 1)$, denoted $E_{LB}$ in what follows, can be calculated as the censored mean

$$E_{LB}(Y^1 \mid L^1 = 1, L^0 = 1) = \frac{\int_{-\infty}^{L} y f_{Y^1}(y) dy}{F_{Y^1}(L)}. \tag{5}$$

Figure 1: Stylized graph of upper and lower bound—General




(a) Lower bound (b) Upper bound

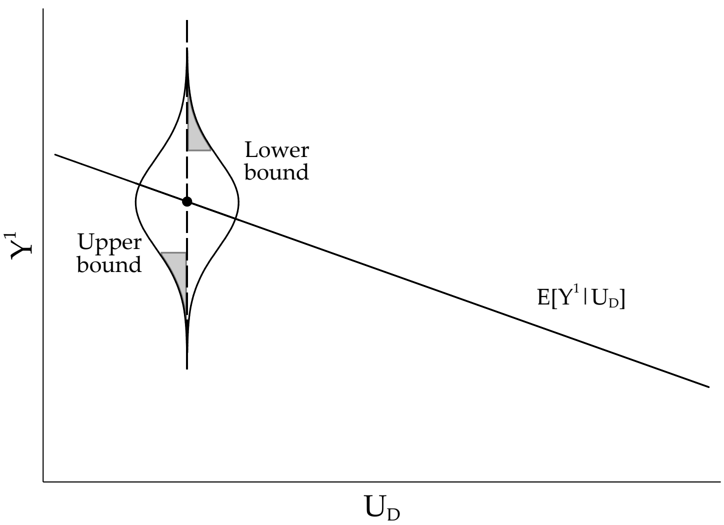
*Notes:* Own illustration.

Vice versa, $U = F_{Y^1}^{-1}\left[\frac{p_1 - p_0}{p_1}\right]$ for the upper bound which can be determined as

$$E_{UB}(Y^1 \mid L^1 = 1, L^0 = 1) = \frac{\int_U^\infty y f_{Y^1}(y)dy}{1 - F_{Y^1}(U)}. \tag{6}$$

We carry out this procedure separately for many different values of $U_D$ (see Figure 2), estimating a distribution $f_{Y^1}$ and the shares of type 1 and type 2 women conditional on $U_D$.

Figure 2: Stylized graph of upper and lower bound—MTEs

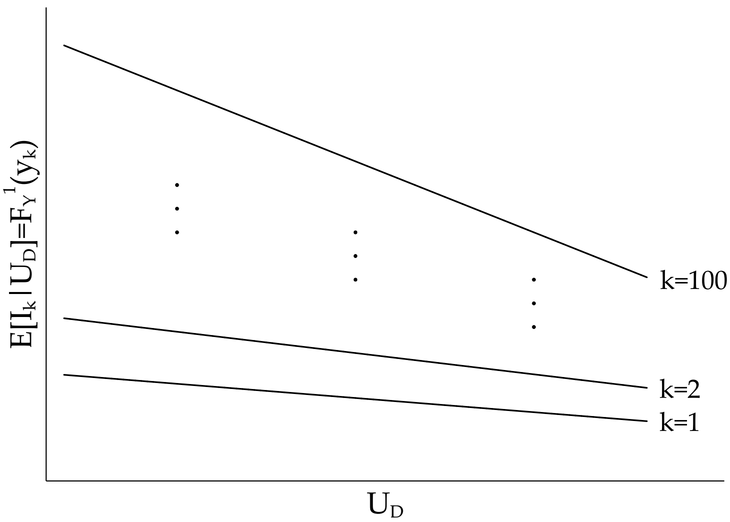



*Notes:* Own illustration.

We know that $F_{Y^1}(L|U_D)$ in the denominator of Eq. (5), conditional on $U_D$, equals $\frac{p_0}{p_1}$, but we also require the corresponding censoring threshold $L$, which is the wage at the quantile $\frac{p_0}{p_1}$. To get it, we proceed as follows. We define a set of indicator variables

$I_k = \mathbb{1}\{Y \leq y_k\}$ for many different values of $y$ distributed over the support of $Y$. Now, estimating $E(I_k) = Pr(Y \leq y_k)$ yields an estimate of the observed outcome's distribution function $F_Y$. To get an estimate of the distribution function of the potential treated outcome $Y^1$, we estimate $k$ separate marginal treatment response functions $E(I_k|\boldsymbol{X}, U_D, D = 1)$ just as $E(Y^1|\boldsymbol{X}, U_D)$ is estimated using the separate estimation approach outlined in Section 2.3 using the same set of control variables as in the main approach.[10] Doing this gives us a large number of values $F_{Y^1}(y_k|U_D)$ and thus, estimates of $F_{Y^1}$ at all places over the entire support of $U_D$ (see Figure 3). For each $U_D$ value, we determine the wage threshold $L$ as the wage $y_k$ for which $E(I_k|\boldsymbol{X}, U_D, D = 1) = \frac{p_0}{p_1}$ holds.

Figure 3: Stylized graph of $F_{Y^1}|U_D$ estimation




*Notes:* Own illustration. We do not restrict the estimates to be linear and instead estimate flexibly without any assumptions on the shape.

The estimation of $\int_{-\infty}^{L} y f_{Y^1}(y) dy$, the numerator of Eq. (5), conditional on $U_D$, proceeds similarly. We first define a set of variables $Y_k = Y \cdot \mathbb{1}\{Y \leq y_k\}$ for the same values $y_k$ as before. We estimate the marginal treatment response functions $E(Y_k|\boldsymbol{X}, U_D, D = 1) = \int_{-\infty}^{y_k} y f_{Y^1}(y) dy$. Now, we use $E(Y_k|\boldsymbol{X}, U_D, D = 1)$ for which the corresponding $E(I_k|\boldsymbol{X}, U_D, D = 1)$ equals $\frac{p_0}{p_1}$ for each value of $U_D$, and normalizing by $F_{Y^1}(L, U_D)$ gives us a lower bound of $E_{LB}(Y^1 \mid L^1 = 1, L^0 = 1, U_D)$.

The estimation of the upper bound is analogous. See Appendix D for a technical remark on the estimation of the upper bound. We also propose an alternative partial identification approach that combines these bounds with an additional linearity assumption on the MTE curvature. We wait explain the approach in Section 5, which follows the baseline results reported in Section 4.

[10]We choose $k = 100$, but this is not a decisive choice.

# 3 Policy environment and data

## 3.1 The German college expansion: Trends and causes

In this paper, we estimate effects of college education on employment and wages of women in Germany. The treatment dummy ($D$ in the previous section) is college education to earn a *Diplom* degree, which is a master's degree equivalent (there has been no bachelor's equivalent in the time under review). While earning a *Diplom* degree was formally possible after 4.5 years, the average study duration was 6 years (see Lundgreen et al., 2008). The alternative to college education in Germany typically is vocational training of 3 years part-time, on-the-job training and part-time vocational schooling with a small salary to cover the basic costs of living. Thus, we will estimate the effects of college education (relative to the next-best alternative, which, in almost all cases, is vocational training).

As exogenous variation ($Z$ in the previous section), we use variation across regions and time in the availability to study as a result of the German college expansion.[11] Between 1961, when the first women in the sample graduated from secondary schooling, and 1990, the number of college districts in Germany doubled from 27 to 54.[12] The number of overall (female) students increased from 208,000 (47,000) to more than 1 million (413,000) (see Figure A1 in Appendix A). While individuals deciding to go to college in the 1960s had to travel an average of nearly 40 km to the closest college, those in the 1980s only had to travel an average of less than 23 km (see Table 2). The basic idea behind instrumenting college education using its availability is that living closer to a college meant spending less money to attend (versus a college that was further away). For instance, cost may decrease because the commute is shorter and there is less need to move. Likewise, if high school graduates have an option to attend college close to them, they may become more aware of the educational choices provided to them.

There are four reasons behind the college expansion (see the review by Bartz, 2007). First, large parts of the population had low levels of education because it was previously not appreciated by the political systems of the Third Reich and Imperial Germany (see Picht, 1964). In neither of these systems did education constitute a natural way of climbing up the social ladder. Second, the outspread of new and more complex production technologies increased firms' demand for highly educated employees, not unlike the development that the U.S. experienced a couple of decades earlier, as described in Goldin and Katz

[11] The description of the college expansion and its reasons in this section closely follows Kamhöfer et al. (2019).

[12] The number of universities (German *Universitäten*) is taken from the Higher Education Compass of the German Rectors' Conference (2016). The number of students is taken from the German Statistical Yearbooks 1961–1991 (German Federal Statistical Office, 1991). Differences to Kamhöfer et al. (2019) emerge, as they only use the latter source and derive openings from the number of students. However, those differences affect only small colleges and the results are robust against this alternation. For the treatment of *Fachhochschule* institutes and other non-traditional institutions, we also refer to Kamhöfer et al. (2019).

(2008). Third, the prerequisite of expanding higher education was expanding secondary schooling. As shown in, for instance, Jürges et al. (2011) and Kamhöfer and Schmitz (2016), the number of academic secondary schools that (unlike the other school tracks at the time) allowed students to enroll in higher education had developed similarly to the number of universities. Fourth, in the time under review, Germany was still divided and education was both outcome and means in West Germany's efforts to outperform the competing East German socialist system (see Picht, 1964, and Jürges et al., 2011).

While those four reasons applied to all German states that were (and are) in charge of education policy, the pace of the educational expansion differed between the states.[13] In spite of expansive research on the subject, we were not able to detect a universal pattern shaping the expansion in higher education. Some states, such as Bavaria, reinstalled historical universities (e.g., the University of Bamberg was founded in 1647, downgraded to a theological seminary in 1803, closed in 1939, and reinstalled in 1972[14]). Other states, like North Rhine-Westphalia, founded universities in rural and industrial regions to serve as "advertisement for education" (*Bildungswerbung*; see NRW, 1971).[15] Moreover, the federal structure within the federal states caused negotiations and required cooperation between the district governments and the state government. Those negotiations introduced additional variation in the timing of the university openings (see, e.g., the account of Weisser, 2005, for anecdotal evidence on the role of election cycles and the local provision of construction sites). In the Supplementary Materials we provide more descriptive statistics on the college openings and the temporal and spatial variation in the ability to attend college. We also provide more information on the political process of college openings.

While the build-up of colleges and capacities demonstrates the relevance of the instrument, remaining questions may concern the validity and the exclusion restriction. We first discuss validity and then the exclusion restriction. Even though the political process of opening a college induced some degree of randomness, a major concern is that regions where colleges were opened differed compared to those without openings in regard to differences in levels of sociodemographic variables and dynamics and changes over time therein. In Table 3 (taken from Kamhöfer et al., 2019), we report descriptive statistics on the regional level at the year 1962 using German Micro Census data. Apart from population density (large cities already had colleges before the educational expansion, and very sparsely populated regions never opened any colleges), sociodemographic characteristics did not

[13] Some of these reasons might have been more pronounced in some states, e.g., because of a different industry structure. However, such initial differences should be captured by including a full set of fixed effects for the district of residence at the time of secondary schooling and district-specific linear trends in the empirical analyses.

[14] See `https://www.uni-bamberg.de/universitaet/profil/geschichte-und-tradition/geschichte/`.

[15] Initial differences in the affected population are, again, captured by including district fixed effects in the model.

Table 2: Distance to closest college districts (in km) by time of college decision

| | Average distance (in km) | | | Average student density | | |
|---|---|---|---|---|---|---|
| Decision to study: | 1958–70 | 1971–80 | 1981–90 | 1958–70 | 1971–80 | 1981–90 |
| 1st closest college district | 38.9 | 26.7 | 22.7 | 32.7 | 32.6 | 58.3 |
| 2nd | 73.6 | 52.4 | 46.7 | 34.2 | 34.2 | 56.4 |
| 3rd | 100.4 | 71.0 | 60.9 | 30.6 | 33.0 | 58.6 |
| 4th | 118.8 | 85.9 | 74.8 | 32.5 | 34.6 | 56.3 |
| 5th | 138.3 | 98.6 | 85.1 | 32.0 | 31.2 | 53.5 |
| 6th | 154.4 | 110.8 | 95.4 | 40.6 | 38.2 | 61.3 |
| 7th | 167.9 | 121.8 | 104.7 | 37.1 | 37.8 | 66.2 |
| 8th | 183.3 | 133.0 | 113.3 | 30.3 | 38.3 | 59.7 |
| 9th | 195.8 | 143.8 | 122.4 | 35.7 | 37.1 | 61.3 |
| 10th | 207.6 | 153.8 | 132.8 | 32.1 | 39.1 | 63.5 |

Notes: Own calculation based on German Federal Statistical Office (1991) and German Rectors' Conference (2016).

differ strongly among the groups of districts. Yet population density decreased over time in the districts that opened a college, while this was not the case in districts that did not. Note, however, that the magnitudes are small. Nevertheless, to account for these remaining differences, we include district fixed effects and district-specific linear time trends in the analysis.

Another potential problem is general equilibrium effects. Given the college expansion's large scale, all college students (e.g., through peer effects or teaching quality) may be affected in addition to marginally entering ones (see Bianchi, 2020). Moreover, the expansion may have had spillover effects on those without a college education (Moretti, 2004). In general, one may expect effects on local labor markets both in the short and long run. As is standard in the literature, we cannot fully rule this out; however, there is one potential concern that we can address—labor demand effects related to the college expansion. Building new colleges may induce local labor demand shocks, affecting both potential outcomes with and without college education ($Y^1$ and $Y^0$). To determine if the creation of new colleges had effects on the local wage structure around the time of the college opening, we exploit a data source that includes wages at that time.

We use individual-level data from the Sample of Integrated Labour Market Biographies (SIAB).[16]. The data set provides social security records on a daily basis for about 2 percent of all individuals who have either been in employment subject to social security or received unemployment benefits. We use the spell data to generate a panel data set that includes only women, the district they live in and their daily wages on January 1 of each year for the period 1975–1990; 1975 is the first year the data are available for.

[16]This study uses the factually anonymous Sample of Integrated Labour Market Biographies (version 1975–2017). Data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

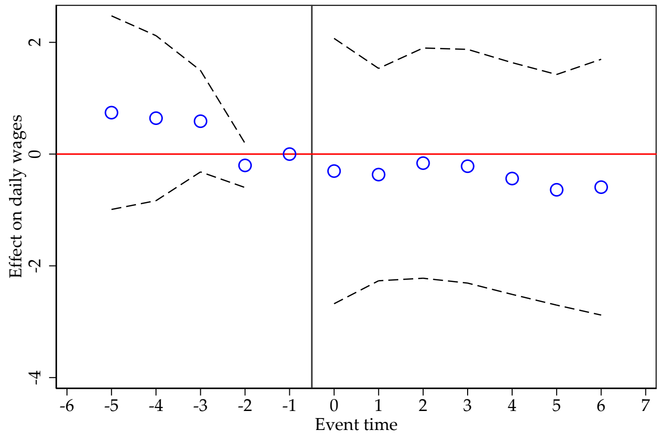
Table 3: Comparison of regions with and without college openings before college openings using administrative data

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | College opening... | | | | | |
| | before 1958 | | between 1958–1990 | | after 1990 or never | |
| | mean | s.d. | mean | s.d. | mean | s.d. |
| *Observations* | | | | | | |
| Number of regions | 27 | | 30 | | 190 | |
| *Sociodemographic characteristics* | | | | | | |
| Female (in %) | 53.0 | *(2.0)* | 53.0 | *(1.4)* | 52.9 | *(4.3)* |
| Average age (in years) | 37.2 | *( 1.1)* | 37.0 | *( 1.1)* | 36.6 | *( 1.9)* |
| Singles (in %) | 38.8 | *(2.5)* | 37.7 | *(2.3)* | 38.9 | *(4.6)* |
| Population density per km$^2$ in 1962 | 1381.9 | *(1076.7)* | 1170.1 | *(1047.3)* | 327.1 | *(479.7)* |
| Change in population density 1962 to 1990 | 1.6 | *(186.3)* | −71.0 | *(202.8)* | 31.5 | *(98.5)* |
| Migrational background (in %) | 2.7 | *(3.0)* | 1.6 | *(1.5)* | 2.1 | *(2.3)* |
| *Socioeconomic characteristics* | | | | | | |
| Share of employees to all individuals (in %) | 47.0 | *(3.6)* | 45.3 | *(4.2)* | 46.2 | *(5.2)* |
| Employees with income $>$ 600 DM (in %) | 27.3 | *(3.8)* | 24.8 | *(5.3)* | 25.9 | *(6.4)* |
| Employees by industry (in %) | | | | | | |
| – primary | 2.1 | *(5.2)* | 5.2 | *(5.2)* | 2.8 | *(5.5)* |
| – secondary | 52.9 | *(8.4)* | 54.7 | *(6.2)* | 54.3 | *(8.9)* |
| – tertiary | 45.0 | *(9.3)* | 40.1 | *(8.3)* | 42.9 | *(9.6)* |
| Employees in blue collar occup. (in %) | 53.6 | *(9.4)* | 59.0 | *(7.9)* | 56.5 | *(9.3)* |
| Employees in academic occup. (in %) | 22.0 | *(4.4)* | 17.5 | *(4.3)* | 20.3 | *(5.9)* |

*Notes:* This table is taken from Kamhöfer et al. (2019). Calculations are based on the Micro Census 1962; see Lengerer et al. (2008). Regions are defined through administrative Regierungsbezirk entries and the degree urbanization (Gemeindegrößenklasse) and may cover more than one district. College information is aggregated at the regional level, and a region is considered to have a college if at least one of its districts has a college. Calculations for population density and change in population density based on district-level data were acquired through personal correspondence with the statistical offices of the federal states. Data are available on request. The variables "employees in blue collar occup." and "employees in academic occup." state the shares of employees in the region in an occupation that is usually conducted by a blue collar worker/college graduate, respectively. Standard deviations (s.d.) are given in italics in parentheses.

Figure 4 shows event study results where we regress individual-level daily gross wages on indicators of college openings in the district. The event time is centered around the year of the college opening and equals zero for the opening year. We can consider 17 openings between 1975 and 1990 and allow for college opening effects five years before until six years after the opening. We also account for calendar year and federal state fixed effects. Please see the notes to Figure 4 for detailed information on the estimations. As this figure shows, wages seem to be largely unaffected by the college openings. In particular, wages did not increase due to college openings before the first students even graduated (i.e., within the first six years). Given that the vast majority of women at that time did not have a college degree, we interpret this as evidence that non-college wages were not affected by college openings in the short run.

Figure 4: Wage developments before and after college openings




*Notes:* Own illustration based on SIAB data, 1975–1990. The graph shows the results of event study regressions. The sample consists of individual-level daily wage data in West Germany. The event is a college opening which took place in 17 districts in the years 1975–1990. The dependent variable is gross daily wage as of January the 1st each calendar year, measured in prices of 1990. The sample mean is 52 Euros. We include binary indicators for the event times –5 to 6, where –1 is the reference category. Event times smaller than –5 and larger than 6 are captured by two additional indicators. The horizontal axis depicts years relative to a college opening in event time 0, and the vertical axis represents effects on daily wages relative to the year immediately before the opening, –1. Additional controls are calendar year and federal state fixed effects, individual age, and a migrational background indicator. 95 percent confidence bands are constructed using district-level-clustered standard errors. The sample includes all women who participate in the labor force. Restricting the analysis to women without a college degree yields similar results, as the majority does not have a college degree at this time.

In Section 3.2.2 we show how we collapse information on college openings and expansions over time into instruments. In brief, we will use variation in the set of university distance and capacity measures outlined in Table 2 to instrument the decision to enroll into college.

## 3.2 Data and descriptives

For the main analysis, we combine the administrative records the expansion of college opportunities with individual-level data from the Adult Starting Cohort of the German National Educational Panel Study (NEPS); see Blossfeld and von Maurice (2011).[17] In ten waves (starting in 2007/2008), the Adult Starting Cohort assesses longitudinal information on individual educational decisions, labor market participation and performance, current and past socioeconomic environment, and competence development for 17,000 individuals born between 1944 and 1989 (see LIfBi, 2015). Compared to other individual- or household-level surveys (e.g., the German Socio-economic Panel Study or SIAB that we use for explorative evidence), the main advantage of this data set is that it includes the district of residence at the time of secondary school graduation (e.g., the year an individual decides about going to college or not). This information is crucial for assigning individuals to the

[17]This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, 10.5157/NEPS:SC6:8.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

correct college distances and sizes they faced when making the decision, since where they live today might be determined by whether or not they decide to enroll in to college (e.g., if they move to a district to attend college; Card, 1995).

We use data from the fourth wave with interviews in 2011/12 from the Adult Starting Cohort and only keep individuals who graduated from a West German secondary school before 1990 because we focus on the West German educational expansion between 1960 and 1990.[18] We also drop individuals aged 65+ in 2012 and those without information on the district of residence before the college enrollment decision. The final sample includes labor market information for 3,409 women.

The explanatory variable is a binary indicator of college education, taking on a value of one if a woman has any higher educational degree, and zero otherwise. Dropouts are treated as individuals without a college education. As Table 4 reveals, 22 percent of the women in our sample have a college education.

#### 3.2.1 Dependent variables: Wages and employment

We consider two labor market outcomes: working for pay and hourly wages. The variable working for pay takes on a value of one if an individual reports to work five or more hours per week for pay at the time of the interview and zero otherwise. The non-working category includes voluntary and involuntary unemployment, participation in active labor market policies, and early retirement.[19] The working hours underlying the hourly wage are calculated based on the actual weekly hours (including overtime) spent with employed or self-employed work or, if missing, the contractual working hours. The summary statistics of the dependent variables in Table 4 reveal a rather low employment rate of women. On average, about 70 percent of the women report working for pay. What is more interesting is the employment gradient by college education. The difference between women with and without college education amounts to roughly 6.5 percentage points (or 9 percent).

Following the literature (e.g., Pischke and von Wachter, 2008), we define the gross hourly wage as gross monthly income divided by 4.3 weeks times the average weekly working hours. Income tax-free marginal employment jobs (with a monthly income up to 450 euros) are disregarded, and if an individual has more than one job at the time of the interview, we sum up all information (see the notes to Table 4 for all restrictions). The descriptive college gradient in earnings in Table 4 exhibits that working women with a college education earn, on average, more than one-third (8.82 euros) more than their peers without a college education.

[18] Wave 4 includes the most observations because of a refreshment sample. West Berlin is excluded from the analysis due to its special status.

[19] While individuals on parental leave would also fall into this category, the long-term perspective makes parental leave unlikely (only 5 percent of the individuals in the final sample are below age 40).

Table 4: Descriptive statistics and OLS estimates

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Mean values | | | OLS estimate of college education | |
| | all | w/o college | with college | coeff-icient | unit of measure |
| *Working for pay sample* | | | | | |
| Observations | 3,409 | 2,665 | 744 | | |
| **Working for pay (in %)** | 69.46 | 68.18 | 74.06 | 3.2** (2.2) | in percentage points |
| *Wage sample* | | | | | |
| Observations | 1,953 | 1,449 | 504 | | |
| Gross hourly wage (in €) | 24.97 | 22.69 | 31.51 | | |
| **Log gross hourly wage** | 2.93 | 2.84 | 3.17 | 0.374*** (0.046) | in log points |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. The top panel (working for pay sample) includes all women. Working for pay takes a value of one if an individual reports to work five or more hours a week and zero otherwise. The bottom panel (wage sample) only includes women with complete wage information. Earnings are in euros. The gross hourly wage is calculated as monthly gross income divided by 4.3 weeks times weekly working hours. To avoid outliers, we do not consider hourly wages derived from gross monthly incomes below 450 euros (the marginal employment threshold) or above the 99$^{\text{th}}$ quantile. Additionally, hourly wages below 5 euros are dropped, as they are likely caused by a measurement error. The weekly working hours include contractual working hours and hours worked overtime in employed and self-employed work using information on all jobs at the time of the interview. We allow for individuals to be employed despite missing wage information. Final outcomes measures used in the analysis are in bold font. Column 1 provides the unconditional mean values of the variables on the left. Columns 2 and 3 provides the mean values conditioning on individuals with and without college education, respectively. Column 4 provides the OLS estimates for the effect of college education on the outcome variables, and column 5 indicates their unit of measurement. Regressions include the standard set of control variables as discussed in Section 3.2.3 and as also used throughout all analyses that follow. The wage sample is conditioned on working for pay. District-clustered standard errors are in parentheses; $^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

In column 4 of Table 4, we also report results from OLS regressions of the dependent variables on college education and controls. The table shows that women with a college education are 3.2 percentage points more likely to work for pay than without a college education. The hourly wage increase associated with college education is 0.374 log points.[20]

### 3.2.2 Instruments

The college expansion increased both the number of colleges and the capacities of existing ones. A priori, it is not clear how to capture these changes in the most efficient way. The simplest way was chosen by the early studies that use college availability as an instrument. Card (1995) and Currie and Moretti (2003) use a binary variable to indicate a college in the district or the local labor market. Using a binary indicator, however, ignores information on capacity. Moreover, when deciding to enroll in a college, many individuals

[20]Which, transferred to an annualized percent, is $100 \times (e^{0.374/6} - 1) = 6.4$ percent per annum (p.a.).

not only take their home district into account but also nearby districts and even remote ones. Kamhöfer et al. (2019) boil down the information on the German college expansion into a single variable that sums up the distance-Kernel-weighted share of students per inhabitant over all districts. The calculation of this college availability index implicitly assumes some structure of how the college expansion may affect the college enrollment decision. Moreover, while the results are robust against several alternative specifications, such a definition, in principle, entails some researcher discretion.

Mullainathan and Spiess (2017) suggest the use of variable selection techniques adopted from machine learning to estimate the first stage in an IV setting (i.e.,, the propensity score estimation in the MTE estimation) and even point out the possible application of variable selection techniques when estimating the effect of a changing college environment on the college enrollment decision as a promising example. Following Belloni et al. (2012), the first-stage estimation can be seen as a prediction problem where many college expansion indicators potentially explain the individual's decision, but only few of these indicators actually matter. The list of potentially relevant indicators (e.g., college in home district, college in neighborhood district, more remote colleges, college size, variable interactions, etc.) can become very large if the researcher wants to put as little structure as possible on the problem and let the data speak. Empirically, we address this problem by running an auxiliary regression using the least absolute shrinkage and selection operator (Lasso) that minimizes the sum of squared residuals plus a penalty term for the number of parameters (see, e.g., James et al., 2013, for an introduction). The advantage of Lasso as a variable selection device is that parameters of variables deemed to have no significant predictive power are set to zero and will thus be discarded in what follows. The full analysis will then be performed with the subset of "relevant" variables.

As potential instruments, we consider ten variables on the distance to the ten college districts in closest proximity (if a college is in the home district, the first variable takes on a value of zero) and ten variables for the number of students per 1,000 inhabitants in these districts. Distances are measured from centroid to centroid of the 326 West German districts. The opening of new colleges in proximity to the home district changes the order of the closest college districts, while increased capacities are reflected in the student density measures. These potential instruments thereby account for both decreasing distances and increasing capacities. Additionally, we include all squared and cubic distances as well as all squared and cubic student densities of the college districts. Finally, we consider all pairwise interactions between the distance and the student density within the same order of proximity. All of these variables sum up to 70 potential instruments.

Table 2 reports average values of the potential instruments separately for different decades. The table shows that the distance to colleges has declined strongly over time. As previously mentioned, the average distance to the closest college has decreased over time from 38.9 to

22.7 kilometers. The selection of the instruments is based on the larger working-for-pay sample.[21] The chosen variables are, in that order, the distance to the $2^{nd}$ and $7^{th}$, the student density of the $7^{th}$, and the distance to the $9^{th}$ and $10^{th}$ nearest university.[22] In Section 6 we report results using other specifications of the instrument.

#### 3.2.3 Control variables

Our data include rich information on (pre-college) control variables; see Table A1 in the Appendix for variable definitions and means by college education. The first group of control variables, *general information*, includes the woman's year of birth (in the analysis captured by a full set of fixed effects), migrational background, and whether she lived in a rural area when she decided to go to college (also captured by fixed effects). *Pre-college living conditions* are all measured at the individual's age of 15. They cover indicators like parental unemployment and whether the parents have never been unemployed as well as indicators for the family's living arrangements (e.g., nuclear family or single parent). *Pre-college education* controls for the average grade points of the highest school track and grade repetition. Finally, *parental sociodemographics* include birth cohort fixed effects for the mother and father as well as indicators for their highest educational degrees.

Table A1 in the Appendix indicates that individuals with a college education had better grades in secondary school and less siblings and were more likely to live in an urban area and have no migrational background. College graduates are also more likely to have parents with jobs that require more skills. In terms of other variables, the difference between both groups seems to be smaller.

As stated above, in addition to cohort fixed effects, we also include a full set of district fixed effects (for the districts at the time of the college enrollment decision) in all regressions (see Mazumder, 2008, for an assessment of the importance of including this). Moreover, we also include district-specific linear time trends (see Stephens and Yang, 2014). Doing this addresses the potential problem that colleges are opened in certain regions with potentially different wage dynamics. Table 3 suggests that economically disadvantaged regions were affected by college openings. Quantitatively, the differences between districts with and without college opening are small, however. Nevertheless, district fixed effects and district-specific time trends should largely account for this.

[21] We use the Stata ado `lassoShooting` provided by Christian Hansen on his website. Any errors are our own responsibility. The Lasso regressions include full sets of control variables.

[22] This supports the finding of Kamhöfer et al. (2019) that university openings (reflected in the distance) are sufficient for a strong first stage. Thus, even if one is willing to believe in the exogeneity of university openings but not of their capacities, this does not alter our interpretation of the first stage.

# 4 Baseline results

## 4.1 Marginal treatment effects

Table A2 in the Appendix reports results of the propensity score estimation,[23] while Figure A2 shows that the propensity score contains sufficient variation to identify MTEs over a wide range of the unit interval. However, we only have a very small amount of identifying variation at very high values of the propensity score. That is, some individuals have such a high distaste for college education that even nearby colleges cannot persuade them to enroll. To avoid the possibility that marginal effects with low support are mistaken for reliable estimates, we trim the 5 percent of the observations with the lowest and the highest estimated propensity scores, respectively. Trimming the 5 percent of observations with the lowest propensity score results in disregarding marginal effects below the 1.8$^{\text{th}}$ percentile of the distaste-to-study index ($U_D$). Similarly, trimming the highest 5 percent of observations results in disregarding observations with an unobserved distaste for college education above the 67$^{\text{th}}$ percentile of the distribution. In other words, trimming the propensity score leads to a range of $U_D$ from 0.018 to 0.67. Neglecting observations outside this range is less critical than it appears at first glance. The trimmed quantiles at the upper end contribute little to the estimation of the most important conventional treatment effects (the ATT, the LATE, and any realistic policy-relevant treatment effect; see the argument in Section 4.3).

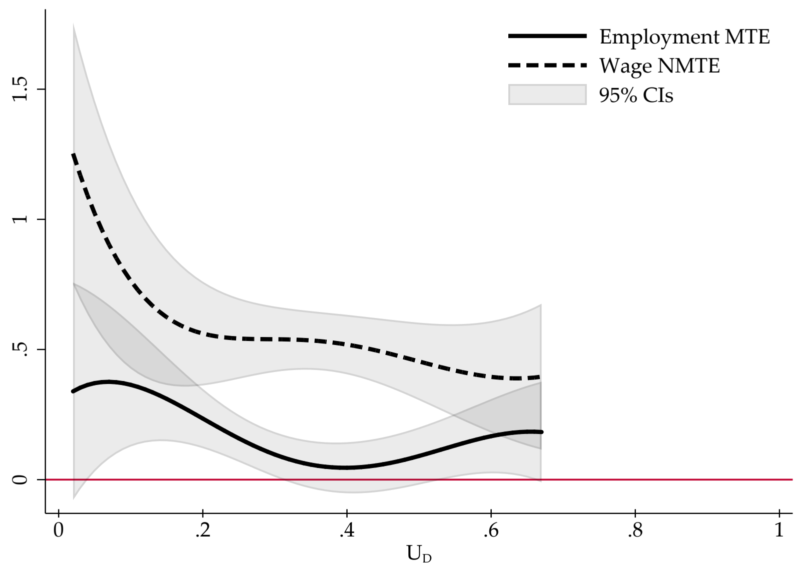
To illustrate the effects of the instruments for identification, we use the propensity score estimates, keep all control variables fixed at the sample means and only vary $Z$ to predict the propensity score. When using the average distances in the period 1971–1980 (reported in Table 2), we obtain a propensity score of 0.164, while we obtain 0.347 when using average distances in period 1981–1990 (note that all other variables such as time trends are fixed). Increasing the distances to 300 km for the second closest college and 500 for the seventh to tenth closest college, for instance, reduces the propensity score to basically zero, meaning that only extremely motivated individuals would choose college education in this scenario. Changing the average distance to 0 km for the second closest college and 100 for the seventh to tenth closest college increases the propensity score to 0.68. Of course, once the $\boldsymbol{X}$ variables are also varied, there is a multitude of combinations that may lead to very high or very low propensity scores.

Figure 5 presents the MTEs of college education on employment (solid curve) and wages (dashed curve). We start by interpreting the findings for employment where the vertical axis denotes percentage points. If one is willing to accept the assumptions put on the instruments, these effects are causal. For women with a high preference for college

[23] To assess the instruments' power, Table A3 in the Appendix states the instruments' partial $F$-statistic at the first stage of a conventional 2SLS approach.

education (a low $U_D$), there are very large effects (around 37.5 percentage points) and negligible effects (less than 3 percentage points) for women with a medium preference for college education. The pattern indicates a considerable effect heterogeneity, indicating that there is a substantial number of women who are induced to work as a result of having a college education (particularly for low $U_D$ values). This heterogeneity opens the door for a considerable selection bias when estimating the ATE of college education on wages for *always employed*.

Figure 5: (Naïve) marginal treatment effects on employment and wages




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. This figure depicts marginal treatment effects (MTEs) on the probability to work for pay (solid curve) and on hourly wages (dashed curve). The effects on wages are measured in log points. To have the correct percentage interpretation, note that the MTE on wages needs to be transformed by the following formula: $100 \times (e^{\text{MTE in log points}} - 1)$. The MTEs are evaluated at values of the propensity scores after trimming 5 percent on each tail of the propensity score distribution. Calculations are based on a local linear regression where the influence of the control variables was isolated using a semiparametric Robinson estimator (Robinson, 1988). The outcome is then evaluated by local linear regression separately by college education according to the method proposed by Carneiro and Lee (2009). Standard errors are calculated using the joint estimation approach.

The dashed curve in Figure 5 reports MTEs on wages (again, not taking CISE into account). The measure of the vertical axis is log points. One percent of a log point only approximately corresponds to a one percent increase. For larger log-point changes, this one-to-one mapping is not appropriate anymore (usually the uniformity between both scales is accepted up to 0.4 log points). Thus, if we interpret the effects (often above 0.4 log points) in percentage points, we need to transform the reported log-point effect first.[24] We call this dashed curve the *naïve* MTE in Section 2. Interpreting this effect without the theoretical reasoning and the quantitative support that there is (at least for some regions of $U_D$) a considerable share of type 2 women in our data, individuals with the highest taste for

[24]The transformation is $100 \times (e^{\text{log points}} - 1)$ = percentage points. When we report annualized returns, we divide the log points by the average study duration of six years first. A return of 1.2 log points, for instance, would translate into an annualized return of $100 \times (e^{1.2/6} - 1) = 22.1$ percent p.a.

college education earn 22.1 percent more per year of college education completed. This effect is strongly declining along the distribution of the taste for college (which is a sign of selection into gains). For the largest part of the distribution, the wage premium per year of college education is about 8.7 percent (0.5 log points in total for 6 years of college education). As discussed in Section 2, however, the level of these effects, as well as the slope, may be confounded by CISE.

Before moving to the bounds, we make a note on the p.a. effects of college education. These effects are per year of education completed and are not the increase in wages per year over the life cycle due to college education. As we pool women of all age groups between around 40 and 65, we can only identify average effects of education on wages at one point in time. The Roy model and our separate estimation of potential outcomes allow for different earnings dynamics of women with and without college education. Yet, due to sample size, we cannot identify life cycle effects or different effects for women with different tenure. However, we do not confound short- with long-run effects, as all women in our sample have been on the labor market for at least 10 years.

## 4.2 Lee bounds using MTEs

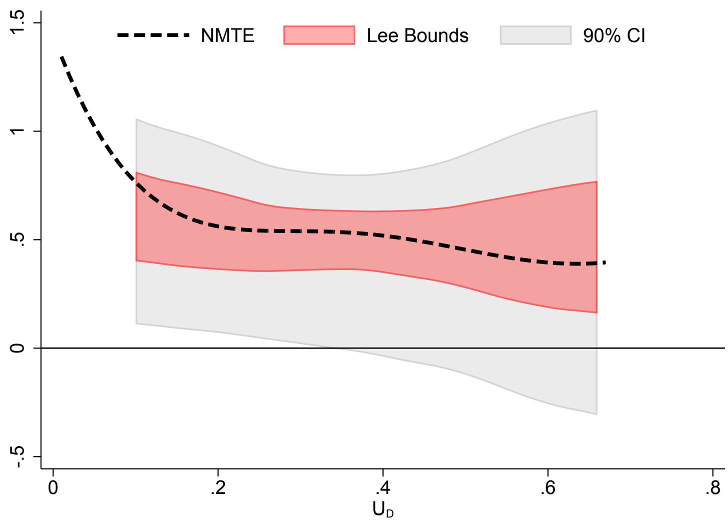
To account for CISE, we perform the bounds approach described in Section 2.4. If the identifying IV assumptions are valid, these bounds do not require any additional assumptions and hence are true worst-case bounds. Figure 6 plots these bounds around the estimated NMTE. The corresponding bootstrapped 90 percent confidence interval for the NMTE surrounds these bounds.[25]

For the lowest identified quantile of unobserved heterogeneity, $U_D = 0.1$, the bounds range from 0.40 to 0.81 log points. This translates into wage returns ranging from 6.9 to 14.5 percent p.a. Unfortunately, we cannot identify the Lee bounds for $U_D$ values below 0.1, as the identifying variation does not suffice to identify quantile MTE effects. We observe the widest bounds for $U_D = 0.66$, ranging from 0.16 to 0.77 log points (2.7 to 14.0 percent p.a.). Naturally, we observe the narrowest bounds for the area of $U_D$ where the employment effects are the lowest, from $U_D = 0.35$ to $U_D = 0.55$. For $U_D = 0.4$, for example, the bounds range from 0.35 to 0.63 log points (6.0 to 11.1 percent p.a.). Overall, the bounds are strikingly narrow, and most importantly, the wage effects of college education are positive throughout the identified area.

Despite the tightness of the bounds, they may still be quite wide to precisely pin down wage returns. In the following section, we try to complement these bounds by an additional restriction. Doing this is particularly important in order to also extrapolate the bounds to

[25] See Imbens and Manski (2004) for details on how to compute the asymptotic confidence interval that contains the treatment effect (as opposed to the identified set) with a certain probability.

Figure 6: Lee bounds for MTE




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. This figure depicts marginal treatment effects on hourly wages. *Naïve* effect (dashed curve) are taken from Figure 5 and area of adjusted Lee bounds as outlined in Section 2.4.

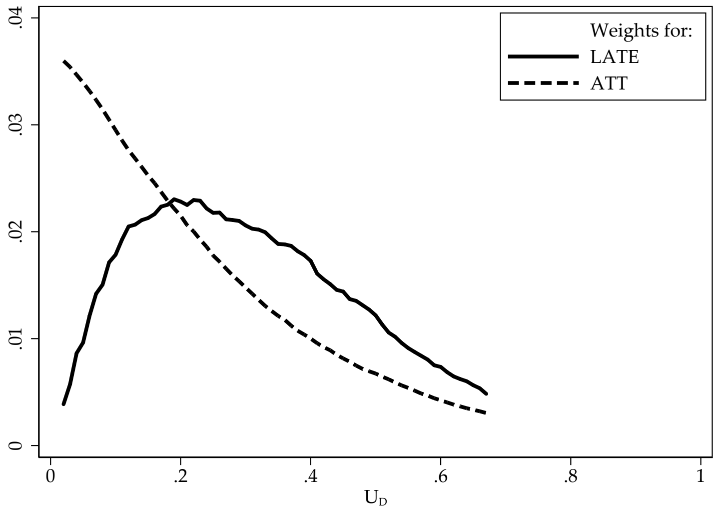
areas where $U_D < 0.1$, which is necessary to compute more general treatment effects such as the ATE and ATT.

### 4.3 Treatment parameters

Using MTEs, it is possible to calculate conventional treatment parameters such as the ATE; the ATT; and the LATE, for our instruments as given by 2SLS estimation. To calculate these parameters, we follow the approach of Heckman and Vytlacil (2007) and weight the MTEs at each quantile of $U_D$ with their relative contribution to the average effect. Figure 7 presents the weights for the ATT and the LATE. Given that these weights are basically zero outside the support, it will not matter quantitatively that we cannot identify MTEs on the unit interval. However, it might matter for the ATE, which employs constant weights on the unit interval. For the moment, we restrict the parameters to the ATT and LATE and refer to Section for an alternative approach that uses slightly stronger assumptions to both sharpen the bounds and extrapolate the MTE outside the identified interval so that we can calculate the ATE.

The weighted averages of the MTEs are reported in panel A of Table 5. Whenever possible, the corresponding bootstrapped standard errors are reported in parentheses. The first row shows the effects on working for pay. Both treatment parameters reveal economically and statistically significant effects on working for pay. On average, college education increases the fraction of treated women who are employed by 23.6 percentage points, higher than the estimated LATE of 17.6 percentage points. By definition, only type 2 women are

Figure 7: Treatment parameter weights conditional on the propensity score




*Notes:* Own illustration based on NEPS–Adult Starting Cohort data. The weights are estimated at values of the propensity scores after trimming 5 percent on each tail of the propensity score distribution.

responsible for these effects (assuming the effect of college education on the probability of being employed is always positive or zero but not negative, i.e., no type 3 women). Thus, these effects directly translates into shares of the types of women, which we presented in Table 1. The portion of women who work irrespective of college education—our type 1—ranges between 69 and 77 percent for the displayed treatment parameters (see panel B of Table 5). This finding shows that CISE is non-negligible in our sample—between 23 and 31 percent. Consequently, the impact on both the marginal effects and conventional treatment effects is likely to be non-negligible as well.

The first line in the panel on wage effects in Table 5 shows the wage effects based on the NMTEs and is thus not adjusted for potential CISE. There is some heterogeneity between the two treatment parameters. Women who actually attended college have, on average, a wage premium of 12 percent p.a., whereas the estimated LATE is 10 percent p.a. Both effects are significant.

Now we turn to the proposed bounds. Both treatment parameters are very similar and correspond to an annual return ranging between 5.8 and 12.5 percent. This spread is quite narrow given that these bounds do not require many assumptions. Moreover, even the lower-bound returns are significantly different from zero at a significance level of 10 percent, as reported by the $p$-values in the brackets.

From all considered treatment parameters, the annual college returns range from 6 to 13 percent—within any treatment effect, the spread barely exceeds 6 percentage points. The effect range is quite plausible and is in line with the effects found for men in the previous literature in Sweden, where Nybom (2017) finds an annual return of 7.5 percent (ATT). Our effects are considerably smaller than those identified for men in the U.S.; Carneiro

Table 5: Bounds on the treatment parameters

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | ATT | | LATE | |
| | log points | percent p.a. | log points | percent p.a. |
| *Panel A* | | | | |
| **Working** | | | | |
| Effect (in percentage points) | 23.6 | | 17.6 | |
| | (8.3) | | (6.9) | |
| **Wages** | | | | |
| Effect without considering CISE | 0.682 | 12.0 | 0.572 | 10.0 |
| | (0.242) | | (0.221) | |
| Lee bounds: | | | | |
| – Upper bound: | 0.705 | 12.5 | 0.690 | 12.2 |
| | (0.253) | | (0.231) | |
| – Lower bound: | 0.348 | 6.0 | 0.337 | 5.8 |
| | (0.241) | | (0.220) | |
| Treatment effects p-value: | [0.074] | | [0.063] | |
| *Panel B* | | | | |
| **Share of type 1 individuals** | | | | |
| | Treated | | Compliers | |
| $p^0/p^1$ | 0.687 | | 0.771 | |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. The even columns state the results in log points; the odd columns show the results transformed in percent p.a. The latter is calculated as $100 \times \left[e^{(\text{log points}/6)} - 1\right]$, assuming an average time spend in college of 6 years. The parentheses below the coefficients report cluster-bootstrapped standard errors based on 250 replications. As pointed out by Imbens and Manski (2004), the confidence interval for the treatment parameters within the identified sets are tighter than for conventional tests at any confidence level.

et al. (2011) report 24.2 percent p.a. (ATT) and Zimmerman (2014) assesses the effect at 22 percent per year for men (a marginal effect of U.S. individuals who are indifferent between college and the next best alternative). However, Zimmerman (2014) finds 0 percent p.a. for women.

Compared to the CISE-unadjusted naïve treatment parameters, our results also suggest that these naïve effects are most likely upper bounds. CISE will most likely induce type 2 women with high earnings potential to work, which will eventually lead to upwardly biased naïve effects.

Our results so far have two shortcomings that are both related to the fact that the $U_D$ support only ranges from approximately 0.1 to 0.65. First, we cannot derive the ATE, as observations outside of this range also have a weight clearly larger than zero when the ATE is calculated. Second, and more generally, we cannot determine MTEs outside of the observed interval (e.g., to learn about the potential effects of future reforms). College expansions are still an ongoing process in Germany, and future expansions are likely to address women with larger values of $U_D$, somewhere in the range of $U_D$ between 0.5 and 0.7. In the next section, we propose a refinement of the approach that combines the Lee bounds with assumptions on the shape of the MTEs. Doing this allows us to sharpen the bounds, extrapolate them on the full unit interval, and calculate ATEs.

# 5 Alternative approach: Combining Lee bounds with a MTE linearity assumption

## 5.1 Approach

Because MTE-based Lee bounds are very data hungry, it is not possible to identify them for the area $U_D < 0.1$. This subsection presents a complementary approach that allows us to not only bound the wage returns for small $U_D$ values but also further tighten them. This approach bounds type 1 wage returns by assuming a functional form of the selection-into-gains pattern that is compatible to Lee's approach.

We first go back to the discussion in Section 2.2 where we talk about estimating $E\left(Y^1 \mid L^1 = 1\right)$ and $E\left(Y^0 \mid L^0 = 1\right)$. $E\left(Y^1 \mid L^1 = 1\right)$ is a mixture of type 1 and type 2 women with shares that can be determined. Solving this for the type 1 women, we get

$$\begin{aligned} E\left(Y^1 \mid L^1 = 1\right) &= \overbrace{E\left(Y^1 \mid L^1 = 1, L^0 = 1\right)}^{\text{Type 1}} \frac{p^0}{p^1} + \overbrace{E\left(Y^1 \mid L^1 = 1, L^0 = 0\right)}^{\text{Type 2}} \frac{(p^1 - p^0)}{p^1} \\ \Leftrightarrow E\left(Y^1 \mid L^1 = 1, L^0 = 1\right) &= E(Y^1 \mid L^1 = 1)\frac{p^1}{p^0} - E(Y^1 \mid L^1 = 1, L^0 = 0)\frac{(p^1 - p^0)}{p^0}. \end{aligned} \tag{7}$$

By assumption of no type 3 women (which will be justified and relaxed below), we have that

$$E(Y^0 \mid L^0 = 1) = E(Y^0 \mid L^1 = 1, L^0 = 1). \tag{8}$$

Thus, by combining (7) and (8), we can write the causal effect of college education on wages for the always workers as

$$E(Y^1 - Y^0|L^1 = 1, L^0 = 1) = \left[E(Y^1 \mid L^1 = 1) - E(Y^0 \mid L^0 = 1)\right] \frac{p^1}{p^0} - \left[\underbrace{E(Y^1 \mid L^1 = 1, L^0 = 0)}_{\substack{\text{College outcome for type 2} \\ \text{individuals: } E(Y^1|②)}} - E(Y^0 \mid L^0 = 1)\right] \frac{(p^1 - p^0)}{p^0}. \tag{9}$$

Equation (9) has a MTE-equivalent if we only condition on $U_D$. We leave this implicit here. The equation makes explicit that if we could observe $Y^1$ for type 2 women, we could then estimate the effect for type 1 women. An alternative strategy to the Lee bounds for MTE could be to just plug in maximum or minimum observed values of $Y$ for the missing $Y^1$ of type 2 women (e.g., in the spirit of Horowitz and Manski, 2000). However, doing this would lead to larger bounds than in the previous section without making weaker assumptions. Instead, we try to make progress by relying on a linearity assumption on the shape of the MTE.

Although this approach is feasible for any (continuous) functional form, we demonstrate the approach for a linear selection into gains.[26] Apart from simplifying the formal expressions, the data and previous literature speak in favor of linearity (see details below). Assume that $MTE(\text{type } 1, U_D) = E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1, U_D) = a + bU_D$. That is, once we have identified the two unknown parameters $a$ and $b$, we then know the wage returns to college education for type 1 individuals. If appropriate, this restriction is a powerful step towards higher external validity, as it also identifies effects for regions with insufficient observations (Mogstad et al., 2018), such as $U_D < 0.1$ in the application at hand. Eq. (9) then simplifies to the following:[27]

$$a + bU_D = NMTE(U_D)\frac{p^1}{p^0} - \left[E(Y^1|L^1 = 1, L^0 = 0, U_D) - E(Y^0 \mid L^0 = 1, U_D)\right] \frac{p^1 - p^0}{p^0}.$$

[26] In Figure S8 in the Supplementary Materials we allow for a linear selection into losses. This slightly increases the upper bound but does not alter our conclusion.

[27] Note that in $p^0 = E(L^0|U_D)$ and $p^1 = E(L^1|U_D)$ we leave conditioning on $U_D$ implicit for sake of legibility.

Solving for the unknown $E(Y^1 \mid L^1 = 1, L^0 = 0, U_D)$, and making explicit that this expression now is in part determined by the two parameters for the linear MTE $\{a, b\}$, we get

$$E(Y^1|\text{type } 2, U_D, a, b) \quad = \quad NMTE(U_D)\frac{p^1}{p^1 - p^0} + E(Y^0|L^0 = 1, U_D) - (a + bU_D)\frac{p^0}{p^1 - p^0}.$$

Since we can directly estimate $NMTE(U_D)$, $E(Y^0 \mid U_D)$, $p^0 = E(L^0|U_D)$ and $p^1 = E(L^1|U_D)$, $E(Y^1|\text{type } 2, U_D)$ is determined once the parameters $a$ and $b$ are set and can then be plugged into Eq. (9) to get the effect for type 1.

Although a linear true selection into gains is not necessary for this approach (any parametric functional form will do), there are several reasons why linearity is a good approximation when bringing this strategy to the data:

- From an empirical point of view, linearity seems to be a fair assumption given the data at hand. This becomes apparent from the $U_D$ area where the college effect has no (statistically significant) effect on employment. We refer to this area as the "CISE-free" $U_D$ range, going from 0.35 to 0.55 in Figure 5. Linearity seems a good approximation for the functional form of the wage MTE returns in this CISE-free area. Although statistical insignificance does not necessarily imply a zero effect, the exact $U_D$ range in Figure 5 seems to play a minor role when looking at the selection-into-gains pattern.[28]
- Linearity has been found in the previous literature: semiparametric estimations of MTEs of college education on wages for men (where CISE probably does not play a large role)—that do not require any a priori fixed shape of the MTE—were fairly linear in Carneiro et al. (2011) and Nybom (2017). Kamhöfer et al. (2019) find linear cognitive skill and health returns. Cornelissen et al. (2018) require a parametric functional form and find that a linear pattern provides a better fit than a quadratic or cubic functional form.
- To extend the research frontier while keeping things as simple as possible, methodological papers frequently assume linearity. Brinch et al. (2017) suggest restricting effect heterogeneity to be linear in order to be able to estimate MTEs using a binary instrument (even with further assumptions on control variables or multinomial instruments, some form of parametric restriction remains). Other seemingly non-linear distributions are often also approximately linear. For instance, the normal selection

[28] Following this line of reasoning, differences between linear functional form and the observed functional form of the wage MTE returns outside the CISE-free area are driven by selection into employment. In Appendix Figure A3 we extrapolate the linear selection-into-gains pattern estimated in the CISE-free area to the entire identified $U_D$ range in order to bound type 1 wage returns. This third bounding approach lies in between the Lee bounds and the approach presented in this subsection in terms of both how restrictive the assumptions and how informative the bounds are.

model (see Carneiro et al., 2011, or Heckman, 1979) as a traditional workhorse model for MTEs implies an approximately linear MTE when the extreme tails of the distribution are discarded. Other examples for restricting the functional form of the selection into gains to linearity include Moffitt (2008), French and Song (2014), and Kowalski (2016).

While, in principle, the two parameters $a$ and $b$ (and thus the MTE) can take on any value, note that the resulting $E(Y^1 \mid \text{type } 2)$ cannot. For natural bounds (conditional on a linear MTE), we discard combinations of $a$ and $b$ that imply unreasonable values of $E(Y^1 \mid \text{type } 2)$. To apply this approach to the data, we take the extreme values of the Lee bounds to restrict the set of potential values for $b$:

- $E(Y^1 \mid \text{type } 2, U_D)$ may not take on lower values than those implied by the corresponding value under the lower Lee bound. Thus, we discard all combinations of $a$ and $b$ that lead to lower values.
- Correspondingly, $E(Y^1 \mid \text{type } 2, U_D)$ may not be larger than implied by the upper value for the Lee bounds. Thus, we discard all combinations of $a$ and $b$ that lead to those values where this is the case.

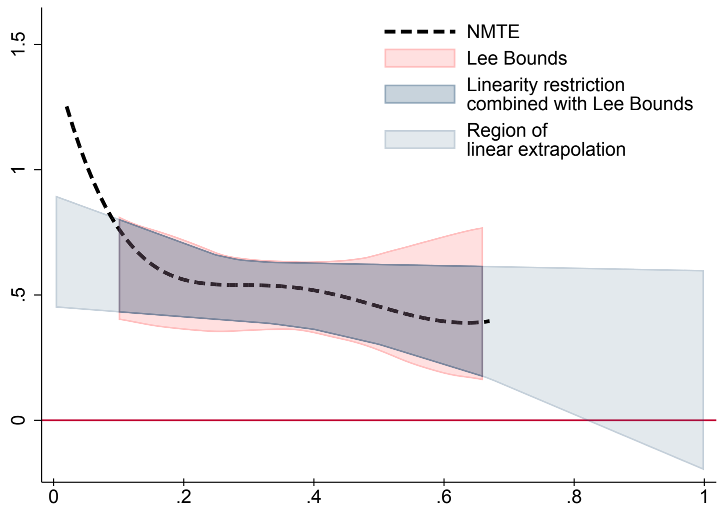
In essence, we want to try any $\{a, b\}$ parameter combination, where $a \in [-\infty, \infty]$ and $b \in [-\infty, 0]$. Because this is computationally unfeasible, we conduct a grid search and evaluate only specific $\{a, b\}$ values at $T_a$ discrete grid points for $a$ and $T_b$ points for $b$ to find the space of admissible MTE lines $MTE(\text{type } 1, U_D) = a + bU_D$.[29] As there are potentially many $\{a, b\}$ combinations that fulfill these restrictions, this strategy bounds the CISE-accounted value of $MTE(\text{type } 1, U_D)$ rather than suggesting a point estimator.

## 5.2 Results

Figure 8 shows the sets of all admissible marginal effect lines for type 1 individuals for these two scenarios: Lee bounds alone and in combination with a presumed linear MTE. Neglecting the fact that this procedure does not take potential sampling errors into account, the dark gray-shaded area in Figure 8 is the set of admissible linear MTE lines allowed by the existing degree of CISE in our data. This means that all linear lines that lie inside this set are possible marginal effect lines for type 1 women. The light gray-shaded area uses extrapolations of these bounds on the unit interval.

[29] We further a priori restrict compatible values of $a$ and $b$ using prior reasoning on potential MTE lines. Doing this yields the following intervals: $a \in [-\max(Y|D = 1), \max(Y|D = 1)]$, $b \in [-2\max(Y|D = 1), 0]$. Moreover, we set $T_a = 200$ and $T_b = 200$ such that $T = 4,000$. However, the resulting bounds are highly insensitive to the choice of the number of iterations $T$. For the sake of clarity and generality, we did not include this in the main text. Also, since these intervals are continuous in nature, we discretize them into small bins of 0.01 increments, essentially conducting a grid search.

Figure 8: Admissible linear MTE lines allowed by the degree of college-induced selection into employment (CISE) in the sample




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. This figure depicts the naïve marginal treatment effect (NMTE) for wages and corresponding bounds using linearity restrictions on the hypothesized shape of the MTE for *always employed*. These bounds are shown by the gray lines that comprise the marginal wage effect. Under this strategy, all linear MTEs that fit within these lines (highlighted by the gray-shaded area) are considered possible.

Figure 8 shows several findings. First, on the interval between 0.1 and 0.65, the bounds are slightly sharper than the Lee bounds alone. Second, the linearly restricted Lee bounds are especially interesting for small $U_D$ values, where the Lee bounds with no further restrictions are not identified. While NMTE estimates are particularly large for $U_D < 0.1$, up to 1.2 log points (22.1 percent p.a.), the bounds lie in a range between 0.45 and 0.9 log points (7.8 and 16.2 percent p.a.). This finding indicates that CISE would result in greatly misleading NMTE wage estimates, especially for always employed women with a high unobserved desire to attend college. Third, we also observe that the upper bound is always quite large and the lower bound only crosses the zero line at very large values of $U_D$. As expected, the bounds grow very large at $U_D$ values of more than 0.8 and should not be used to draw inferences on this population. However, this is a population that is hard to motivate to attending college anyway and most likely do not respond to any reasonable education policy.

We now use this bounding approach to derive all conventional treatment parameters (ATE, ATT, and LATE) and report their results in Table 6. As can be inferred from Figure 8, the bounds on ATT and LATE do not differ significantly between the two bound versions (Lee bounds only and Lee bounds with linearity; compare Table 5). They are now slightly narrower, with the ATE a bit smaller with bounds of 3.9 to 11.6 percent. This finding is expected due to the negative slope of the MTE and thus smaller effects at higher values of $U_D$ that had not been taken into account for the ATT and LATE. Nevertheless, these effects are still quite large.

Table 6: Bounds on the treatment parameters

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | ATT | | LATE | | ATE | |
| | log points | percent p.a. | log points | percent p.a. | log points | percent p.a. |
| Lee bounds in combination with a linear MTE: | | | | | | |
| – Upper bound: | 0.731 | 13.0 | 0.685 | 12.1 | 0.661 | 11.6 |
| – Lower bound: | 0.395 | 6.8 | 0.371 | 6.4 | 0.228 | 3.9 |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. The even columns state the results in log points, and the odd columns show the results transformed in percent p.a. The latter is calculated as $100 \times \left[e^{(\text{log points}/6)} - 1\right]$, assuming an average time spend in college of 6 years. As the linearly restricted MTEs are non-pivotal statistics, we do not report standard errors for them.

# 6 Robustness checks

## 6.1 Assumption of no type 3 individuals

In Section 2.2, we assumed that no women is of type 3; i.e., there are no women that do *not* work in as a result of having a college education but do work if they have a college education ($\Pr(L^1 \geq L^0) = 1$). The main reason for this assumption is simplicity. Without this assumption, the exposition of the problem is more complicated and we would need bounds for a second unobservable outcome. Here, we aim to show that this assumption may be justified and, maybe more important, our results do not crucially hinge on it. A major threat to this assumption is assortative mating and the infamous "Mrs. Degree". Women who entered college have been more likely to meet future husbands with a higher earnings potential, allowing the women to reduce their labor market participation compared to not having a college education. We now (*i*) provide some evidence in favor of this assumption based on results along the dimension of unobserved heterogeneity, (*ii*) provide results along the dimension of observed heterogeneity, and (*iii*) show that allowing for minor deviations from this assumption does not affect the findings significantly.

(*i*): Returning to the employment MTE in Figure 5, these results indirectly lend support to our assumption of no type 3 women. While we cannot test this, we can at least see that over the entire range of $U_D$, there is no area with a negative average effect of education on employment, providing stronger evidence than if we only found a positive 2SLS coefficient because we observe a whole distribution of mean effects.

(*ii*): In Table 7 we provide evidence on heterogeneity along observable characteristics. We report 2SLS coefficients of the effect of college education on working for pay for different subgroups based on all control variables used in the analysis. To do this, we go through all

Table 7: 2SLS estimates of college education on employment by subgroups

| | (1) | (2) | (3) |
|---|---|---|---|
| | | 2SLS effect of college education on employment if characteristic | |
| Characteristic | Share (in percent) | does not apply | applies |
| **Full sample benchmark** | | | |
| All women | | 0.171** (0.077) | |
| **General information** | | | |
| Born after 1960 (mean year of birth) | 53.2 | 0.190** | 0.124 |
| | | (0.092) | (0.136) |
| Grew up in rural area | 23.7 | 0.088 | 0.480** |
| | | (0.088) | (0.223) |
| Migrational background | 8.7 | 0.171** | 0.178 |
| | | (0.083) | (0.163) |
| **Pre-college living conditions (R = respondent)** | | | |
| Any siblings | 84.2 | 0.088 | 0.185 |
| | | (0.132) | (0.120) |
| First born | 30.7 | 0.182** | 0.152 |
| | | (0.088) | (0.097) |
| Age 15: not lived in core family | 9.3 | 0.189** | −0.002 |
| | | (0.083) | (0.156) |
| Parents: both employed at R's age of 15 | 45.5 | 0.209** | 0.115 |
| | | (0.090) | (0.086) |
| Parents: both never unemployed until R's age of 15 | 61.0 | 0.218** | 0.134 |
| | | (0.099) | (0.091) |
| **Pre-college education** | | | |
| Final school grade: excellent or good | 22.5 | 0.204** | 0.049 |
| | | (0.086) | (0.109) |
| Repeated at least one grade | 18.6 | 0.197** | 0.072 |
| | | (0.086) | (0.120) |
| **Parental sociodemographics** | | | |
| Parents: vocational training | 40.6 | 0.123 | 0.253*** |
| | | (0.080) | (0.092) |
| Parents: further training | 12.6 | 0.183** | 0.094 |
| | | (0.085) | (0.132) |
| Parents: college education | 6.0 | 0.174* | 0.150 |
| | | (0.089) | (0.184) |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. The sample includes information on 3,409 observations. Each row of the table gives the effect of college education on the probability of being employed, differentiated by the individual characteristic given on the left. All characteristics are binary, taking on a value of one if the characteristic applies and zero otherwise. The mean value of the characteristic (i.e., the share values of 1) is given in the first column. Instruments and control variables are the same as in the baseline specifications of the paper. District-clustered standard errors are in parentheses; $^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

variables in Table A1 and group them to binary variables in order to split the sample in this dimension. Table 7 reports results for all subgroups with at least a share of 5 percent in the sample (as too small numbers make IV regressions infeasible). For nearly all subgroups, we find striking evidence for non-negative effects of college education on working for pay.

Of course, this is not a proof of non-existence of type 3 women. Thus, we continue in part (*iii*) by calculating back-on-the-envelope results when we allow for a small share of type 3 women.

(*iii*): In an extended analysis in Appendix C we first derive Eq. (9) without assuming that $p_D = 0$. The new equation is defined through two unobserved expected values, one for type 2 and one for type 3 women. We then, for observed parts of the formula, plug in sample averages from our data set. For the unobserved factors, we define plausible maximum and minimum values to calculate bounds. Thus, we do not carry out another regression analysis but instead plug in numbers to learn about the potential magnitudes of the effects that small shares of type 3 women can have on the bounds. We calculate these bounds for three sets of plausible values: a baseline version that assumes $p_D = 0$ but uses sample averages for the observed factors, version 2 that assumes $p_D = 0.01$ and thus a share of type 3 women who represent 1 percent of the population, and version 3 that assumes $p_D = 0.02$. The resulting bounds for type 1 wages are the following (see Appendix C):

– Baseline, $p_D = 0.00$: 0.37–0.70,

– Version 2, $p_D = 0.01$: 0.34–0.71, and

– Version 3, $p_D = 0.02$: 0.30–0.72.

The baseline bounds are in line with those in Table 5 (this is how we calibrated the model). Apparently, the bounds only slightly increase when there is a small number of type 3 women. The intuition is obvious: with a small number of type 3 individuals, the influence of even large effects in this group on the results are comparably small.

The underlying model in Appendix C is a very simple one. Its goal is not to produce numbers that can be taken at face values but to instead show that slight deviances of our assumption of no type 3 women have only minor consequences. Of course, this also holds for 3, 4, or 5 percent shares of type 3 women. Yet, based on the results above, we have no reason to assume that the shares of type 3 women are substantially larger.

## 6.2 Instrument

While the machine-learning-based instrument selection has certain advantages (see Section 3.2.2), the lack of theoretical foundation with regard to the chosen subset of instruments calls for robustness checks. This subsection discusses the sensitivity of the estimated treatment parameters for two alternative instrument specifications. Table 8 compares the ATT Lee bounds from the baseline specification in column 1 (taken from Table 5) with the two other specifications in columns 2 and 3.

We first use the distances to the ten nearest universities without further instrument selection. While using less relevant distances (those not chosen by the Lasso) adds noise to the fitted value of the college decision, it is more intuitive to consider all ten college distances rather than a theoretically unjustified subset. Column 2 of Table 8 reports the resulting ATT when using the ten instruments in the selection equation. The upper and lower Lee bounds are 0.716 and 0.387 log points, respectively. This is very close to Lee bounds for the baseline specification, with an upper bound of 0.705 and a lower bound of 0.348.

Next, we boil down the college expansion into a single instrument, the college availability index $Z_{it}$, that links university distances and their capacities. It is used in Kamhöfer et al. (2019) and is formally calculated as

$$Z_{it} = \sum_{j}^{326} K(\text{dist}_{ij}) \times \left( \frac{\#\text{students}_{jt}}{\#\text{inhabitants}_{jt}} \right).$$

For each year $t$, we calculate the college capacities as the number of students per inhabitant across all $j = 1, ..., 326$ West German districts. When there is no college in district $j$ in year $t$, the capacities are zero, resulting in a heat map of college availability. While this heat map only varies over time, we weight the exposure to the college expansion by the kernel-weighted distance ($K(\text{dist}_{ij})$) from an individual's $i$ home district (the district of high school graduation) to each district $j$. Using the Gaussian kernel with a bandwidth of 250 km as additional weight ensures that districts in closer proximity to the home district matter relatively more compared to more remote districts.[30] Column 3 of Table 8 shows that the Lee bounds are, again, close to the baseline specification, with an upper bound of 0.720 and a lower bound of 0.516.

The Supplementary Materials includes the LATEs corresponding to the ATTs in Table 8. Figure A4 in the Appendix plots the LATE weights for all three instrument specifications (similar to Figure 7). The range of $U_D$ that can and cannot be identified using the college expansion do not seem to depend on the specification of the instrument.

All in all, the alternative instrument specifications deliver bounds estimates that are very comparable to the ones presented in the previous section. Again, all estimates suggest positive and economically meaningful wage increases due to women having a college education.

[30]While the home district has a kernel weight of 0.4, this weight reduces to 0.37 for colleges 100 km away from the home district and 0.24 for colleges 250 km away (for perspective, the north–south distance from Hamburg to Munich is about 600 km, and the east–west distance from Berlin to Cologne is 480 km). Thus, for more remote colleges to matter as much as nearer ones, they have to be relatively more important in terms of their capacities. Kamhöfer et al. (2019) show that the results are stable with respect to changes in the instrument specification (e.g., using a different kernel bandwidth or using a binary college indicator instead of capacities).

Table 8: ATT bounds for alternative instrument specifications

| | (1) | (2) | (3) |
|---|---|---|---|
| | Instrument specification | | |
| | Lasso (baseline) | All ten dist. (*iv*) | Index (*v*) |
| Lee bounds | | | |
| – Upper bound | 0.705 | 0.716 | 0.720 |
| | (0.253) | (0.204) | (0.212) |
| – Lower bound | 0.348 | 0.387 | 0.516 |
| | (0.241) | (0.178) | (0.209) |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. This table only states ATTs; the Supplementary Materials include the LATE results. Column 1 repeats the baseline results from Table 5 for comparison.

## 6.3 Earnings as an outcome

One might want to use monthly earnings as an outcome instead of hourly wages. While wages are usually seen as a productivity indicator, earnings are a compound measure of productivity and labor supply. Moreover, CISE is not a problem when zero earnings for women who do not work are included in the data; however, interpretations of the results do change.

Column 1 in Table 9 reports 2SLS estimates of college education on monthly earnings (in euros, including zero earnings for non-employed women). We report 2SLS for simplicity. As before, the ATTs are very similar to the 2SLS effects. Having graduated from college increases earnings by about 1,700 euros per month for the compliers. Compared to the mean earnings for those who work of 2,400 euros, this translates into an increase of 71 percent, which is a p.a. effect of 9.3 percent and is thus in line with the previous baseline results using wages, even though this is the average effect for type 1 and type 2 women. Compared to the overall mean of 1,526 euros, this amounts to an increase of 13.3 percent p.a. This effect is larger than our baseline bounds, as it is a compound of three effects: an increase in hourly wages, an increase in labor force participation, and—as can be seen in column 4—an increase in working hours, conditional on positive hours.

Column 4 reports the effect of college education on weekly working hours for the employed. College education increases weekly working time by about 7 hours, about one-quarter of the mean (see column 5). This increase indicates that there is a college-induced labor supply effect that goes beyond the binary decision to work and is due to women switching from part- to full-time work. Given that this effect is estimated conditional on working, this is a naïve effect, again, in the terminology of Section 4.

Table 9: 2SLS results for monthly earnings and weekly working hours

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Monthly earnings (in euros, incl. 0) | | | Weekly working hours | |
| | 2SLS coeff. | Mean incl. 0 | Mean if >0 | 2SLS coeff. | Mean if > 0 |
| College education | $1,695.7^{**}$ | 1,526 | 2,411 | $7.1^{**}$ | 28.4 |
| | (774.0) | | | (3.0) | |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. Monthly earnings are measured in euros and are set to zero if someone reports not to work (working hours missing). Earnings is based on 2,937 observations; 1,885 have a non-zero income. Weekly working hours are defined in the same way as for the hourly wage calculation; see note to Table 4 (2,316 observations). District-clustered standard errors are in parentheses; $^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

# 7 Conclusion

Unlike most of the previous literature, this paper explicitly aims to identify causal wage returns to college education for women. Looking at women's returns is interesting not least because women have benefited most from educational expansions in recent decades (see the tightening or even reversal of gender education gaps). Aside from selection into college education, an additional identification problem emerges. We call this manifestation of a well-established sample selection problem *college-induced selection into employment* (CISE), where individuals self-select into employment because of their increased education. This problem may be particularly severe for women and renders any instrument assignment non-random in wage samples.

To study the problem, we combine information on the college expansion in Germany with representative data from the German National Educational Panel Study. Our outcome measures are primarily the hourly wage and an indicator for working for pay. We find that women with a college education are, on average, more likely to be employed by more than 17 percentage points. This finding translates into a share of 22 percent of all women who—due to being college educated—spill into the labor market without having an appropriate counterfactual in any conventional IV wage regression. This finding also clearly demonstrates that the CISE margin of labor supply is potentially hazardous for any naïvely estimated treatment effect, also calling into question our naïve IV wage returns of roughly 10 percent p.a.

To address this problem explicitly, we develop a novel partial identification approach. We complement the bounds by Lee (2009), which we adjust to non-random assignment of college education, with restrictions on certain behavioral patterns along the essential heterogeneity, more concretely, restrictions like linearly downward-sloping marginal treatment effects (MTEs). Doing this helps us learn about the true effect for always employed women, those who are employed irrespective of their college education. Applying these bounds to our data, we find that the female relative wage return per year of college is 6.0–12.5 percent (average treatment effects on the treated). These bounds may appear wide, but they are considerably tighter than the heterogeneity in causal estimates (not suffering from CISE) on male labor market returns of college education in the literature. Of course, the bounds may also be tighter because German women are a more homogeneous sample. The estimated treatment parameters without taking CISE into account are closer to the upper bound than to the lower bound, which may suggest that women who work as a result of having a college education have an above-average earnings potential.

Our results suggest considerable college wage returns for German women. Average effects are considerable, and we find positive MTEs throughout the entire MTE distribution, implying beneficial (gross) average returns—also for those who have a lower willingness to obtain a college education. The results also point at selection into gains where women who benefit most from education also have the highest willingness to obtain more education, but the bounds do not rule out that there is no selection into gains.

Note, however, that our sample of women is older than in comparable studies interested in life cycle income. Unfortunately, due to sample size, we cannot estimate the effects for younger women only. Bhuller et al. (2017), for instance, find that returns to education (for men) increase over the lifespan, implying that we should expect our estimates to be larger than in comparable studies with a younger sample. Of course, comparability is not easy anyway, as this is the first paper that addresses CISE-based heterogeneity in individual returns.

# References


Angrist, J. D. and Krueger, A. B. (1992). Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery. NBER Working Papers 4067, National Bureau of Economic Research.

Angrist, J. D. and Krueger, A. B. (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Arellano, M. and Bonhomme, S. (2017). Quantile Selection Models: With an Application to Understanding Changes in Wage Inequality. *Econometrica*, 85(1):1–28.

Bar, M., Kim, S., and Leukhina, O. (2015). Gender Wage Gap Accounting: The Role of Selection Bias. *Demography*, 52(5):1729–1750.

Bartz, O. (2007). Expansion und Umbau – Hochschulreformen in der Bundesrepublik Deutschland zwischen 1964 und 1977. *Die Hochschule*, 16(2):154–170.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6):2369–2429.

Bhuller, M., Mogstad, M., and Salvanes, K. (2017). Life Cycle Earnings, Education Premiums and Internal Rates of Return. *Journal of Labor Economics*, 35(4):993–1030.

Bianchi, N. (2020). The Indirect Effects of Educational Expansions: Evidence from a Large Enrollment Increase in University Majors. *Journal of Labor Economics*, forthcoming.

Björklund, A. and Moffitt, R. (1987). The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics*, 69(1):42–49.

Blau, F. D. and Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, 55(3):789–865.

Blossfeld, H.-P. and von Maurice, J. (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, 14(2):19–34.

Blundell, R., Costa Dias, M., Meghir, C., and Shaw, J. (2016). Female Labor Supply, Human Capital, and Welfare Reform. *Econometrica*, 84(5):1705–1753.

Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007). Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds. *Econometrica*, 75(2):323–363.

Blundell, R. and MaCurdy, T. (1999). Chapter 27 - Labor Supply: A Review of Alternative Approaches. volume 3 of *Handbook of Labor Economics*, pages 1559 – 1695. Elsevier.

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, 125(4):985–1039.

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In Christofides, L., Grant, K., and Swidinsky, R., editors, *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, pages 201–222. University of Toronto Press.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(6):2754–2781.

Carneiro, P. and Lee, S. (2009). Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality. *Journal of Econometrics*, 149(2):191–208.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2016). From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions. *Labour Economics*, 41(C):47–60.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, 126(6):2356–2409.

Currie, J. and Moretti, E. (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *The Quarterly Journal of Economics*, 118(4):1495–1532.

de Chaisemartin, C. (2017). Tolerating Defiance? Local Average Treatment Effects Without Monotonicity. *Quantitative Economics*, 8(2):367–396.

Dong, Y. (2019). Regression Discontinuity Designs With Sample Selection. *Journal of Business & Economic Statistics*, 37(1):171–186.

Eckstein, Z. and Lifshitz, O. (2011). Dynamic Female Labor Supply. *Econometrica*, 79(6):1675–1726.

Eckstein, Z. and Wolpin, K. I. (1989). Dynamic Labour Force Participation of Married Women and Endogenous Work Experience. *The Review of Economic Studies*, 56(3):375–390.

Fernandez-Val, I., van Vuuren, A., and Vella, F. (2018). Nonseparable Sample Selection Models with Censored Selection Rules: An Application to Wage Decompositions. IZA Discussion Papers 11294, Institute for the Study of Labor (IZA).

French, E. and Song, J. (2014). The Effect of Disability Insurance Receipt on Labor Supply. *American Economic Journal: Economic Policy*, 6(2):291–337.

German Federal Statistical Office (various issues, 1959–1991). Statistisches Jahrbuch für die Bundesrepublik Deutschland. Technical report, German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Rectors' Conference (2016). Hochschulkompass – List aller Hochschulen. Technical report, German Rectors' Conference (Hochschulrektorenkonferenz), Bonn.

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

Goldin, C., Katz, L., and Kuziemko, I. (2006). The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, 20(4):133–156.

Gronau, R. (1974). Wage Comparisons-A Selectivity Bias. *Journal of Political Economy*, 82(6):1119–1143.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Heckman, J. J. (2010). Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature*, 48(2):356–398.

Heckman, J. J. and Pinto, R. (2018). Unordered Monotonicity. *Econometrica*, 86(1):1–35.

Heckman, J. J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6 of *Handbook of Econometrics*, chapter 71, pages 4875–5173. Elsevier.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association*, 95(449):77–84.

Huber, M. and Melly, B. (2015). A Test of the Conditional Independence Assumption in Sample Selection Models. *Journal of Applied Econometrics*, 30(7):1144–1168.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Imbens, G. W. and Rubin, D. B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies*, 64(4):555–574.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning – with Applications in R*. Springer, New York Heidelberg Dordrecht London.

Jürges, H., Reinhold, S., and Salm, M. (2011). Does Schooling Affect Health Behavior? Evidence from the Educational Expansion in Western Germany. *Economics of Education Review*, 30(5):862–872.

Kamhöfer, D., Schmitz, H., and Westphal, M. (2019). Heterogeneity in Marginal Non-monetary Returns to Higher Education. *Journal of the European Economic Association*, 17(1):205–244.

Kamhöfer, D. A. and Schmitz, H. (2016). Reanalyzing Zero Returns to Education in Germany. *Journal of Applied Econometrics*, 31(5):912–919.

Kowalski, A. E. (2016). Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments. NBER Working Papers 22363, National Bureau of Economic Research.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lemieux, T. and Card, D. (2001). Education, Earnings, and the 'Canadian G.I. Bill'. *Canadian Journal of Economics*, 34(2):313–344.

Lengerer, A., Schroedter, J., Boehle, M., Hubert, T., and Wolf, C. (2008). Harmonisierung der Mikrozensen 1962 bis 2005. Technical report, GESIS-Methodenbericht 12/2008, GESIS–Leibniz Institute for the Social Sciences, German Microdata Lab, Mannheim.

LIfBi (2015). Startkohorte 6: Erwachsene (SC6) – Studienübersicht Wellen 1 bis 5. Technical report, Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

Lundgreen, P., Scheunemann, J., and Schwibbe, G. (2008). *Berufliche Schulen und Hochschulen in der Bundesrepublik Deutschland 1949-2001, Teil II: Hochschulen*, volume 8. Vandenhoeck und Ruprecht.

Machado, C. (2017). Unobserved Selection Heterogeneity and the Gender Wage Gap. *Journal of Applied Econometrics*, 32(7):1348–1366.

Mazumder, B. (2008). Does Education Improve Health? A Reexamination of the Evidence from Compulsory Schooling Laws. *Economic Perspectives*, 32(2):2–16.

Moffitt, R. (2008). Estimating Marginal Treatment Effects in Heterogeneous Populations. *Annales d'Économie et de Statistique*, 91/92:239–261.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters. *Econometrica*, 86(5):1589–1619.

Moretti, E. (2004). Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data. *Journal of Econometrics*, 121(1-2):175–212.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Mulligan, C. B. and Rubinstein, Y. (2008). Selection, Investment, and Women's Relative Wages over Time. *The Quarterly Journal of Economics*, 123(3):1061–1110.

NRW (1971). Sachstandsbericht des Ministers für Wissenschaft und Forschung. Technical report, Ministry of Science and Research of the state of North Rhine-Westphalia (NRW), March 2, 1971, Düsseldorf.

Nybom, M. (2017). The Distribution of Lifetime Earnings Returns to College. *Journal of Labor Economics*, 35(4):903–952.

Parro, F. (2012). International Evidence on the Gender Gap in Education over the Past Six Decades: A Puzzle and an Answer to It. *Journal of Human Capital*, 6(2):150–185.

Pekkarinen, T. (2012). Gender Differences in Education. *Nordic Economic Policy Review*, 1:165–197.

Picht, G. (1964). *Die deutsche Bildungskatastrophe: Analyse und Dokumentation*. Walter Verlag.

Pischke, J.-S. and von Wachter, T. (2008). Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation. *The Review of Economics and Statistics*, 90(3):592–598.

Riphahn, R. and Schwientek, C. (2015). What Drives the Reversal of the Gender Education Gap? Evidence from Germany. *Applied Economics*, 47(53):5748–5775.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Schwiebert, J. (2015). Estimation and Interpretation of a Heckman Selection Model with Endogenous Covariates. *Empirical Economics*, 49(2):675–703.

Stephens, M. J. and Yang, D.-Y. (2014). Compulsory Education and the Benefits of Schooling. *American Economic Review*, 104(6):1777–1792.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, 70(1):331–341.

Wagner, G., Frick, J., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127(1):139–169.

Weisser, A. (2005). 18. Juli 1961 – Entscheidung zur Gründung der Ruhr-Universität Bochum. Technical report, Internet-Portal Westfälische Geschichte.

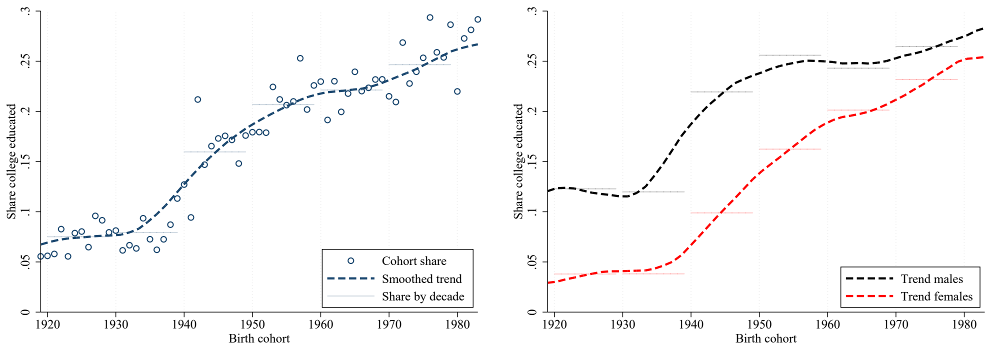
Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books. The MIT Press.

Zimmerman, S. D. (2014). The Returns to College Admission for Academically Marginal Students. *Journal of Labor Economics*, 32(4):711 – 754.

# Appendix

## A. Figures and tables

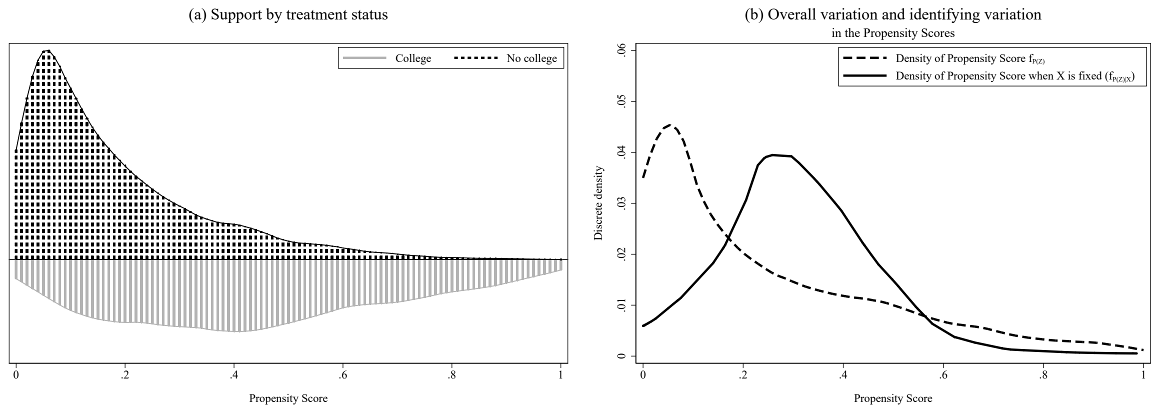
Figure A1: College education by cohort in Germany




(a) Overall trend (b) Trend by gender

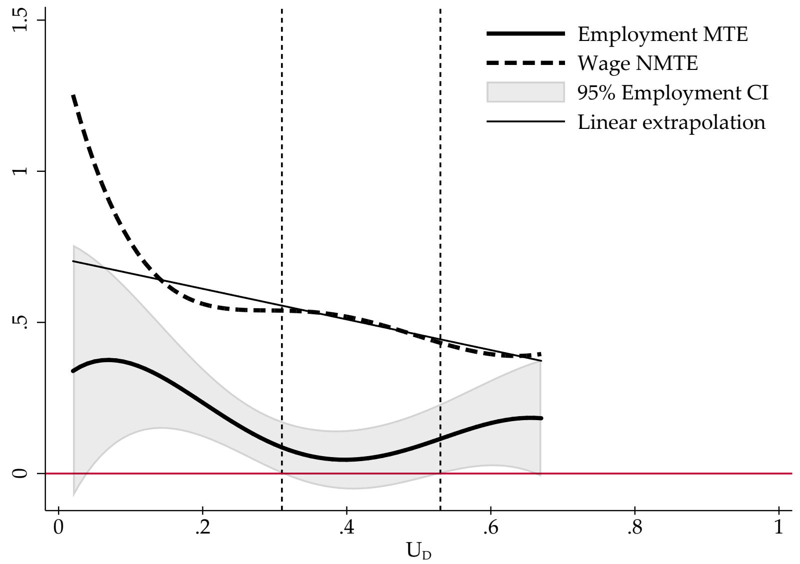
*Notes:* Own illustration based on representative data from the German Socio-economic Panel Study (SOEP), see Wagner et al. (2007).

Figure A2: Distribution and marginal distribution of the Propensity Score and college decision



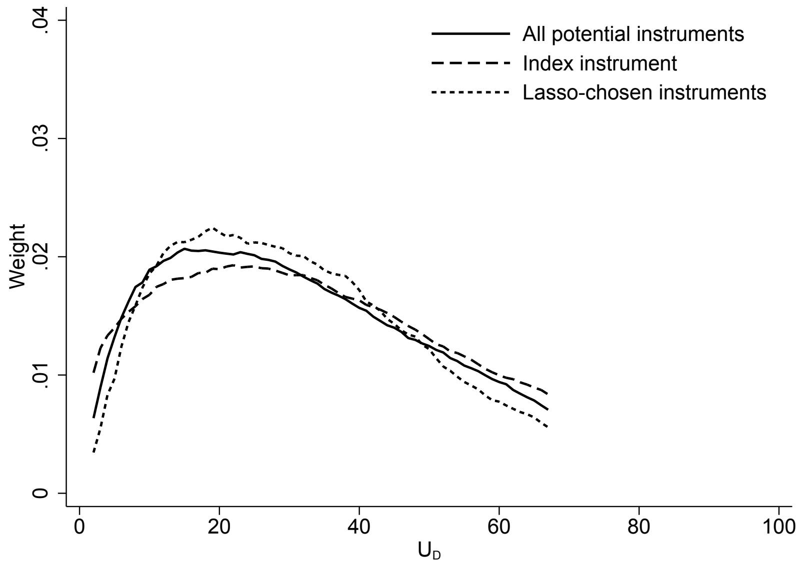

*Notes:* Own illustration based on NEPS–Adult Starting Cohort data. The left panel depicts the estimated density of the propensity score separately by college degree. The right panel shows the effective variation in the propensity score that we use for identifying marginal effects. In this measure, we integrated out all influences from all control variables but the instruments. As marginal effects may get unstable at the tails, we restrict our analysis to the middle part of the propensity score distribution, by omitting the 5 percent lowest and 5 percent highest propensity score observations. The respective cutoff levels are .018 (lower tail) and .670 (upper tail).

Figure A3: Extrapolation of the selection into wage gains using CISE-free MTE estimates




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. This graph depicts marginal treatment effects on the probability to work for pay (solid line) and on hourly wages (dashed line). The effects on wages are measured in log points. See notes to Figure 5 and the text for details on the calculation.

Figure A4: LATE weights for different instrument specifications




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. The instruments correspond to these discussed in the robustness checks. Calculation similar to Table 7.

## Tables

Table A1: Control variables and means by college degree

| Variable | Definition | Respondents: with college degree | Respondents: w/o college degree |
|---|---|---|---|
| **General information (R=respondent, M=mother, F=father)** | | | |
| Age (FE) | R's age in 2011/12 | 53.2 | 52.7 |
| Born abroad | =1 if R was born abroad | 0.01 | 0.01 |
| No native speaker | =1 if mother tongue is not German | 0.00 | 0.00 |
| **Pre-college living conditions** | | | |
| Siblings | Number of siblings | 1.63 | 1.86 |
| Older siblings | Number of older siblings | 1.10 | 1.27 |
| First born | =1 if R was the first born in the family | 0.34 | 0.30 |
| Married before college | =1 if R got married before the year of the college decision or in the same year | 0.01 | 0.00 |
| Parent before college | =1 if R became a parent before the year of the college decision or in the same year | 0.00 | 0.00 |
| Age 15: single parent | =1 if R was raised by single parent | 0.06 | 0.05 |
| Age 15: patchwork | =1 if R was raised in a patchwork family | 0.01 | 0.03 |
| Age 15: orphan | =1 if R was an orphan at the age of 15 | 0.01 | 0.02 |
| Age 15: M employed | =1 if M was employed at the R's age of 15 | 0.48 | 0.51 |
| Age 15: M never unemployed | =1 if M was never unemployed until the R's age of 15 | 0.60 | 0.63 |
| Age 15: F employed | =1 if F was employed at the R's age of 15 | 0.94 | 0.90 |
| Age 15: F never unemployed | =1 if F was never unemployed until the R's age of 15 | 0.99 | 0.97 |
| **Pre-college education** | | | |
| Final school grade: excellence | =1 if the overall grade of the highest school degree was excellent | 0.04 | 0.02 |
| Final school grade: good | =1 if the overall grade of the highest school degree was good | 0.25 | 0.19 |
| Final school grade: satisfactory | =1 if the overall grade of the highest school degree was satisfactory | 0.13 | 0.19 |
| Final school grade: sufficient or worse | =1 if the overall grade of the highest school degree was sufficient or worse | 0.01 | 0.01 |
| Repeated one grade | = if needed to repeat one grade in elementary or secondary school | 0.15 | 0.18 |
| Repeated two or more grades | =1 if needed to repeat two or more grades in elementary or secondary school | 0.02 | 0.01 |
| **Parental sociodemographics** | | | |
| M: year of birth (FE) | Year of birth of the R's M | 1931.2 | 1932.1 |
| M: migrational background | =1 if M was born abroad | 0.06 | 0.05 |
| M: vocational training | =1 if M has vocational training | 0.26 | 0.25 |
| M: further training | =1 if M has further job qualification (e.g., *Meister* degree) | 0.05 | 0.02 |
| M: college education | =1 if M has a college degree | 0.04 | 0.01 |
| M: education unknown | =1 if M's highest educational degree is unknown | 0.01 | 0.03 |
| M: occupation unknown | =1 if M's occupation is unknown | 0.46 | 0.46 |
| F: year of birth (FE) | Year of birth of the R's F | 1928.2 | 1929.2 |
| F: migrational background | =1 if F was born abroad | 0.06 | 0.05 |
| F: vocational training | =1 if F has vocational training | 0.22 | 0.33 |
| F: further training | =1 if F has further job qualification (e.g., *Meister* degree) | 0.14 | 0.10 |
| F: college education | =1 if F has a college degree | 0.14 | 0.03 |
| Number of observations (employment sample) | | 744 | 2,665 |

*Notes:* Information taken from NEPS–Starting Cohort 6, wave 4. Mean values refer to the employment sample. In the case of binary variables, the mean gives the percentage of 1s. FE = variable values are included as fixed effects in the analysis.

Table A2: Marginal effects of the main propensity score estimation

| Variable | (1) Coefficient | (2) Marginal Effect |
|---|---|---|
| Distance to 2$^{nd}$ nearest uni | $-0.0285^{***}$ | $-0.0039^{***}$ |
| | (0.0097) | (0.0011) |
| Distance to 7$^{th}$ nearest uni | 0.0097 | 0.0013 |
| | (0.0074) | (0.0010) |
| Students 7$^{th}$ nearest uni | $0.0121^{***}$ | $0.0016^{***}$ |
| | (0.0023) | (0.0003) |
| Distance to 9$^{th}$ nearest uni | $-0.0206^{**}$ | $-0.0032^{**}$ |
| | (0.0098) | (0.0013) |
| Distance to 10$^{th}$ nearest uni | $-0.023^{***}$ | $-0.003^{***}$ |
| | (0.005) | (0.001) |
| Migrational background | $-0.2080$ | $-0.0281$ |
| | (0.6330) | (0.0856) |
| No native speaker | $-0.0281$ | 0.0038 |
| | (1.1280) | (0.1520) |
| First born | $0.4500^{***}$ | $0.0609^{***}$ |
| | (0.1290) | (0.0173) |
| Siblings | $-0.2160^{***}$ | $-0.0292^{***}$ |
| | (0.0441) | (0.0059) |
| Married before college | 0.6390 | 0.0864 |
| | (0.704) | (0.0951) |
| Parent before college | 0.0824 | 0.0111 |
| | (0.9450) | (0.1280) |
| Final school grade: excellent | $1.645^{***}$ | $0.223^{***}$ |
| | (0.363) | (0.0485) |
| Final school grade: good | $1.0440^{***}$ | $0.1410^{***}$ |
| | (0.2310) | (0.0309) |
| Final school grade: satisfactory | $0.3200^{*}$ | $0.0433^{*}$ |
| | (0.2420) | (0.0328) |
| Final school grade: sufficient or worse | 0.221 | 0.0298 |
| | (0.5960) | (0.0806) |
| Grade repetition: 1 grade | $-0.1940$ | $-0.0263^{*}$ |
| | (0.1440) | (0.0195) |
| Grade repetition: 2+ grades | $-0.0256$ | $-0.00346$ |
| | (0.4290) | (0.0580) |
| Age 15: lived by single parent | 0.420 | 0.0567 |
| | (0.2640) | (0.0357) |
| Age 15: lived in patchwork family | $-1.0970^{**}$ | $-0.1480^{**}$ |
| | (0.4700) | (0.0633) |
| Age 15: orphan | $-1.2660^{**}$ | $-0.1710^{**}$ |
| | (0.6150) | (0.0830) |
| Age 15: mother unemployed | $-0.0461$ | $-0.00623$ |
| | (0.169) | (0.0235) |
| Age 15: mother never employed | $-0.1020$ | $-0.0138$ |
| | (0.1740) | (0.0235) |
| Age 15: father unemployed | 0.3590 | 0.0485 |
| | (0.2520) | (0.0241) |
| Age 15: father never employed | 0.1480 | 0.0200 |
| | (0.5670) | (0.0767) |
| M: migrational background | $0.421^{*}$ | 0.0570 |
| | (0.2430) | (0.0329) |
| M: further training | $0.8900^{***}$ | $0.1200^{***}$ |
| | (0.298) | (0.0402) |
| M: vocational training | $0.2970^{*}$ | $0.0402^{***}$ |
| | (0.1610) | (0.0217) |
| M: college education | $2.0880^{***}$ | $0.282^{***}$ |
| | (0.4760) | (0.0217) |
| F: migrational background | 0.270 | 0.0366 |
| | (0.2390) | (0.0323) |
| F: further training | $0.9620^{***}$ | $0.1300^{***}$ |
| | (0.2510) | (0.0336) |

*Continued on next page*

Table A2 – *continued*

| Variable | (1) Coefficient | (2) Marginal Effect |
|---|---|---|
| F: vocational training | −0.1100 | −0.0148 |
| | (0.2250) | (0.0305) |
| F: college education | 1.6240*** | 0.2200*** |
| | (0.2950) | (0.0392) |

*Notes:* Own calculations based on NEPS-Adult Starting Cohort data for wave 4 in 2011/12.

Table A3: Partial *F*-statistic for instruments

| | Instrument | | |
|---|---|---|---|
| | Lasso-chosen instruments | Distances to ten nearest colleges | College availability index |
| Partial first-stage *F*-statistic | 35.8 | 14.9 | 265.0 |
| Number of instruments | 5 | 10 | 1 |

*Notes:* Own calculations based on NEPS–Adult Starting Cohort data for wave 4 in 2011/12. This table gives the partial *F*-statistics of the instruments at the first stage of a 2SLS estimation. The sample is for working for pay as dependent variable at the second stage.

## B. Intuition behind Eq. (3) and (4)

What is the intuition behind the stated formulas in the separate estimation approach? We start by Eq. (3):

$$E(Y^1|U_D = p) = E(Y^1|U_D \leq p, D = 1) + p \times \frac{\partial E(Y^1|U_D \leq p, D = 1)}{\partial p},$$

Say, we consider an instrument that affects individuals at $p = 0.5$. Individuals with unobserved costs of college education below the median have already opted for college at a lower value of $p$. We can think of them as the "always-takers," who take college education irrespective of a marginal instrument change at $p = 0.5$. The $U_D$ of these always-takers at 0.5 is below 0.5, otherwise they would not have opted for college education according to the selection question. We observe $E(Y|P(\boldsymbol{Z}) = 0.5, D = 1)$, which is the average wage for the always-takers at $p = 0.5$. Now, if we vary the instrument slightly at $P(\boldsymbol{Z}) = 0.5$, new individuals, those at the margin of college education at $p = 0.5$, spill in. They are the compliers at 0.5. This may cause $E(Y|P(\boldsymbol{Z}) = 0.5, D = 1)$ to change, which is captured by the marginal derivative $\frac{\partial E(Y|P(\boldsymbol{Z})=0.5,D=1)}{\partial P(\boldsymbol{Z})}$. As we are interested in the level rather than the difference, we add the marginal derivative to the baseline wage $E(Y|P(\boldsymbol{Z}) = 0.5, D = 1)$. Note that the derivative in the second term is multiplied by the weight $p$. As we want to compare the (theoretically infinitesimal) small of group of individuals at the margin of taking college education with the bigger group of always-takers, we scale the derivative by $p$ to adjust to the marginal individuals. Intuitively, the higher $p$, the less sensitive $E(Y|P(\boldsymbol{Z}) = p, D = 1)$ is to the inflow of new individuals.

Similarly, we can infer from $E(Y|P(\boldsymbol{Z}) = p, D = 0)$ and the selection equation that only individuals with larger values of $U_D$ than $p$ will not opt for college education. To get the mean wage of indifferent non-graduates with a rank equal to $p$, we need to adjust $E(Y|P(\boldsymbol{Z}) = p, D = 0)$ as follows:

$$E(Y^0|U_D = p) = E(Y^0|U_D > p, D = 0) - (1 - p) \times \frac{\partial E(Y^0|U_D > p, D = 0)}{\partial p}.$$

The difference to (3) is twofold. First, the derivative is deducted, not added. This is because, as the instrument is varied, complying women opt for college education and leave the untreated group. Second, the weight is adjusted to the $(1 - p)$ share of women whose rank in $U_D$ is higher than $p$. We estimate $E(Y^1|U_D \leq p, D = 1)$ and $E(Y^0|U_D > p, D = 0)$ by estimating local polynomial regressions for college and non-college-educated women separately. We evaluate these non-parametric regressions at up to 100 different values of $p$ along the support of $P(\boldsymbol{Z}$.

## C. Allowing for small shares of type 3 individuals

We now repeat the derivation of Eq. (7) to (9) from Section 5.1 without imposing the assumption that $p_D$=0 we get the following equations. Eq. (7) becomes:

$$\begin{aligned}
E(Y^1 \mid L^1 = 1) &= E(Y^1 \mid L^1 = 1, L^0 = 1)\frac{p_A}{p^1} \\
&\quad + E(Y^1 \mid L^1 = 1, L^0 = 0)\frac{p_C}{p^1} \\
\Leftrightarrow E(Y^1 \mid L^1 = 1, L^0 = 1) &= E(Y^1 \mid L^1 = 1)\frac{p^1}{p_A} \\
&\quad - E(Y^1 \mid L^1 = 1, L^0 = 0)\frac{p_C}{p_A} \qquad (7')
\end{aligned}$$

Eq. (8) becomes:

$$\begin{aligned}
E(Y^0 \mid L^0 = 1) &= E(Y^0 \mid L^1 = 1, L^0 = 1)\frac{p_A}{p^0} \\
&\quad + E(Y^0 \mid L^0 = 1, L^1 = 0)\frac{p_D}{p^0} \\
\Leftrightarrow E(Y^0 \mid L^0 = 1, L^1 = 1) &= E(Y^0 \mid L^0 = 1)\frac{p^0}{p_A} \\
&\quad - E(Y^0 \mid L^0 = 1, L^1 = 0)\frac{p_D}{p^A} \qquad (8')
\end{aligned}$$

Eq. (9) becomes:

$$\begin{aligned}
E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1) &= E(Y^1 \mid L^1 = 1)\frac{p^1}{p_A} - E(Y^0 \mid L^0 = 1)\frac{p^0}{p_A} \qquad (9') \\
&\quad - \left[ \underbrace{E(Y^1 \mid L^1 = 1, L^0 = 0)}_{\text{Type 2}} \frac{p_C}{p_A} - \underbrace{E(Y^0 \mid L^0 = 1, L^1 = 0)}_{\text{Type 3}} \frac{p_D}{p_A} \right]
\end{aligned}$$

Thus, if $p_D > 0$, we end up with two unobserved numbers (one for type 2, one for type 3) when we want to determine $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$. To gain a better understanding of how sensitive our bounds are with respect to the existence of a small percentage of type 3 individuals, we consider some back-of-the-envelope calculations in what follows. Note, that these calculations are not carried out for different values of $U_D$ as in the MTE estimations. Instead, to keep things simple and illustrative, we calculate a point "estimate" of $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$.

We calibrate Eq. (9′) such that it produces results similar to the ATT results without type 3 women in Table 5. The numbers are reported in Table C1 (note that wages $Y$ are in logs).

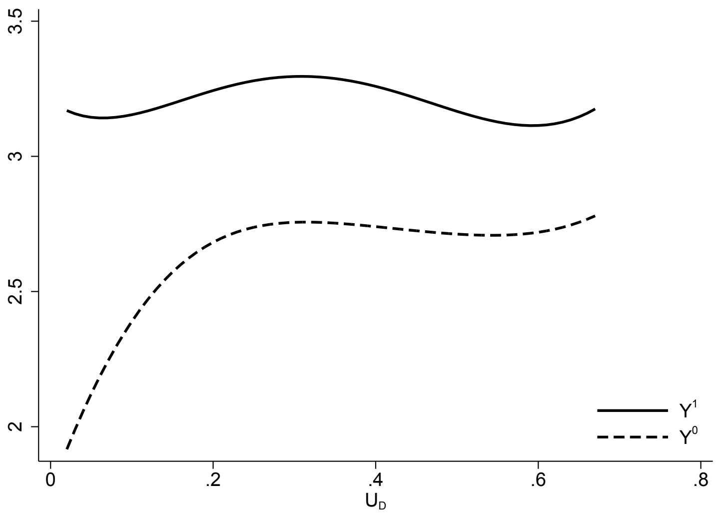
Why these numbers? Figure C1 reports estimates of the potential outcomes for wages using the separate estimation approach, following Carneiro and Lee (2009). The upper curve reports $E(Y^1|L^1 = 1, U_D)$, the lower curve $E(Y^0|L^0 = 1, U_D)$. The difference between the two curves yields the NMTE estimate for wages in Figure 5. We use the broad averages of 3.1 and 2.5 for the numbers in rows 1 and 2 in Table C1. The difference broadly corresponds to the ATT in Table 5.

Table C1: Parameter settings for the RHS of Eq. (9′)

| | Parameter | Value | Source |
|---|---|---|---|
| 1 | $E(Y^1 \mid L^1 = 1)$ | 3.1 | Broad average of $Y^1$ over $U_D$ in Figure C1 |
| 2 | $E(Y^0 \mid L^0 = 1)$ | 2.5 | Broad average of $Y^0$ over $U_D$ in Figure C1 |
| 3 | $p^1$ | 0.8 | Table 5 |
| 4 | $p^0$ | 0.6 | Table 5 |
| 5 | Lower bound $E(Y^0 \mid L^1 = 0, L^0 = 1)$ | 1.6 | Min of $Y^0 \times 0.85$ in Figure C1 |
| 6 | Upper bound $E(Y^0 \mid L^1 = 0, L^0 = 1)$ | 2.6 | Max of $Y^0 \times 0.93$ in Figure C1 |
| 7 | Lower bound $E(Y^1 \mid L^1 = 1, L^0 = 0)$ | 2.8 | Min of $Y^1 \times 0.90$ in Figure C1 |
| 8 | Upper bound $E(Y^1 \mid L^1 = 1, L^0 = 0)$ | 3.8 | Max of $Y^1 \times 1.15$ in Figure C1 |

*Notes:* Own illustration. Values from Figure C1 and Table 5 are approximations. Parameters 5 to 8 are weighted to mimic the benchmark results in Table 5. Table 5 does not directly state $p^1$ and $p^0$ but information on $p^0/p^1$ and $p^1 - p^0$ allows calculating $p^1$ and $p^0$.

Figure C1: Potential outcomes for wages




*Notes:* Own illustration using NEPS–Adult Starting Cohort data. This graph depicts the levels of the potential outcomes for wages, $Y^1$ and $Y^0$. Estimates according to Carneiro and Lee (2009).

$p^0$ and $p^1$ in rows 3 and 4 Table in C1 are set such that they both lead to a difference of 0.2 (broadly the ATT on working in Table 5) and to a share of type 1 individuals of 0.75 (see panel B, Table 5).

For the unobserved wage of type 2 and type 3 individuals (rows 5 to 8 Table in C1) we use two extreme cases to calculate bounds: the minimum and the maximum of the potential outcomes in Figure C1. We weight the minimum and maximum values such that they produce bounds for the benchmark case of no type 3 women which are similar to those in Table 5. As there is no theoretical justification for choosing exactly these numbers and because we keep we same weights when allowing for small shares of type 3 women, this calibration does not affect the interpretation of our results. Thus, we get four different combinations of extreme values for type 2 and type 3 women.

Now, we present three versions: We allow $p_D$ to be zero percent (baseline case, no type 3), one percent, and two percent. Then, the resulting employment rates for the types are:

- $p_D = 0.00 \rightarrow p_A = 0.80$ and $p_C = 0.20$ (thus, $p^1 = p_A + p_C = 0.80$ and $p^0 = p_A + p_D = 0.60$),
- $p_D = 0.01 \rightarrow p_A = 0.59$ and $p_C = 0.21$, and
- $p_D = 0.02 \rightarrow p_A = 0.58$ and $p_C = 0.22$.

Table C2 calculates $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$ by inserting all possible combinations of type 2 and type 3 extreme wages as well as the three employment rates considered for type 3 in (9′). While the columns state the different employment rates, the rows of the table vary in the extreme values used.

Table C2: Sensitivity of the bounds with respect to different type 3 employment rates

| | Extreme values for | | Wage effect $E(Y^1 - Y^0 \mid L^1 = 1, L^0 = 1)$ for | | |
|---|---|---|---|---|---|
| | Type 2 $E(Y^1 \vert L^1 = 1, L^0 = 0)$ | Type 3 $E(Y^0 \vert L^1 = 0, L^0 = 1)$ | Baseline $p_D = 0.00$ | Version 2 $p_D = 0.01$ | Version 3 $p_D = 0.02$ |
| Combination 1 | 3.8 | 2.6 | 0.37 | 0.35 | 0.34 |
| Combination 2 | 3.8 | 1.6 | 0.37 | 0.34 | 0.30 |
| Combination 3 | 2.8 | 2.6 | 0.70 | 0.71 | 0.72 |
| Combination 4 | 2.8 | 1.6 | 0.70 | 0.69 | 0.68 |

*Notes:* Own calculations. The wage effects are based on Eq. (9′).

The resulting bounds for type 1 wages are

- Baseline, $p_D = 0.00$: 0.37–0.70,
- Version 2, $p_D = 0.01$: 0.34–0.71, and
- Version 3, $p_D = 0.02$: 0.30–0.72.

This is a simplistic model. Its goal is not to produce numbers that can be taken at face values. Instead its goal is to show that slight deviances of our assumption of no type 3 women have only minor consequences. Of course, this also holds for 3, 4, or 5 percent shares of type 3 women.

## D. Technical remark on the estimation of the upper bound

The estimation of the censored mean of the upper bound is similar to the respective estimation for the lower bound. But because of the sake of computational time, we use the same definitions of $Y_k$ and $I_k$ (instead of defining new indicators using $\geq$ in the indicator function to acknowledge that the censored mean is larger than $U$ for the upper bound) and apply the following formula:

$$\begin{aligned}
E_{UB}(Y^1 \mid L^1 = 1, L^0 = 1) &= \frac{\int_U^\infty y f_{Y^1}(y)dy}{1 - F_{Y^1}(U)} = \frac{\int_{-\infty}^\infty y f_{Y^1}(y)dy - \int_{-\infty}^U y f_{Y^1}(y)dy}{1 - F_{Y^1}(U)} \\
&= \underbrace{\int_{-\infty}^\infty y f_{Y^1}(y)dy}_{E(Y^1|X,U_D)} \underbrace{\frac{1}{1 - F_{Y^1}(U)}}_{A} - \frac{\overbrace{\int_{-\infty}^U y f_{Y^1}(y)dy}^{B}}{\underbrace{1 - F_{Y^1}(U)}_{A}}
\end{aligned}$$

Where $E(Y^1|X, U_D)$ simply is the unadjusted marginal college wage and $A$ and $B$ are defined and estimated as:

**A:** (referring only to the term in the denominator) $\frac{p_1 - p_0}{p_1}$, which we estimate using the marginal potential employment probabilities by college state.

**B:** $E(Y_k|X, U_D)$, where $Y_k = Y \cdot \mathbb{1}(Y \leq y_k)$

The question remains how we find the correct threshold $y_k$. We estimate $E(I_k|X, U_D)$, where $I_k = Y \leq y_k$ for many different values of $y_k$ and take the one for which $E(I_k|X, U_D, D = 1) = \frac{p_1 - p_0}{p_1}$. We then take the respective $y_k$ and estimate the corresponding expression $E(Y_k|X, U_D, D = 1)$.