

Fitzenberger, Bernd; de Lazzer, Jakob

**Article — Published Version**

## Changing selection into full-time work and its effect on wage inequality in Germany

Empirical Economics

**Provided in Cooperation with:**

Springer Nature

*Suggested Citation:* Fitzenberger, Bernd; de Lazzer, Jakob (2021) : Changing selection into full-time work and its effect on wage inequality in Germany, Empirical Economics, ISSN 1435-8921, Springer, Berlin, Heidelberg, Vol. 62, Iss. 1, pp. 247-277, <https://doi.org/10.1007/s00181-021-02098-0>

This Version is available at:

<https://hdl.handle.net/10419/286846>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



<https://creativecommons.org/licenses/by/4.0/>



# Changing selection into full-time work and its effect on wage inequality in Germany

Bernd Fitzenberger<sup>1,3</sup> · Jakob de Lazzer<sup>2</sup>

Received: 31 July 2019 / Accepted: 15 July 2021 / Published online: 7 December 2021  
© The Author(s) 2021

## Abstract

To explore whether changes in the selection into full-time work among German men were a driver in the rise in wage inequality since the mid-1990s, we propose a modification of selection-corrected quantile regressions. Addressing Huber and Melly's (J Appl Econom 30(7):1144–1168, 2015) concerns, this modification allows us to estimate the effects of selection with respect to both observables and unobservables. Our findings show that those employed in 1995 would have had lower wages in 2010 than those employed in 2010 and wage dispersion would have been higher, suggesting that full-time workers have become less heterogeneous over time.

**Keywords** Wage inequality · Full-time employment · Selection into employment

**JEL Classification** J31 · J20 · J60 · J21

---

We thank Martin Biewen and Matthias Seckler for very helpful discussions as well as Blaise Melly and an anonymous referee for very helpful comments in the review process. We are also grateful for helpful discussions and suggestions at the Conference on “Economic Applications of Quantile Regression 2.0” 2019, the IRTG Summer Camp 2018, and various presentations in the Priority Programme of the German Science Foundation on “The German Labor Market in a Globalized World—Challenges through Trade, Technology, and Demographics” (SPP 1764). We thank the Research Data Center at IAB for excellent support with the data access through the CADAL project, which is part of SPP 1764. We acknowledge financial support of this project by the German Science Foundation (DFG) through the project “Accounting for Selection Effects in the Analysis of Wage Inequality in Germany” (Project Nos.: BI 767/3-1 and FI 692/16-1). The responsibility for all errors is, of course, ours.

---

✉ Bernd Fitzenberger  
bernd.fitzenberger@iab.de

<sup>1</sup> IAB and FAU Erlangen-Nürnberg, IFS, CESifo, IZA, ROA, Berlin, Germany

<sup>2</sup> Humboldt University Berlin, Berlin, Germany

<sup>3</sup> IAB (Institute for Employment Research), Regensburger Straße 100, 90478 Nuremberg, Germany

## 1 Introduction

Germany experienced a considerable increase in wage inequality until 2010 (Dustmann et al. 2009; Card et al. 2013; Möller 2016; Biewen et al. 2018). For an assessment of what factors drive the observed changes in the wage distribution and the wage differences between labor market groups, it is necessary to take into account that the selection into paid work may change over time and that it may differ across groups. Selection may work through the changing composition of the workforce with respect to easily observable characteristics, such as educational qualifications, work experience, or age. It may also work through selection based on unobserved factors like motivation, social skills, or the ability to adapt to changing circumstances.

This paper estimates selection-corrected quantile regressions to address two research questions regarding wage inequality among German men in 1995 and 2010. First, we consider the shape of the wage distribution and the magnitude of inequality in wages which would have prevailed if all unemployed had been working full-time. Because full-time employment is selective and likely based on earnings prospects, we would expect wage inequality to be higher if both the unemployed and the employed were working full-time. Our second question addresses the changes over time: How would wage inequality have developed if selection into full-time employment had not changed over time?

If the distribution of observed and unobserved characteristics affecting wages were the same among unemployed and employed, we would not have to correct for selection. However, full-time workers are likely to differ considerably from unemployed workers. A common approach is to apply sample selection corrections for mean regressions based on Heckman (1979), an approach which cannot be easily generalized to the analysis of the entire distribution even under the normality assumption. There exists a small but growing literature on how to account for unobservables in the analysis of wage distributions. For instance, Card et al. (2013) estimate worker and firm fixed effects accounting for unobservable persistent differences between workers and between firms. However, the study does not account for the selection into employment due to unobservables.

A limited number of approaches have been suggested to correct entire distributions for selection due to unobservables. Most applications of selection-corrected quantile regressions so far employ a control function approach, as in Buchinsky (2001; 1998), Albrecht et al. (2009), Bollinger et al. (2011) and Picchio and Mussida (2011), which we also apply for wage regressions based on German administrative data (see also Das et al. 2003 for semi-parametric selection models). Huber and Melly (2015) point out that this selection correction approach is only valid if the error terms in the selection equation and the wage equation are independent conditional on the selection probability. This conditional independence assumption implies equal slope coefficients for the determinants of wages in the selection-corrected quantile regressions of wages.

As our methodological contribution, we propose to respecify the estimated selection-corrected quantile regressions by transforming the dependent variable with the goal that equality of the slope coefficient then holds. The transformation is estimated based on the identification-at-infinity assumption which is plausible in our application. As a modification of the two-step approach by Buchinsky (1998), our

approach includes an additional step to address the concern raised by Huber and Melly. A version of their test of equality of slope coefficients is used to guide the choice of the transformation, this way ensuring conditional independence in our application. With the control function approach augmented by a transformation of the dependent variable, we estimate quantile regressions which are corrected for selective movements between unemployment and full-time work. Undoing the transformation based on the selection-corrected quantile coefficients and employing the decomposition technique of Melly (2006) and Chernozhukov et al. (2013), we then estimate counterfactual wage distributions.

Our approach transforming the dependent variable is similar to the approach suggested in the companion paper Biewen et al. (2020) [henceforth, BFS], which estimates the selection bias in employment for the estimation of the gender wage gap. However, the actual implementation of the transformation approach and the specific application differs between the two papers. BFS investigate a small-scale application, and the transformation is applied to both the dependent variable and the covariates. Our application involves a much larger dataset, and the selection probabilities are much higher. For this reason, we will use of the identification-at-infinity approach to obtain selection-corrected coefficient estimates, and we then use these to estimate the selection correction for the transformed model. Furthermore, our transformation involves only the dependent variable but not the covariates. As described later in this paper, BFS make some assumptions on the link between the model specification and the transformation factor, which we do not make here. Due to the large sample size, small substantive (economic) deviations from the conditional independence property for the transformed model lead to a rejection of our implementation of the Huber and Melly test, which means that the challenge to find an appropriate transformation is much higher in the application here compared to BFS. It turns out that, in fact, we were unable to find an appropriate transformation for the full sample that passes the Huber/Melly test. This is an informative finding by itself, and we then successfully apply the approach separately for two subsamples with different selection mechanisms. Regarding the substantive economic research question, BFS analyze the gender wage gap accounting for selection into employment among females while this paper investigates the role of selection into employment in explaining the increase in wage inequality among males. In sum, this paper and BFS involve independent contributions both regarding the implementation of the transformation estimator and different substantive applied research questions.

In a recent important paper, Arellano and Bonhomme (2017) suggest a copula based method to provide consistent estimates of quantile regressions with selection correction. They estimate quantile regressions while assuming a fixed copula between the conditional rank in the wage distribution and the rank in the error term of the selection equation. The approach amounts to estimating rotated quantile regressions, which relate the  $\tau$ th quantile regression in the nonselected sample to a rotated value, which represents the rank of the  $\tau$ th unselected quantile in the selected sample, thus linking the two for estimation purposes. This is an alternative to Buchinsky's selection corrections approach which estimates the difference between the  $\tau$ th quantiles in the two samples. The approach of Arellano and Bonhomme (2017) has two disadvantages. First, the authors estimate the copula while assuming a specific functional form, and

they allow only for the covariates to have a limited impact on the joint distribution of ranks. Second, the estimation of the copula is computationally very involved. We view the two approaches to model the relationship between the ranks or the quantiles, respectively, in the unselected and selected sample as complementary, both having to address the dependence of this relationship upon covariates.

As an alternative to a quantile regression approach, Chernozhukov et al. (2019) and Fernandez-Val et al. (2019) discuss the estimation of selection-corrected distribution regressions based on a local parametric approximation. Distribution regressions can be inverted to estimate the distribution of the dependent variables under counterfactual selection rules. These counterfactuals are then used to decompose the changes in wage inequality over time. Similar to the assumed copula in Arellano and Bonhomme (2017), these studies assume a specific functional form regarding the link between the selection equation and the distribution regression. This link (modeling the so-called selection sorting effect) is allowed to vary as a function of a linear index of the covariates, and the coefficients in the index are estimated. D'Haultfoeuille et al. (2014) suggest an approach in which identification relies on the independence between covariates and selection for large values of the outcome, and on the homogeneity of the estimand across the distribution. This assumption does not seem plausible in our application.

Regarding our first research question, the unemployed prove to be a negative selection of the workforce in times of low unemployment as in the years 1995 and 2010. The counterfactual wage quantiles, if everyone were working full-time, would be lower than the observed ones and wage inequality would be considerably higher. Concerning our second research question, those employed in 1995 would have had lower wages in 2010 than those employed in 2010 and wage dispersion would have been higher for them. Overall, this implies that full-time workers have become less heterogeneous with regard to the factors driving wages as well as the selection into full-time work.

The remainder of this paper is organized as follows: Sect. 2 describes the data used and provide descriptive evidence of trends in wages and unemployment, as well as the instrumental variables used for the control function approach. Section 3 outlines in detail our econometric approach for estimating selection-corrected quantile regressions and calculating counterfactual wage distributions. We apply this approach to our data and discuss the results in Sect. 4. Section 5 concludes.

## 2 Data and descriptive evidence

Our analysis uses the SIAB, the factually anonymous Sample of Integrated Labour Market Biographies (version 1975–2010, henceforth denoted by SIAB710).<sup>1</sup> This is an administrative dataset based on German social security records. It contains a 2% sample of all dependent employees who are subject to social security, all individuals receiving unemployment benefits, but no self-employed or civil servants. We restrict the analysis to those aged 25–55 working in West Germany. Wages are available as daily wages in Euros, which we deflate to the level of 1990. Since these wages

<sup>1</sup> We used the Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB), see Vom Berge et al. (2013) for the data documentation.

are collected from administrative data sources, the measurements are very precise and there are no problems of selective nonresponse or measurement error, which wages reported in survey data sets may suffer from. Following of literature, we restrict attention to full-time employees, because our dataset does not contain information on hours worked (Dustmann et al. 2009; Card et al. 2013; Möller 2016).

As a consequence of the unavailability of comparable wages for part-time workers and because we have no reliable information on individuals being out of labor force, we perform our analysis only on the data for males. For males, the majority of selective movements during working age occur between unemployment and full-time employment. However, this is not the case for females. For them, part-time employment and absence from the labor force affect large shares of the working age population, so an analysis which restricts attention to the selection between unemployment and full-time employment is not well suited to studying the effects of selection on female wages.

We analyze wages for the years 1995 and 2010. These years represent the start and end of the strong rise in lower-tail wage inequality for German workers, as well as the turning point in the development of unemployment (Biewen et al. 2018; Möller 2016). Table 1 involves descriptive statistics on the samples used for our analysis.

Levels of education are aggregated into three categories based on highest degrees obtained: (i) high-educated: college (university/university of the applied sciences), (ii) medium-educated: high school and/or vocational training, and (iii) low-educated: no/other degree. These are the standard education categories used in the literature on wage inequality for Germany based on the SIAB (see, e.g., Dustmann et al. 2009, 2014; Biewen et al. 2018).

We capture an individual's labor market history by the number of days spent in full-time employment and part-time employment, respectively, aggregated over the last 5 years. Episodes of part-time and non-employment are important determinants of individual wage development (Paul 2016) and of changes in wage inequality in general (Biewen et al. 2018). All wages above the contribution threshold for social security, which lies between the 85th and 90th percentile, are censored in the sample. For the analysis of wage quantiles above the threshold, we impute wages, similar to the method of Gartner (2005). The imputed wages are based on the fitted values of a Tobit model for censored data and take into account the heteroscedastic variance of the Tobit model.<sup>2</sup> However, because of the severe censoring for the high-educated, we restrict our analysis to the medium and low-educated.

## 2.1 Wage inequality

From the early 1990s onward, wage inequality increased substantially, as measured for instance by the gap between the top quartile and the bottom quartile of the distribution of gross wages. The top left panel of Fig. 1 shows that, relative to their levels in 1995, male workers near the bottom of the wage distribution suffered a decline in real wages,

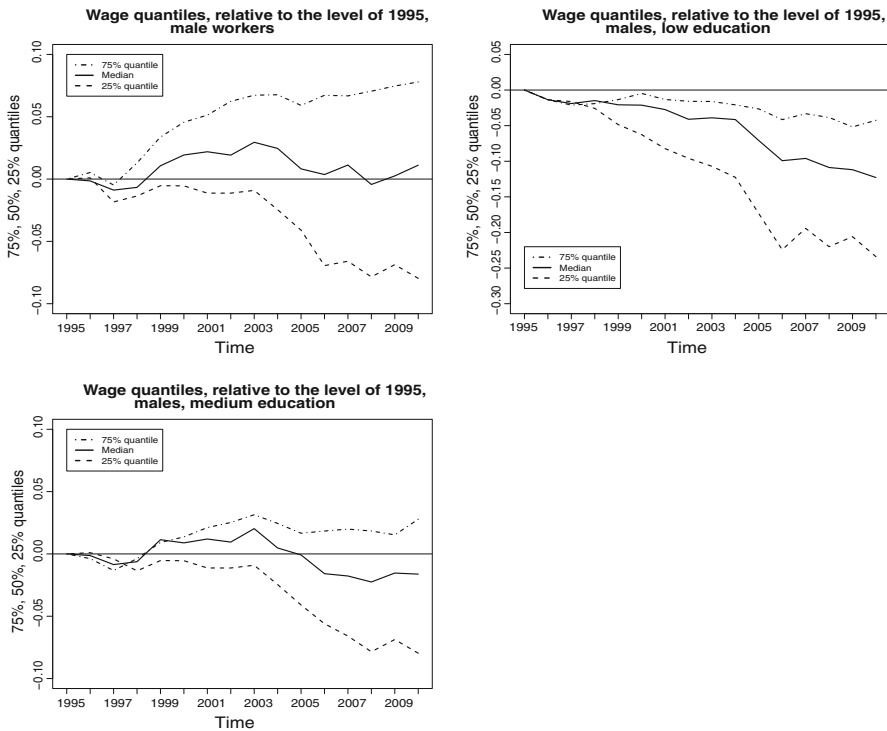
<sup>2</sup> To account for censoring, one may consider estimating censored quantile regression. Since estimating censored quantile regression involves a major computational problem and since predicting top conditional quantiles may show unsatisfactory results (see the Monte Carlo evidence in Fitzenberger and Winker 2007), we refrain from doing so.

**Table 1** Sample descriptives

	1995 Employed	1995 Unemployed	2010 Employed	2010 Unemployed
Age	38.92	38.98	41.34	39.98
Low-educated	.1325	.2738	.1196	.3609
Medium-educated	.7564	.6738	.716	.5993
High education	.1111	.0524	.1645	.0398
Days in PT last 5 years	5.703	16.61	17.18	37.5
Days in FT last 5 years	1580	885.4	1597	518.3
N	169,243	27,833	145,196	29,846
<i>Low-educated</i>				
Age	39.37	39.64	39.16	33.39
Days in PT last 5 years	4.695	32.36	29.98	10.94
Days in FT last 5 years	1444	466.3	1346	143
N	25,148	14,121	20,766	11,256
<i>Medium-educated</i>				
Age	38.59	43.37	41.43	38.06
Days in PT last 5 years	4.069	0	14.53	34.13
Days in FT last 5 years	1587	1536	1608	514.8
N	144,095	13,712	124,430	18,690

Notes: Averages of explanatory variables by subgroup. 'Days in PT/FT last 5 years' measure the number of days in part-time/full-time employment during the last 5 years.

Source: SIAB7510, own calculations



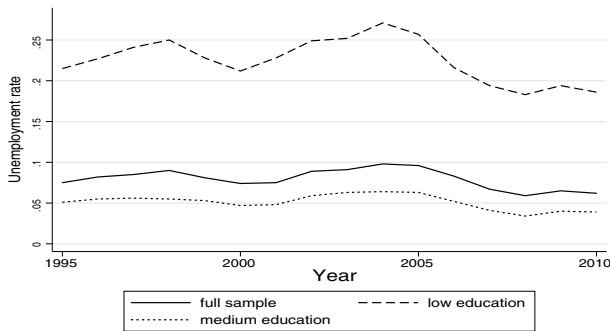
**Fig. 1** Change in wage inequality over time. *Notes:* The graphs show the changes of the cross-sectional quantiles of real wages in logs over time relative to 1995. Source SIAB7510, own calculations

while those near the top experienced an increase. The median wage basically stagnates over the entire period from 1995 to 2010. A part of the increase in inequality can be attributed to an aging population and increased shares of highly educated workers (Dustmann et al. 2009; Biewen et al. 2018). Among policy makers, the observed increase in inequality is often viewed as a negative development, because it reflects falling earnings for low-wage workers. This has caused great concern, which has contributed to the introduction of a statutory minimum wage for Germany in 2015 (Caliendo et al. 2019). Even within education groups, the wage distributions have widened since the mid-1990s. As shown in panels 2 to 3 of Fig. 1, wage inequality increased strongly both for the low-educated and the medium-educated. For the low-educated, real wages fell even up to the top of the wage distribution, even above the upper quartile, and the decline of the median real wage between 1995 and 2010 amounts to about 10 log points.

## 2.2 Unemployment

Our analysis focuses on the unemployed receiving unemployment benefits. The benefit entitlement period amounts typically to at most 12 months for those individuals who





**Fig. 2** Unemployment rate. *Notes:* Source IAB Labor market report 10/2017. Unemployment rate among male workers in West Germany

previously had a spell of dependent employment.<sup>3</sup> The registered unemployment rate for German men changed substantially between 1995 and 2010. Starting from 7.5% in 1995, it reaches its peak of 9.8% in 2004. After 2004, there is first a strong decline and then a slight increase in the aftermath of the financial crisis. The development for the medium-educated is almost parallel to the aggregate unemployment rate. The unemployment rate of low-educated is generally higher, especially before 2005, but declines afterward even more strongly than that of the medium-educated. The strong drop in unemployment between 2004 and 2010 coincides with the rapid increase in wage inequality documented above (Fig. 2).

A common interpretation is that the fall in unemployment could be associated with a stronger inflow of previously unemployed into full-time work (see, e.g., Dustmann et al. 2014). Those previously unemployed individuals might, on average, possess observable and unobservable characteristics which are less highly valued in the labor market than those of the already employed workers. Therefore, the resulting labor force may be more heterogeneous with regard to the drivers of wages (Biewen et al. 2018). Because work incentives for low-wage workers have been strengthened by various labor market reforms in the early 2000s, for instance, through cuts of unemployment benefits, this effect could be particularly strong in the lower tail of the wage distribution, contributing to the decline of the quantiles below the median.

However, it is an open question whether a decline of unemployment benefits necessarily implies a widening of the lower tail of the wage distribution. We would like to mention three possible counter arguments without being able to provide a comprehensive discussion. First, labor market frictions might prevent wages of newly employed to differ substantially from those of the already employed. Second, the cuts in unemployment benefits may also have reduced the bargaining power of the incum-

<sup>3</sup> Unemployment benefits are paid for a longer time period above certain age limits, which applies mostly to workers above age 55. Long-term unemployed are covered by other types of welfare which have undergone multiple reforms over the observation timeframe and are not consistently observed in the dataset. Additionally, not all of those receiving welfare benefits are available for employment (e.g., due to illness or early retirement with pensions below welfare levels). We therefore refrain from including the long-term unemployed in our analysis, as they are not well suited for analyzing counterfactual wages if employment was not selective with respect to worker characteristics.

bent workforce. Third, rising rates of retirement, a falling supply of younger workers, and higher wage flexibility among younger workers may reduce unemployment but not widen the wage distribution.

### 2.3 Instruments for selection

Semiparametric identification of selection effects in quantile regressions of wages requires at least one instrument satisfying an exclusion restriction (compare Buchinsky 1998), analogous to a Heckman sample selection model for mean regression. The instruments need to provide exogenous variation in the selection probability into employment without affecting wages. Since the SIAB7510 data do not contain individual level variables, which we think are suitable as instruments, we use instead four additional variables merged to the SIAB7510 at the regional district level (Kreisebene). These variables are cohort sizes of young adults aged 18–24 and 25–30 as well as graduation rates in lower secondary and higher secondary education. These instruments reflect exogenous shocks to the labor supply in the respective region and year, affecting individual employment chances. We believe the exclusion restrictions to be credible, because it is unlikely that wages respond in the short run to labor supply differences between regional districts. Wage rigidities prevent short-term adjustment in response to labor supply variations due to new entrants into the labor market (compare Bauer et al. 2007). This is partly because wage contracts generally span multiple years and wages of new employees are not independent of wages for current employees, after accounting for individual differences in employment history. Additionally, collective bargaining in Germany work at the level of the industry or large firms and therefore does not allow for a wage response to shocks at the district level. District level data on the instruments are obtained from the Federal Statistics Office's regional database.<sup>4</sup> Our analysis will rely on an identification-at-infinity assumption, meaning that the support of the instrument includes with positive probability cases, for which the selection probability is close to one (Heckman 1990).

## 3 Methodological approach

### 3.1 Model setup

The setup follows Huber and Melly (2015). The wage equation for all individuals (employed or unemployed) is

$$Y^* = X\beta + v, \quad (1)$$

<sup>4</sup> Data source: Regionaldatenbank des Statistischen Bundesamtes. If we include individuals aged 20–60, the strength of the instruments increases. This means that individuals in their early 20s and late 50s are more strongly affected by labor supply shocks of young workers entering the labor force. However, we restrict our empirical analysis to those 25–55 years old because different nonemployment states cannot be distinguished well in our data. Many individuals aged between 20 and 24 are still in education and individuals in their late 50s start leaving the labor force through early retirement.

where  $Y^*$  denotes the latent log wage in the absence of selection,  $X$  the vector of observable covariates, being determinants of wages,  $v$  the error term, and  $\beta$  the vector of coefficients. We assume that  $\beta_{0.5} = \beta$ , i.e.,  $\beta$  represents the median coefficients and  $v$  represents the residual of a median regression. Assuming a linear quantile regression, the conditional  $\tau$ -quantile of the latent wage  $Q_\tau(Y^*|X)$  is specified by

$$Q_\tau(Y^*|X) = X\beta + Q_\tau(v|X) = X\beta_\tau, \quad (2)$$

which also means that  $Q_\tau(v|X) = X(\beta_\tau - \beta)$  is a linear function of  $X$ . Correspondingly, the  $\tau$ th quantile regression of  $Y^*$  is  $X\beta_\tau + v_\tau$ , with  $v_\tau = v - Q_\tau(v|X) = v - X(\beta_\tau - \beta)$ .

The selection problem arises because we only observe wages for employed individuals. Let  $Y$  denote the observed wage and  $D$  the selection indicator. We specify

$$D = 1(Z\gamma + \varepsilon \geq 0),$$

where  $Z$  is a strict superset of  $X$ , thus also including instruments for selection, which are excluded in Eq. (1), and  $\varepsilon$  is assumed to be independent of  $Z$ . The probability of selection

$$Pr(D = 1|Z) = Pr(Z\gamma + \varepsilon > 0|Z) \quad (3)$$

is a function of  $Z\gamma$ . For the selective sample, the observation rule is  $Y = Y^*$  ( $Y^*$  observed) only if  $D = 1$ . A conditional quantile in the selected sample is

$$Q_\tau(Y|Z) = X\beta_\tau + Q_\tau(v_\tau|Z, D = 1). \quad (4)$$

The term  $Q_\tau(v_\tau|Z, D = 1)$  denotes the quantile- $\tau$ -specific selection bias, with  $Q_\tau(v_\tau|Z, D = 1) > (<)0$  representing positive (negative) selection. The selection bias can be rewritten as

$$Q_\tau(Y|Z) = Q_\tau(Y^*|Z, D = 1) = X\beta_\tau + \tilde{g}(X, Z\gamma) \quad (5)$$

where  $Q_\tau(v_\tau|Z, D = 1) = \tilde{g}(X, Z\gamma)$  because  $v_\tau$  depends on  $X$  and  $D = 1$  on  $Z\gamma$ .

The control function  $\tilde{g}(X, Z\gamma)$ , which properly accounts for selection bias, should be a flexible function of  $X$  and  $Z\gamma$ , which is challenging because of the curse-of-dimensionality regarding  $X$  being multivariate. Nonparametric identification requires both independent variation in  $Z\gamma$  given  $X$  and identification at infinity. Identification at infinity means that with positive probability, based on the distribution of  $Z\gamma$ , the selection probabilities  $Pr(D = 1|Z)$  is close to one (Das et al. 2003). The selection model above implies that  $Q_\tau(v_\tau|Z, D = 1)$  converges to zero (no selection), if the employment probability  $P(D = 1|Z)$  converges to one, which is equivalent to  $Z\gamma$  going to infinity.

Extending upon Heckman (1990) and Andrews and Schafgans (1998), who consider the case where  $u$  is independent of  $X$ , both the intercept and the slope coefficients  $\beta$  can be identified, if we have observations with a selection probability close to one for each

value of  $X$ . Given the linear specification of  $X\beta_\tau$ , a smaller subspace of the support  $A$  of  $X$  suffices, where  $E[(X'X) \cdot I(X \in A)]$  can be inverted [ $I(\cdot)$  denotes the indicator function] and where the selection probability is close to one with positive probability. In our application, the selection probability is quite large for most observations and the subset of observations with a selection probability close to one (to anticipate: the median (upper quartile) of the selection probabilities lies above 93% (96%) in all four subsamples considered, see Table 3), is sufficiently large to estimate  $\beta_\tau$  consistently. In our application, we will use the coefficient estimates based on the identification-at-infinity sample to characterize the selection bias in the full sample.<sup>5</sup>

### 3.2 Buchinsky's approach

The selection correction approach proposed by Buchinsky (1998; 2001) applies a standard Heckman selection approach with instruments (Heckman 1979; Vella 1998) to quantile regression. Buchinsky specifies the selection correction term in the second stage [Eq. (3)] as a function of the inverse Mills ratio  $\lambda(Z\hat{\gamma})$ . However, even under joint normality of  $\varepsilon$  and  $v$ , the selection correction term  $Q_\tau(v_\tau|Z, D = 1)$  is generally not a linear function in  $\lambda$ . Thus, Buchinsky suggests to approximate the selection correction term  $Q_\tau(v_\tau|Z, D = 1)$  by a power series (polynomial) of  $\lambda$  (see Vella 1998 on semiparametric approaches for selection correction in mean regressions). Further, Buchinsky assumes that the joint distribution of  $v$  and  $\varepsilon$  is independent of  $Z$ , conditional on the probability of selection  $Pr(Z\gamma + \varepsilon > 0|Z)$  (Huber and Melly 2015).

In the second step, the selection-corrected quantile regression

$$Q_\tau(Y|X) = X\beta_\tau + \theta_\tau g(\lambda) \quad (6)$$

is estimated for the selective sample with  $D = 1$ . Equation (6) presumes that  $\theta_\tau g(\lambda)$  represents  $Q_\tau(v_\tau|Z, D = 1)$ .  $g(\cdot)$  is a power series of  $\lambda$ , and thus  $\theta_\tau g(\lambda)$  approximates the selection correction term  $Q_\tau(v_\tau|Z, D = 1)$ .

Without the assumption that the joint distribution of  $v$  and  $\varepsilon$  is independent of  $X$  conditional on  $Z\gamma$ , the selection model specified by Eqs. (2) and (3) implies that the selection correction term  $Q_\tau(v_\tau|Z, D = 1)$  is some unknown function of both  $X$  and  $Z\gamma$ , see discussion of Eq. (5) in Sect. 3.1.

### 3.3 Huber–Melly test for conditional independence

Huber and Melly (2015) propose a quantile regression based test for the conditional independence assumption, which says that the joint density of  $v$  and  $\varepsilon$  is independent of  $Z$  conditional on  $Z\gamma$ . As noted by Huber and Melly (2015), Buchinsky's approach

<sup>5</sup> In principle, one could use the identification-at-infinity sample to estimate selection-corrected quantile regression coefficients consistently. Such an approach hinges on the correctness of the linear specification of the regression model. However, we are also interested in estimating the selection effects explicitly. This allows us to investigate whether our estimation approach fits well the unconditional wage distribution in the selective sample (as a safeguard against an incorrect parametric specification of the quantile regressions) and to estimate the counterfactual unconditional wage distribution for different selection probabilities.

builds upon this conditional independence assumption, which implies homogeneous slope coefficients across all quantiles, see discussion of Eq. (2) in Sect. 3.1.<sup>6</sup>

We illustrate this point in the following. Conditional independence implies for the joint density of  $v$  and  $\varepsilon$

$$f_{v,\varepsilon}(\cdot|Z) = f_{v,\varepsilon}(\cdot|Pr(D = 1|Z)) = f_{v,\varepsilon}(\cdot|Z\gamma). \quad (7)$$

When there is no sample selection, i.e.,  $Pr(D = 1|Z) = 1 \forall Z$ , Eq. (7) implies that  $v$  and  $\varepsilon$  are independent of  $Z$ . Under conditional independence, the quantile regression coefficients  $\beta_\tau$  are identified when controlling for the selection bias term  $Q_\tau(v_\tau|Z, D = 1)$  only by flexible function of  $Z\gamma$  as in Buchinsky (1998, 2001), see also Huber and Melly (2015, Sect. 2.2).

Conditional independence in Eq. (7) also holds for  $v_\tau$  and  $\varepsilon$ , implying that  $Q_\tau(v_\tau|Pr(D = 1|Z), D = 1) - Q_\tau(v|Pr(D = 1|Z), D = 1)$  does not depend upon  $Z$  conditional upon the selection probability. Thus, the term  $X(\beta_\tau - \beta)$  only involves a constant difference in the intercept, meaning that the slope coefficients in  $\beta_\tau$  do not depend upon  $\tau$ .

When the conditional independence assumption does not hold, slope coefficients  $\beta_\tau$  may vary across quantiles, which is typically a motivation as to why researchers apply quantile regression in the first place. This limits the applicability of Buchinsky's approach.

Huber and Melly (2015) suggest a test based on the entire process of quantile regression coefficients to investigate whether the conditional independence assumption holds. They estimate quantile coefficients for a fine grid of quantiles across the distribution and then test the null hypothesis that the slope coefficients are identical. Violations of the null hypothesis are detected by using Kolmogorov–Smirnov (KS) and Cramér–von Mises (CM) test statistics to the coefficient process across quantiles. In practice, Huber and Melly use a grid of quantiles and suggest to implement the test for a range from the 10th to the 90th percentile as a starting point. The first stage is estimated using the semiparametric Klein and Spady (1993) estimator. The sample selection correction is based on a polynomial in the inverse mills ratio of the estimated index function estimated. Inference is based on resampling the influence function of the quantile regression estimator, building on the differentiability of the selection correction function to take account of the first stage estimation error.

### 3.4 Our approach

In short, we first implement Buchinsky's approach based on the original data and then apply the conditional independence test which strongly rejects. This is why we suggest to transform the dependent variable to account of heteroscedasticity in the original data and then apply Buchinsky's approach on the transformed dependent variable. Relying on identification at infinity, the transformation is based on quantile regressions for the subsample with a very high probability of participating. In our application, we are successful in finding a transformation after which the Huber–Melly test passes. Note

<sup>6</sup> The conditional independence assumption is implied by Assumptions C and E in Buchinsky (1998).

that it is not guaranteed to find such a transformation and we perform a specification search to find a proper transformation. If the conditional independence assumption is not rejected for the transformed model, we can use the transformed model to account for selection bias. Transforming back the dependent variable allows us to estimate counterfactual distributions in absence of selection or in the presence of a different selection mechanism.<sup>7</sup>

Now, we describe in detail different steps of our approach:

1. To estimate the probability to be in the selective sample, we estimate a Probit regression  $Pr(D = 1|Z) = \Phi(Z\gamma)$ , assuming that the distribution of  $\varepsilon$  in Eq. (3) is independent of  $Z$ .<sup>8</sup>
2. Based on the Probit estimates in step 1), a subsample of the data is determined for which identification at infinity is plausible, i.e., selection is negligible. We estimate standard quantile regressions based on this identification-at-infinity subsample. Using coefficient estimates  $\delta_u, \delta_l$  at the upper quantile  $u$  and the lower quantile  $l$ , respectively, we then estimate the predicted conditional quantile differences ( $l$  and  $u$  are tuning parameters)

$$\sigma(X, \delta) = X\delta_u - X\delta_l \quad (8)$$

for a worker with characteristics  $X$ . The transformation then involves dividing  $Y$  by  $\sigma(X, \delta)$ .<sup>9</sup>

3. Next, we run selection-corrected quantile regressions for the transformed outcome:

$$Q_\tau \left( \frac{Y}{\sigma(X, \delta)} \middle| X \right) = X\check{\beta}_\tau + g(\theta_\tau, Z\gamma). \quad (9)$$

We specify the selection correction as a piecewise constant function, with  $g(\theta_\tau, Z\gamma) = \sum_{j=1}^4 \theta_{\tau,j} I(Z\gamma \in Q_j)$  involving dummies for four quintiles of the propensity score  $I(Z\gamma \in Q_j)$  and  $\theta_\tau = (\theta_{\tau,j})_{j=1,\dots,4}$  (the highest quintile  $Q_5$

<sup>7</sup> The basic idea to transform the dependent variable is similar to the companion paper Biewen et al. (2020), which estimates the selection bias in employment for the estimation of the gender wage gap. However, there are two key methodological differences. First, our paper only transforms the dependent variable while leaving the covariates unchanged, while Biewen et al. (2020) transform both the dependent variable and the covariates. Second, our approach to determine the transformation factor relies on the identification-at-infinity approach, which is plausible in our setting. Biewen et al. (2020) assume a location-scale model  $Y^* = X\beta + g(x)u$ , where  $u$  is the rank in the conditional distribution of  $Y^*$  given  $x$  and derive a transformation factor based on the estimated conditional dispersion in the selective sample under the assumption that the dispersion of ranks in the selective sample is a function of the first-stage selection probability.

<sup>8</sup> Huber and Melly (2015) use the alternative semiparametric estimator suggested by Klein and Spady (1993), which is also part of the implementation of the test provided by Huber and Melly (2015). We have experimented with both approaches (Probit and Klein and Spady, 1993) for some cases in our application and find little difference between the two with regard to the fitted probabilities (the comparison is available upon request). For simplicity and for computational reasons, the empirical analysis in this paper is based on the Probit regressions for the first stage.

<sup>9</sup> This is analogous to the heteroscedasticity correction approach of Chen and Khan (2003), using a heteroscedasticity correction based on the inter-quartile range of the conditional distribution.

represents the omitted category).<sup>10</sup> Then, as our implementation of the Huber–Melly test for conditional independence, we implement a Wald test of the equality of the slope coefficients  $\check{\beta}_\tau$  along a grid of  $\tau$ .

4. This step assumes that the conditional independence test in the previous step passes. We run OLS for the transformed model for the identification-at-infinity sample and then estimate the selection effect based on quantile regressions of the OLS residuals based on the entire sample.<sup>11</sup> We then use the implied residuals based on entire sample to estimate the selection effects along the distribution.
5. Finally, we undo the transformation by multiplying the coefficients with  $\sigma(X, \delta)$ .

For simplicity, we implement the conditional independence test as a Wald test of the equality of slope coefficients over an equi-spaced grid of quantiles. Our application differs from Huber and Melly (2015) regarding the following three issues, which prevent us from using their implementation. First, bootstrapping the entire estimation process, inference takes account of the estimation error in all stages including the transformation. Second, applying a weighted cluster bootstrap inference avoids non-convergence of the Probit in the first stage and is cluster robust at the regional level, which is the level of the variation in the instruments.<sup>12</sup> Third, we approximate the selection correction term by a piece-wise constant selection correction function which is non-differentiable. Furthermore, implementing the Huber–Melly test for Buchinsky’s estimator using a polynomial in the inverse-Mills-ratio based on the untransformed model requires a lot of computation time due to our large sample size.

If the conditional independence test for the transformed model rejects, we use this for respecifying our estimation approach. Note as a caveat that inference for our Wald tests for homogeneous slopes does not take account of the fact that we search for a transformation such that the conditional independence test passes. Hence, multiple hypotheses testing is a concern given that we search for the proper specification of the transformation model.<sup>13</sup> A key point is that in contrast with the standard concern in the literature about searching for significant effects by running different model specifications, here we search for a transformation of the dependent variable which leads to a non-rejection. Thus, standard approaches (e.g., Bonferroni/Holm) to adjust critical values ( $p$ -values) under the zero hypothesis do not apply—rather power concerns arise. Our approach involves testing different (typically incompatible) zero hypotheses, and the validity of the final estimates hinges on the nonrejected zero hypothesis being true. To explore whether the first-best transformation involves a singular non-rejection, we also report the results for the second-best transformation.<sup>14</sup> The latter prove very close

<sup>10</sup> This specification yields better fits and more reliable findings than using a polynomial in the inverse Mills’ Ratio  $\lambda$  (detailed results are available upon request).

<sup>11</sup> The more salient approach would be instead to apply Buchinsky’s estimator literally by estimating quantile regressions for the transformed model with selection correction based on the full sample. However, doing so yields noisy estimates along the distribution despite not reflecting significant differences according to the Huber/Melly test. In contrast, the OLS estimates yield more satisfactory results regarding the fit of the observed distribution and the predicted counterfactual distribution.

<sup>12</sup> The code provided by Huber and Melly (2015) could be adjusted to provide cluster robust inference.

<sup>13</sup> “Appendix A.1” of the companion paper Biewen et al. (2020) provides a detailed discussion of the issue and what to do about it.

<sup>14</sup> We are grateful to a referee who suggested to investigate further admissible transformations.

to those of the first-best ones, thus strengthening our findings. As an additional robustness check, we perform a random split of the sample into a training sample to estimate the transformation model and a validation sample to perform the conditional independence test and to estimate the selection-corrected quantile regressions. Our findings show that the transformation model from the training sample implies a non-rejection of the conditional independence test when implemented for the validation sample. Also, the model fit in the validation sample is very good. These additional findings are available upon request.

As part of our specification search, we investigate which quantile regression coefficients change strongly across quantiles. To illustrate this point, note that, based on preliminary estimates, the conditional independence tests never passed for a model pooling both education groups. Therefore, we conclude that the nature of the selection bias differs between the two education groups, which motivates us to estimate separate models by education group.<sup>15</sup>

### 3.5 Counterfactual wage distribution under alternative selection rules

We use the estimated selection-corrected quantile regressions to estimate the counterfactual wage distribution under different selection rules. We estimate the counterfactual distribution using a selection-corrected Melly (2006) approach as in Albrecht et al. (2009) (see also Machado and Mata 2005; Chernozhukov et al. 2013), while taking account of the transformation of the outcome. Let  $Z$ ,  $X$ ,  $g(Z\gamma)$  apply to the observed sample and  $\tilde{Z}$ ,  $\tilde{X}$ , and  $g(\tilde{Z}\tilde{\gamma})$  to the counterfactual sample, where  $\tilde{\gamma}$  represents the counterfactual selection rule. Specifically, we estimate two counterfactuals: First, the wage distribution if all individuals in the sample were employed, and, second, the wage distribution if the selection rule of a different calendar year applies. The first counterfactual involves the covariates  $\tilde{X}$  of the entire sample and sets  $g(\theta_\tau, \tilde{Z}\tilde{\gamma})$  equal to zero, i.e.,  $\theta_\tau = 0$ , corresponding to a selection probability of one. For the second counterfactual,  $\tilde{Z}$  and  $\tilde{X}$  represent the employees and  $g(\tilde{Z}\tilde{\gamma})$  their selection rule (implied by the first stage Probit estimates) in the different calendar year.<sup>16</sup>

Our implementation of the Melly (2006) approach uses predictions of conditional quantiles for a fine grid of equi-spaced  $\tau \in [0.01, 0.02, \dots, 0.99]$  for each observation in the counterfactual sample to estimate the conditional distribution of log wages. The

<sup>15</sup> Also, Machado (2017) finds differences in direction of selection across different sociodemographic groups.

<sup>16</sup> Based on estimating selection-corrected distribution functions, Chernozhukov et al. (2019) and Fernandez-Val et al. (2019) propose a decomposition of changes in wage inequality over time which allows to distinguish the effects of changes in observables, in coefficients for observables, and in selection on unobservables (both regarding selection probabilities and the link between selection probabilities and the dependent variable). Chernozhukov et al. (2019) distinguish further between the selection structure effect (the change in selection probabilities) and the change in selection sorting (the aforementioned link). The second counterfactual, we estimate, quantifies the selection structure effect. However, this is done in comparison with a counterfactual which keeps selection of observables as in the base year and we assume that both parts of the selection on unobservables remain unchanged.



counterfactual conditional quantile is

$$Q_{\tau}(Y|\tilde{Z}) = \sigma(\tilde{X}, \delta) \left[ \tilde{X} \check{\beta}_{\tau} + g(\theta_{\tau}, \tilde{Z} \tilde{\gamma}) \right],$$

where  $\check{\beta}_{\tau}$ ,  $\delta$ , and  $g(\theta_{\tau}, \cdot)$  (including the definition of the quintile dummies) are estimates based on the observed sample.

We then stack the 99 predictions for all individual observations in the counterfactual sample represented by  $(\tilde{Z}, \tilde{X})$  and calculate the unconditional empirical quantiles of the entire expanded sample, where the number of observations is 99 times the number of observations in the counterfactual sample. This counterfactual distribution, denoted by  $T_Y(\tilde{X}, \check{\beta}, \delta, \theta, \tilde{\gamma})$ , represents the counterfactual distribution of  $Y$  for the sample with characteristics  $\tilde{Z}$ , the alternative selection rule  $\tilde{\gamma}$ , the selection-corrected coefficients for the transformed model  $\check{\beta}$ , the coefficients of the selection correction terms  $\theta$ , and the transformation coefficients  $\delta$ .

The difference between the observed wage distribution, which is denoted by  $TO_Y$  representing the quantiles of  $Y$  in the selective observed sample with  $D = 1$ , and the counterfactual distribution  $T_Y(\tilde{Z}, \check{\beta}, \delta, \theta, \tilde{\gamma})$  is given by

$$TO_Y - T_Y(\tilde{Z}, \check{\beta}, \delta, \theta, \tilde{\gamma}). \quad (10)$$

This difference measures the total effect of selection relative to the counterfactual.

We can now decompose the total selection effect into a component due to differences in observed characteristics driving wages, i.e., the difference between  $X$  and  $\tilde{X}$ , and a component due to differences in selection based on unobservables. To this end, we calculate the counterfactual distribution denoted by  $T_Y(\tilde{X}, \alpha)$  based on running linear quantile regressions using  $X$  from the observed sample of employees (without transformation) and then predicting the counterfactual distribution for the sample with  $\tilde{X}$  using the Melly (2006) approach as described above. Here,  $\alpha$  involves the quantile regression coefficients for the observed sample.

The total selection effect in Eq. (10) can be decomposed into the effect of changes in observable characteristics

$$TO_Y - T_Y(\tilde{X}, \alpha), \quad (11)$$

and the residual effect of selection on unobservables

$$T_Y(\tilde{X}, \alpha) - T_Y(\tilde{Z}, \check{\beta}, \delta, \theta, \tilde{\gamma}). \quad (12)$$

We now discuss the two cases separately. The first counterfactual wage distribution which would prevail if all observed individuals in a given year, both full-time workers and unemployed, were employed and earning market wages is obtained by setting  $\theta_{\tau}$  equal to zero. Then, Eq. (10) defines the total effect of selection into work, which is decomposed into the selection effect due to observables [Eq. (11)] and the effect of selection on unobservables [Eq. (12)] when contrasting full-time workers with the total sample of full-time workers and unemployed.

The second counterfactual wage distribution allows us to study the effect of changes in selection over time. To estimate this counterfactual, we fix the conditional probability of selection into full-time work, i.e., the index  $Z\gamma$ , and the distribution of observed characteristics fixed at the level of the base year. Using the coefficient estimates obtained in the observation year (in our application the year 2010), we estimate the counterfactual wage distribution under the selection rule of a base year (in our application the year 1995). Let the index  $b$  denote the base year and  $o$  the observation year.

Then,

$$TO_Y^o - T_Y(Z^b, \check{\beta}^o, \delta^o, \theta^o, \tilde{\gamma}^b) \quad (13)$$

is the total selection effect. It can be decomposed as above into the effect of the change between base year and observation year in the selection of observables and in the selection on unobservables, both among full-time workers. To account for the selection of observables, we estimate the counterfactual distribution  $T_Y(\tilde{X}, \alpha)$  [as in Eq. (11)] with observables in the employment sample 1995  $\tilde{X}$  and coefficients  $\alpha$  for wage regressions among the employed in 2010. To account both for selection on observables and unobservables, we estimate  $T_Y(Z^b, \check{\beta}^o, \delta^o, \theta^o, \tilde{\gamma}^b)$  [as in Eq. (12)] where  $\check{\beta}^o, \delta^o, \theta^o$  represent the coefficient estimates of our selection-corrected quantile regressions in ( $o \Rightarrow$ ) 2010.  $Z^b$  are the sample characteristics for the employed in 1995,  $\tilde{\gamma}^b$  the coefficients of the selection model in 1995, and  $Z^b \tilde{\gamma}^b$  determines the 1995 selection probability.

The following standard caveat applies: These counterfactual distributions do not account for general equilibrium effects which might potentially lead to changing returns to skills in response to an influx of previously unemployed into employment [see the detailed discussion in Fortin et al. (2011)]. One likely response to such an influx would be falling returns to those skill levels over-represented among the unemployed, e.g., low levels of education. Therefore, returns to education might increase due to higher relative scarcity. Then, the estimated counterfactual wage distribution would be less dispersed than the one arising when all unemployed are employed and general equilibrium effects operate.

## 4 Empirical application

### 4.1 Selection equation: step 1

Our decomposition method with sample selection correction requires instruments which affect the employment status but which do not affect wages. We run separate Probit regressions of the full-time indicator by education group, i.e., separately for the low-educated and the medium-educated.<sup>17</sup>

<sup>17</sup> This is done for two reasons. First, the propensity scores based on a Probit regression pooling the two education groups and using the same regional instruments differ notably from those based on Probit regressions by education group. Second, we could not find a transformed model passing the Huber–Melly test when we account for selection based on a pooled Probit. Detailed results are available upon request.

For the medium-educated, the Probit regression accounts for the following covariates, which are also allowed to affect wages: Age, age squared, number of days spent in full-time work over the last 5 years, and number of days spent in part-time work over the last 5 years. As instruments, which are measured at the district level (as population shares) and which are excluded in the wage equation, we account for share of lower secondary graduates, share of upper secondary graduates in the district, share of individuals aged 18–24, and share of individuals aged 25–30. The employment history variable account for the recent employment experience being associated with current full-time employment, thus accounting either for state dependence or for unobserved heterogeneity causing persistence in employment outcomes. Later these covariates are also used as control variables in the wage regression accounting for experience effects. We use labor supply instruments at the district level, assuming that wages are not affected by these supply instruments in the short run.<sup>18</sup> Because we account for recent employment experience both in the selection equation and in the wage equation, this is compatible with labor supply changes affecting wages in the medium run through changes in work experience.<sup>19</sup> All covariates in the selection model are highly predictive for full-time employment among the medium-educated and have the expected signs (see Table 2, columns 2 and 3).<sup>20</sup> The excluded instruments are highly significant with an *F*-statistic of 20.4 in 1995 and 29.7 in 2010.

We also estimated the same specification for the low-educated; however, the instruments were nowhere close to being significant.<sup>21</sup> Because the medium-educated are the larger group and the low-educated may be complements to the medium-educated, we use the average fitted employment rate of the medium-educated at the district level based on the estimated selection equation in Table 2, columns 2 and 3, respectively, as alternative instrument for the selection equation of the low-educated. This fitted employment rate is a function of the labor supply measures used for the medium-educated.<sup>22</sup> The results for the low-educated are reported in Table 2, columns 4 and 5.

---

Footnote 17 continued

We conclude that the selection into full-time work and the effect of selection on observed wages differ between the two education groups.

<sup>18</sup> This is an identifying assumption, which cannot be tested, and, unfortunately, we are not aware of auxiliary empirical evidence supporting or questioning this assumption. The exogeneity of the instruments in the short run is plausible because wage setting in Germany responds to local labor market conditions in a sluggish way (e.g., because of collective bargaining).

<sup>19</sup> Even though our labor supply measures are likely to be serially correlated, a possible confounding effect on wages of the population shares of currently young workers in the past is not very likely because the young workers are not yet part of the labor force or have only recently entered the labor force.

<sup>20</sup> At first glance, the positive coefficient for the share of the 25–30 years old may be surprising. Note that the dependent variable is full-time employment among the 25–55 years old. An increase in the share of 18–24 years old shows a negative effect effectively replacing employment among the 25–55 years old. Holding the share of 18–24 years old constant, the positive effect of the share of the 25–30 years old reflects the fact that their employment chances are higher when their share is larger. The sum of the two coefficients is negative reflecting a negative effect of overall labor supply.

<sup>21</sup> We omit the detailed results which are available upon request. Note that in a specification pooling both education groups the instruments are significant. Recall, however, that pooling was rejected by the data.

<sup>22</sup> For the same reasons as for the medium-educated, we do not think that these labor supply measures affect wages of the low-educated in the short run conditional upon the covariates used in the wage equation (see Footnote 18).

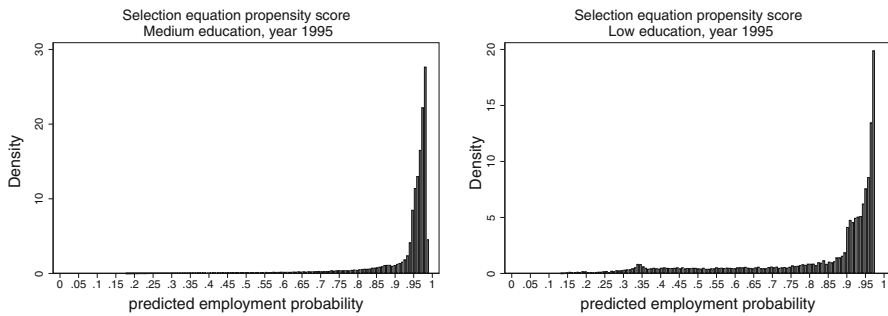
**Table 2** First stage: Probit regression for full-time employment

	Medium-educ 1995	Medium-educ 2010	Low-educ 1995	Low-educ 2010
Age	−0.0468 (0.0556)	−1.302*** (0.0634)	0.517*** (0.103)	−0.559*** (0.105)
Age squared	−0.0185*** (0.00693)	0.139*** (0.00788)	−0.0899*** (0.0128)	0.0481*** (0.0135)
Part-time	0.193*** (0.0239)	0.344*** (0.0128)	0.102** (0.0440)	0.285*** (0.0187)
Last 5 years	0.494*** (0.00352)	0.625*** (0.00336)	0.464*** (0.00584)	0.610*** (0.00556)
Full-time				
Last 5 years				
LS grad rate	9.791 (6.407)	−2.493 (5.174)	—	—
HS grad rate	29.21*** (6.682)	11.10** (4.904)	—	—
Share age	3.888*** (0.911)	3.930*** (0.645)	—	—
25–30 years				
Share age	−6.386*** (2.044)	−5.994*** (1.484)	—	—
18–24 years				
Fitted prob.	—	—	1.513*** (0.442)	1.282*** (0.224)
Medium-educ				
N	158,216	143,120	30,860	32,022
F_stat Instr.	20.4	29.7	11.7	32.9

Notes: Probit coefficients. Standard errors clustered at the district level in parentheses. Instruments for selection (rows 8–11) are measured at district level. The instruments for the medium-educated represent exogenous labor supply shocks due to variation in cohort sizes of individuals entering the labor market. For the low-educated, the instrument 'Fitted Prob. Medium' represents the average fitted employment probability for the district level based on the Probit regressions in the second and third column, respectively. F\_stat Instr. denotes the *F*-Statistic for significance of the instruments. \* $p < 0 : 10$ , \*\* $p < 0 : 05$ , \*\*\* $p < 0 : 01$

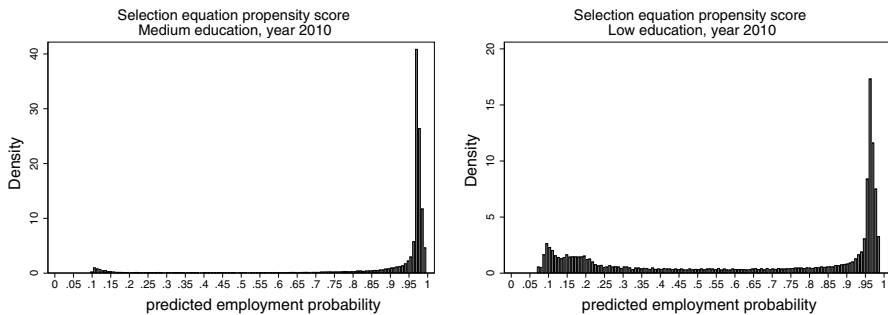
This instrument proves highly significant with an *F*-statistic of 11.7 in 1995 and 32.9 in 2010, implying that a higher employment rate of the medium-educated induced by labor supply changes also increases the employment rate of the low-educated. We interpret this as evidence for the low-educated being complements of the medium-educated.

As discussed in Sect. 3.4, identification at infinity in the outcome model requires that the selective sample of the employed contains a sizeable number of observations with a propensity score close to one, i.e., the regressor matrix restricted to these observations must have full rank. Figures 3 and 4 show that the distribution of the propensity score for the sample of employed and unemployed is concentrated close to one in all cases. Table 3 shows selected quantiles of the distribution of propensity scores for the selective sample of the employed. For the medium-educated, the median is 97% (97%) and the lower quartile is 95% (96%) in 1995 (2010). For the low-educated, the median is 94% (96%) and the lower quartile is 87% (78%) in 1995 (2010). Based



**Fig. 3** Distribution of propensity scores in full sample by education group, year 1995.

*Notes:* Propensity scores for being selected into full-time employment for the full sample of both employed and unemployed individuals based on estimates in Table 2. Source SIAB7510, own calculations



**Fig. 4** Distribution of propensity scores in full sample by education group, year 2010.

*Notes:* Propensity scores for being selected into full-time employment for the full sample of both employed and unemployed individuals based on estimates in Table 2. Source SIAB7510, own calculations

**Table 3** Probability of selection among employed

Year	Medium-educated			Low-educated		
	25% quantile	Median	75% quantile	25% quantile	Median	75% quantile
1995	.945	.965	.977	.868	.936	.964
2010	.962	.970	.977	.778	.957	.967

*Notes:* Median and quartiles of the propensity score distribution restricted to the subset of full-time workers

on these findings, we conclude that the identification-at-infinity approach described above is quite plausible for our application.

## 4.2 Conditional independence test for Buchinsky's approach

We estimate Buchinsky's approach without transformation for selection-corrected quantile regressions using dummies for the quintiles of the propensity score to account for selection. Predicting the observed wage distribution in the employment sample

**Table 4** Conditional independence tests for equality of slope coefficients in selection-corrected quantile regressions without transformation (*P*-Values)

Covariates	Test range for $\tau$	Medium-educated		Low-educated	
		1995	2010	1995	2010
All	80–20	.000	.000	.000	.000
All	70–30	.000	.000	.000	.000
All	60–40	.000	.000	.000	.000
Age + Age squared	80–20	.000	.000	.000	.000
Age + Age squared	60–40	.000	.000	.000	.000
Part-time 5 years	80–20	.000	.000	.000	.000
Part-time 5 years	60–40	.863	.278	.001	.000
Full-time 5 years	80–20	.000	.000	.000	.000
Full-time 5 years	60–40	.000	.000	.000	.000

Notes: *P*-values of specification tests under the null-hypothesis of conditional independence as in Huber and Melly (2015), testing for equality of the slope coefficients  $\beta_\tau$  in Eq. (6) over  $\tau$  (e.g., range 80–20 denotes  $\tau = .2, .25, \dots, .75, .8$ ). Selection-corrected quantile regressions as suggested by Buchinsky (1998). Wald tests on an equi-spaced five-percent-grid over the stated range for  $\tau$  of the conditional distribution

using the Melly (2006) approach yield a close correspondence between the model prediction and the actual distribution.<sup>23</sup>

Our implementation of the Huber–Melly test of equal slope coefficients  $\beta_\tau$  for the selection-corrected quantile regressions involves selected Wald tests, whose results are reported in Table 4. For the test range 80–20 ( $\tau = .2, \dots, .8$ ), the test statistics decisively reject in all cases. This also happens for the narrower test range 60–40 when implementing the test for all covariates. Only for the covariate part-time during the last 5 years, the test does not reject for the narrower test range. The rejection for all covariates is robust to other test ranges in between (detailed results are available upon request). We conclude that Buchinsky’s approach based on quantile regressions for log wages is not applicable for our application.

### 4.3 Transformation and conditional independence test: steps 2 and 3

We use an identification-at-infinity sample to estimate the transformation factor  $\sigma(X, \delta)$  in step 2 of our approach. For this, we use observations with a predicted probability above 90%/85% in 1995/2010 for the low-educated and above 97.5%/98% in 1995/2010 for the medium-educated, respectively. Based on different choices for the quantile range used for the transformation, we estimate quantile regressions with selection correction as in step 3. We use an equi-spaced grid of five-percentile intervals as possible choices for the upper and lower point of the transformation range. Then, we undertake the conditional independence tests and base our choice of the transformation factor, i.e., the choice of  $\delta_l, \delta_u$  for the quantile differences used, on the

<sup>23</sup> In contrast, using a low order polynomial in the inverse Mills ratio did not result in satisfactory within sample fit. Detailed results are available upon request.

**Table 5** Conditional independence tests for equality of slope coefficients in selection-corrected quantile regressions after transformation (*P*-Values)

Covariates	Test range for $\tau$	Medium-educated		Low-educated	
		1995	2010	1995	2010
All	80–20	.385	.000	.000	.000
All	70–30	.956	1.00	.031	.394
All	60–40	.977	1.00	.541	.669
Age + Age squared	80–20	.999	.998	.474	.090
Age + Age squared	60–40	1.00	.992	.620	.751
Part-time 5 years	80–20	.983	1.00	.155	.345
Part-time 5 years	60–40	.993	1.00	.134	.507
Full-time 5 years	80–20	1.00	.997	.035	.345
Full-time 5 years	60–40	.985	.939	.672	.972
Range ( $u - l$ ) for Transformation ( $\delta_u, \delta_l$ )		80–40	75–35	80–30	75–25

Notes: *P*-values of specification tests under the null-hypothesis of conditional independence as in Huber and Melly (2015), testing for equality of the slope coefficients  $\beta_\tau$  in Eq. (6) over  $\tau$  (e.g., range 80–20 denotes  $\tau = .2, .25, \dots, .75, .8$ ). Selection-corrected quantile regressions based on transformed model, with transformation based on predicted quantile difference  $\sigma(X, \delta)$  in the identification-at-infinity sample, with  $\delta = (\delta_u, \delta_l)$  for range ( $u - l$ ). Wald tests on an equi-spaced five-percent-grid over the stated range for  $\tau$  of the conditional distribution

test results. The findings are reported in Table 5 for our preferred models passing the conditional independence test.

In all cases, the conditional independence test passes for the narrow range 60–40 [ $u - l = 60\% - 40\%$ ] and for all individual covariates for both reported ranges. For the medium-educated, the test passes for all covariates for 70–30 and also in 1995 for 80–20. For the low-educated, the test passes for 70–30 in 2010 and barely so at a 3%-level in 1995. There are a three clear rejections for 80–20 considering all covariates, even though for the individual covariates the test passes in all cases. Note that Huber and Melly (2015) caution themselves regarding the behavior of their conditional independence tests when moving into the tail of the distribution. The comparison between Tables 4 and 5 shows that the transformation does a very good job in reducing the differences in slope coefficients.

To reduce concerns about a potential multiple testing problem, Table 6 reports the test results for the second-best set of transformation factors  $\sigma(X, \delta)$  from the grid of possible choices. Therefore, significance levels are slightly higher, but the test passes under the same conditions, for the same education groups and the same years as in 5.

We conclude that the conditional independence assumption is plausible for the transformed model, and the evidence is somewhat stronger for the medium-educated than for the low-educated.<sup>24</sup> Keeping this in mind, we will be very cautious in not to over-interpret the estimated selection effects for the low-educated.

<sup>24</sup> Note that for quantile regressions after transformation pooling low-educated and medium-educated the conditional independence test is nowhere close to pass, i.e., the conditional independence test is not guaranteed to pass after a mechanical transformation. Detailed results are available upon request.

**Table 6** Conditional independence tests for equality of slope coefficients in selection-corrected quantile regressions after second-best transformation (*P*-Values)

Covariates	Test range for $\tau$	Medium-educated		Low-educated	
		1995	2010	1995	2010
All	80–20	.000	.000	.000	.000
All	70–30	.937	1.00	.026	.381
All	60–40	.999	1.00	.384	.615
Age + Age squared	80–20	.899	.962	.095	.083
Age + Age squared	60–40	1.00	.949	.186	.826
Part-time 5 years	80–20	.990	1.00	.198	.910
Part-time 5 years	60–40	.958	1.00	.370	.705
Full-time 5 years	80–20	.932	.906	.012	.492
Full-time 5 years	60–40	.980	.885	.455	.228
Range ( $u - l$ ) for second-best Transformation ( $\delta_u, \delta_l$ )		70–40	85–40	80–20	65–25

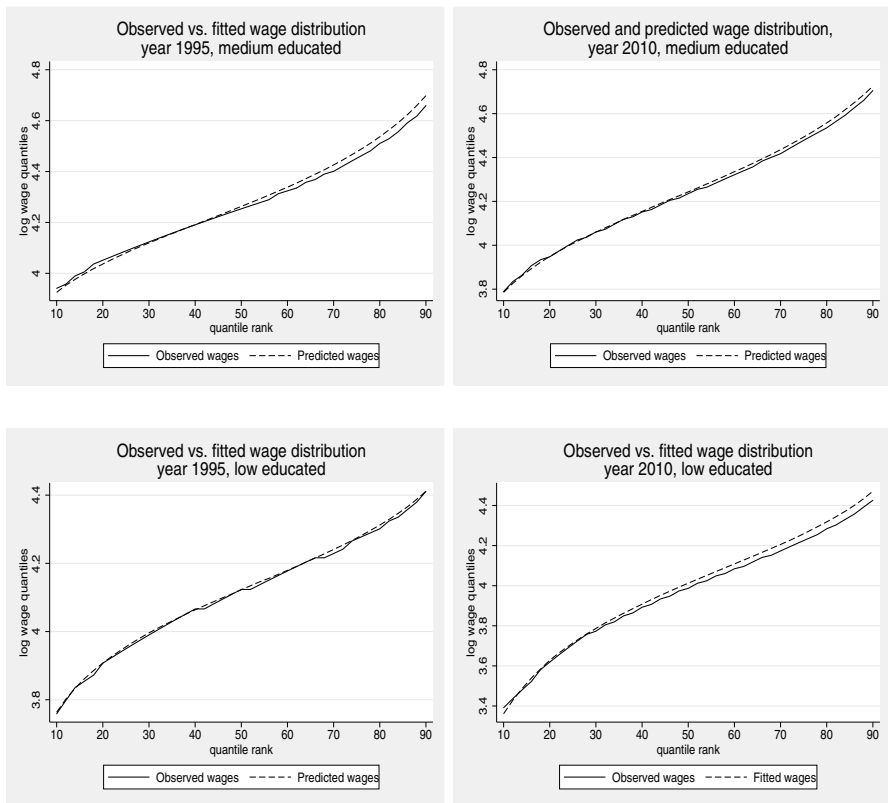
Notes: *P*-values of specification tests under the null-hypothesis of conditional independence as in Huber and Melly (2015), testing for equality of the slope coefficients  $\beta_\tau$  in Eq. (6) over  $\tau$  (e.g., range 80–20 denotes  $\tau = .2, .25, \dots, .75, .8$ ). Selection-corrected quantile regressions based on transformed model, with transformation based on predicted quantile difference  $\sigma(X, \delta)$  in the identification-at-infinity sample, with  $\delta = (\delta_u, \delta_l)$  for range ( $u - l$ ). Wald tests on an equi-spaced five-percent grid over the stated range for  $\tau$  of the conditional distribution

#### 4.4 Goodness of fit and impact of selection: steps 4 and 5

Assuming that conditional independence holds, we run OLS regressions without selection correction on the identification-at-infinity sample after the transformation. Then, we calculate residuals for the entire employment sample based on the OLS coefficient estimates. For these residuals, we then run quantile regressions on an intercept and the selection correction terms. Under conditional independence, this focuses on the evolution of the selection effects along the conditional distribution. Adding the OLS-fitted values to the fitted values of the quantile regressions for the residuals provide the quantile regression fits for the transformed model, which then can be used to simulate the wage distribution for the employed as well as the counterfactual wage distribution if all unemployed were also employed. These simulations are based on the Melly (2006) approach.

Contrasting the actual and simulated wage distribution for the employed allows to assess the goodness-of-fit for the observed unconditional wage distribution. For all cases in Fig. 5, the fitted distribution closely tracks the actual distribution. Note that this is by no means obvious in light of our multi-step estimation approach. If the identification-at-infinity assumption was inappropriate or the transformation model/the model estimated for the transformed data were misspecified, the fitted distributions could differ from the actual distribution. The close fit between the actual and the fitted wage distribution also adds credibility to the estimated counterfactual





**Fig. 5** Actual and fitted wage distributions for employed.

*Notes:* Fitted wage distribution based on Melly (2006) approach. We use the model estimates for the transformed data and then undo the transformation. For the transformed data, we run OLS regressions based on identification-at-infinity sample and quantile regressions with selection correction based on the entire sample of employed

distributions discussed below.<sup>25</sup> Note that Fig. 5 shows the rise in wage inequality from 1995 to 2010. The 90–10 differential increases by about 15 log points for the medium-educated and by about 40 log points for the low-educated, with sizeable real wage losses in the lower tail of the distribution, especially for the low-educated.

What is the nature of the estimated selection effects? Table 7 reports the estimated average conditional selection effect  $[\sigma(X, \delta)g(\theta_\tau, Z\gamma)]$  for log wages after undoing the transformation for selected values of  $\tau$  for different values of the selection probability  $Pr(D = 1|Z) = \Phi(Z\gamma)$ .<sup>26</sup> Table 7 covers a wide range of selection prob-

<sup>25</sup> We also obtained the fitted wage distributions for the second-best transformations, see Table 6. The fit is equally good—as shown in Fig. 5—for all cases, except for the medium-educated in 1995 for whom the fit was only slightly worse in the upper tail of the distribution. These findings add credibility to our findings for the first-best transformation. The detailed findings are available upon request.

<sup>26</sup> Note that the calculation of average selection effects for a given selection probability across all sample observations does not account for the statistical dependence between actual selection probabilities and the covariates  $X$ . Because of the high selection probabilities in our application (Table 3), the fairly high average

abilities representing most of their support in the employment sample. For a very high selection probability of 99%, the selection effects are zero and they grow with smaller selection probabilities. Around the median selection probability, the selection effects are in the order of 10–20 log points across all quantiles showing sizeable positive selection into employment. Incidentally, the selection effects vary with  $\tau$ ; however, there is no common pattern across the four cases. For the medium-educated, they tend to increase with  $\tau$ , except for a very low selection probability in 1995. This suggests that for medium-educated selection effects grow with the rank in the conditional wage distribution. For the low-educated, the pattern along the conditional wage distribution is less clear. The selection effects are more similar for different  $\tau$ 's. Specifically, for very low selection probabilities the selection effect falls with  $\tau$ , similar to the medium-educated in 1995, but the selection effect increases slightly with  $\tau$  for intermediate values of the selection probability. While the estimated selection effects imply that there is strong positive selection into employment when selection probabilities are around 93%–97%, the range of the median in the four cases, these results do not allow us to quantify the selection effects along the unconditional wage distribution, which is what comes next.

Based on Sect. 3.5, we estimate the counterfactual distribution  $T_Y(\tilde{X}, \alpha)$  to account for the different selection of observables in the total sample  $\tilde{X}$ , where  $\alpha$  involves the quantile regression coefficients of log wages on  $X$  among the employed without selection correction. To account both for selection on observables and unobservables, we estimate  $T_Y(\tilde{Z}, \tilde{\beta}, \delta, 0, \tilde{\gamma})$  setting  $\theta = 0$ , because there is no need for selection correction when using the full sample. Effectively, we predict wages based on the transformation coefficients  $\delta$ , selection-corrected coefficient estimates  $\tilde{\beta}$ , and sample characteristics  $X$ . Figure 6 displays the counterfactual wage distributions if both the unemployed and the employed were working full-time. The distribution labeled with 'Sel. on observables' accounts for the differences in observables between employed and unemployed and 'full employment' accounts for both observables and unobservables.

There are three key similarities across the four cases. First, all counterfactual full-employment distributions lie for the most part below the corresponding observed wage distributions, except for the absence of selection on observables among the medium-educated in 1995. This means that the employed in the sample are positively selected with regard to wages. Thus, the counterfactual wage quantiles for the unemployed are lower than the corresponding wage quantiles for the employed. Second, the distribution accounting for observables typically lies between observed wages and the full employment distribution, implying that there is positive selection among employees both on observables and on unobservables. Third, the gap between observed wage quantiles and counterfactual wages is largest in the lower tail of the distribution; it falls along the distribution and closes in the upper tail. Hence, the negatively selected unemployed are concentrated in the lower tail of the wage distribution.

At the same time, there are some noteworthy differences across the four cases. The figures in the upper panel of Fig. 6 show that for the medium-educated in 1995 there is

---

selection effects for the lower selection probabilities considered may thus overstate the actual selection effect for those individuals in the sample for whom the selection probability applies.

**Table 7** Average conditional selection effect for log wages by selection probability

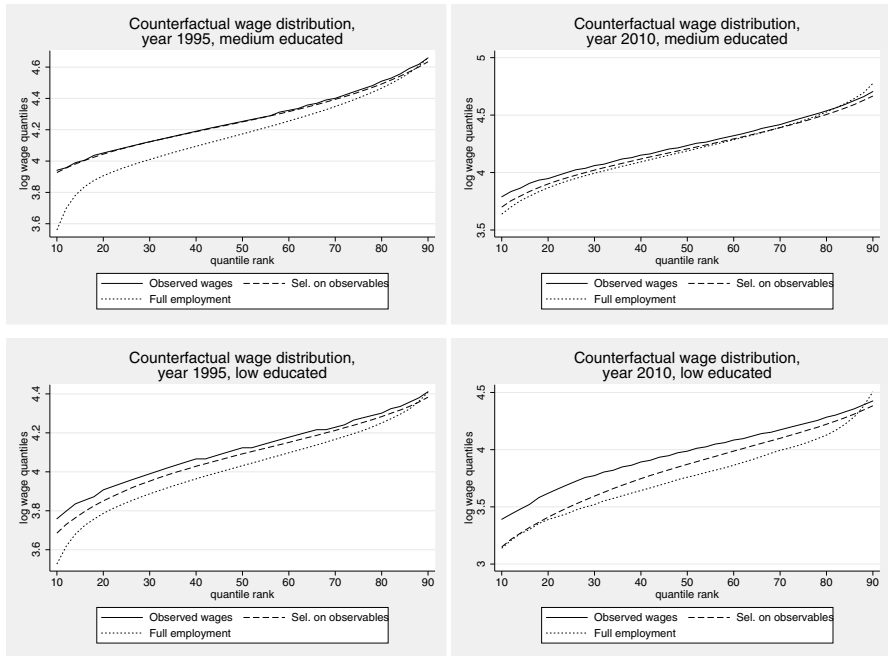
Selection probability (%)	Conditional quantile $\tau$				
	10%	25%	50%	75%	90%
<i>Medium-educated 1995</i>					
99	0.000	0.000	0.000	0.000	0.000
96	0.094	0.136	0.175	0.199	0.198
95	0.192	0.209	0.234	0.252	0.242
85	0.383	0.244	0.212	0.227	0.223
<i>Medium-educated 2010</i>					
99	0.000	0.000	0.000	0.000	0.000
97	0.139	0.186	0.247	0.300	0.323
96	0.258	0.257	0.278	0.301	0.303
85	0.417	0.415	0.469	0.543	0.582
<i>Low-educated 1995</i>					
99	0.000	0.000	0.000	0.000	0.000
95	0.117	0.119	0.125	0.119	0.121
93	0.228	0.239	0.243	0.236	0.240
88	0.362	0.292	0.253	0.238	0.243
<i>Low-educated 2010</i>					
99	0.000	0.000	0.000	0.000	0.000
96	0.094	0.136	0.175	0.199	0.198
95	0.192	0.209	0.234	0.252	0.242
85	0.383	0.244	0.212	0.227	0.223

Notes: Estimated average conditional selection effect  $[\sigma(X, \delta)g(\theta_\tau, Z\gamma)]$  for log wages after undoing the transformation. The selection effects are a function of the selection probability  $Pr(D = 1|Z) = \Phi(Z\gamma)$ . Further, they differ by the  $\tau^{th}$  quantile regression and by the transformation factor  $\sigma X, \delta$ . We calculate the average of  $[\sigma(X, \delta)g(\theta_\tau, Z\gamma)]$  among all workers in the employment sample for a given selection probability, irrespective of the worker's actual selection probability. Table 3 reports all quartiles of the selection probabilities

no selection on observables and strong positive selection on unobservables. The results differ for 2010, where we find small but positive selection on observables and much smaller positive selection on unobservables than in 1995. Further, the total selection effect over most of the distribution falls over time. Also for the low-educated, there are changing patterns of selection (lower panel of Fig. 6). While both types of selection seem almost equally important in 1995, the selection on observables dominates in the lower tail of the distribution and both types of selection become stronger above the median. We conclude that while selection on observables increased over time for both education groups the importance of selection on unobservables fell.

#### 4.5 Keeping selection as of 1995

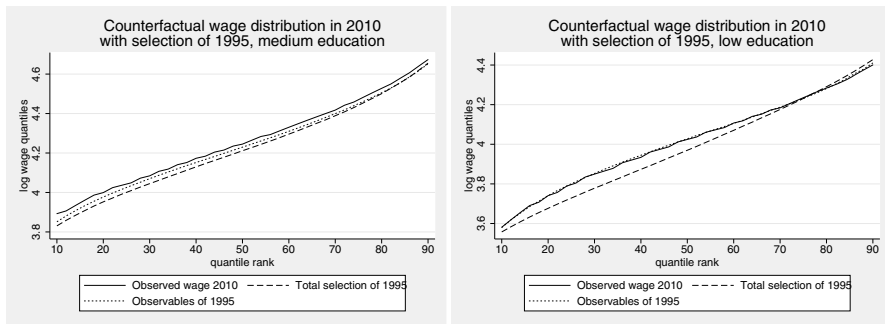
As the last step of our analysis, we estimate counterfactual wage distributions for 2010 assuming that either selection on observables or selection on observables and



**Fig. 6** Actual and counterfactual wage distributions.

*Notes:* Counterfactual wage distributions based on Melly (2006) approach, see Sect. 3.5. 'Full employment' and 'Sel.[ection] on observables' represent counterfactual wage distributions when both the unemployed and the full-time employed are working full-time. 'Sel. on observables' represents the situation where the wages are predicted based on standard quantile regressions on observed characteristics  $X$ , thus only accounting for differences in observable  $X$ . 'Full employment' represents the situation where wages are predicted based on the estimated quantile regressions with selection corrections, thus accounting both for differences in observables  $X$  and in unobservables. The counterfactual wage distributions also use predicted wages for the full-time employed

unobservables had remained at its values as of 1995, as described in Sect. 3.5. Figure 7 displays the two counterfactual wage distributions keeping selection as of 1995 together with the actual distribution in 2010.  $T_Y(\tilde{X}, \alpha)$  is denoted as 'Observables of 1995' and  $T_Y(Z^b, \check{\beta}^o, \delta^o, \theta^o, \tilde{\gamma}^b)$  as 'Total selection of 1995.' For both education groups, the effect of the change in the selection between 1995 and 2010 is small relative to the total selection effects within both years as shown in Fig. 6. A second common finding is that the counterfactual wage distribution under the total selection as of 1995 lies below the 2010 distribution. This applies to the total range of the distribution for the medium-educated and to the range below the 70%-quantile for the low-educated. For the medium-educated, the distribution with observables as of 1995 lies between the distribution observed in 2010 and the distribution with total selection of 1995. Further, both counterfactual distributions show slightly larger wage dispersion, as measured by the implied quantile differences. For the low-educated, the counterfactual with observables as of 1995 basically corresponds to the distribution of 2010; thus, the change in the selection of observables does not seem to have an impact. However, the distribution under total selection of 1995 shows lower wages below the 70%-quantiles



**Fig. 7** Actual wage distribution in 2010 and counterfactual wage distribution keeping selection as of 1995. *Notes:* Counterfactual wage distributions based on Melly (2006) approach, see Sect. 3.5. 'Observables of 1995' and 'Total selection of 1995' represent counterfactual wage distributions. 'Observables of 1995' represents the situation where the wages are predicted based on standard quantile regressions on observed characteristics  $X$ , thus only accounting for differences in observable  $X$  and assuming that selection on unobservables is as in 2010 (this is the counterfactual  $T_Y(\bar{X}, \alpha)$  defined in Sect. 3.5). 'Total selection of 1995' represents the situation where wages are predicted based on the estimated quantile regressions with selection corrections, thus accounting both for differences in observables  $X$  and in selection probabilities between 1995 and 2010 while keeping the selection coefficients as of 2010 (this is the counterfactual  $T_Y(Z^b, \tilde{\beta}^o, \delta^o, \theta^o, \tilde{\gamma}^b)$  defined in Sect. 3.5)

with a maximum gap around the 30%-quantile. This means that wage dispersion in the middle of the distribution, e.g., as measured by the interquartile differences, would have been higher under the selection as of 1995. However, the increase is lower when moving to the tails of the distribution.

Summing up, we conclude that with the selection of employees as of 1995 wage inequality would have been slightly higher in 2010. Despite the strong increase in wage inequality between 1995 and 2010, this finding suggests that the fall in unemployment up to 2010 by itself has not been associated with a change in the selection of employed toward higher inequality. Further, despite the strong fall in real wages in the lower tail of the distribution, the selection of the employed has changed toward higher wages.

## 5 Conclusions

As its methodological contribution, this paper proposes and implements a modification of selection-corrected quantile regressions. This modification addresses Huber and Melly's (2015) concern that using a control function approach as suggested by Buchinsky (1998) is only valid under equality of the slope coefficients on the determinants of the outcome variable, which is only observed in the selected sample. We propose estimating a transformation of the outcome variable based on the identification-at-infinity assumption and then estimate selection-corrected quantile regressions for the transformed dependent variable with the goal that equality of the slope coefficient then holds. A version of the test suggested by Huber and Melly (2015) is used to guide the choice of the transformation. We emphasize that whether the transformation approach works is specific to the application. Undoing the transformation provides nonlinear

selection-corrected quantile regressions for the outcome variable of interest which can be used to estimate counterfactual distributions.

Regarding the empirical analysis of wage inequality in Germany based on the suggested modification of selection-corrected quantile regressions, this paper addresses two questions. The first one is: What would the wage distribution be if all unemployed were working full-time? Our analysis focuses on medium- and low-educated in the years 1995 and 2010. As to be expected, the selection of the unemployed differs strongly from the full-timers. The unemployed are negatively selected in terms of wages with respect to both observed characteristics and unobservables driving the employment probability. If the unemployed were working full-time, they would be over-represented at the bottom of the wage distribution, and therefore, the overall wage dispersion would be higher. Negative selection is stronger among the low-educated than it is among medium-educated workers.

Our second question is: How would the wage distribution have developed if selection into full-time employment had not changed from 1995 to 2010? We find that for this counterfactual the level of wages in 2010 would have been lower in the lower and middle part of the wage distribution and wage inequality would have been slightly higher. Put differently, over time full-time workers have become less heterogeneous with regard to the factors driving wages as well as the selection into full-time work. This finding seems surprising in light of the existing literature emphasizing the role of composition changes in driving wage inequality (see Lemieux 2006; Dustmann et al. 2009; Biewen et al. 2018, among others). Further, selection due to unobservables did not contribute in a substantial way to the rise in within-group inequality for the medium-educated. Overall, our results suggest that the rise in wage inequality is not driven by previously unemployed individuals, who are negatively selected, entering full-time work. Two limitations regarding our findings are that we omit the high-educated because of the severe censoring in this group and that we only analyze within-education group inequality, both of which possibly explain some of the differences to the previous literature.

**Funding** Open Access funding enabled and organized by Projekt DEAL. We acknowledge financial support of this project by the German Science Foundation (DFG) through the project “Accounting for Selection Effects in the Analysis of Wage Inequality in Germany” (Project Nos.: BI 767/3-1 and FI 692/16-1).

## Declarations

**Conflict of interest** The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

**Financial interests** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted

by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Albrecht J, van Vuuren A, Vroman S (2009) Counterfactual distributions with sample selection adjustments: econometric theory and an application to the Netherlands. *Labour Econ* 16:383–396
- Andrews DW, Schafgans MM (1998) Semiparametric estimation of the intercept of a sample selection model. *Rev Econ Stud* 65(3):497–517
- Arellano M, Bonhomme S (2017) Quantile selection models with an application to understanding changes in wage inequality. *Econometrica* 85(1):1–28
- Bauer T, Bonin H, Goette L, Sunde U (2007) Real and nominal wage rigidities and the rate of inflation: evidence from West German micro data. *Econ J* 117(524):F508–F529
- Biewen M, Fitzenberger B, de Lazzer J (2018) The role of employment interruptions and part-time work for the rise in wage inequality. *IZA J Labor Econ* 7(1):10
- Biewen M, Fitzenberger B, Seckler M (2020) Counterfactual quantile decompositions with selection correction taking into account Huber/Melly (2015): an application to the German gender wage gap. *Labour Econ* 67
- Bollinger C, Ziliak JP, Troske KR (2011) Down from the mountain: skill upgrading and wages in Appalachia. *J Labor Econ* 29(4):819–857
- Buchinsky M (1998) The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *J Appl Econom* 13:1–30
- Buchinsky M (2001) Quantile regression with sample selection: estimating women's return to education in the U.S. *Empir Econ* 26:87–113
- Caliendo M, Schröder C, Wittbrodt L (2019) The causal effects of the minimum wage introduction in Germany-An overview. *Ger Econ Rev*
- Card D, Heining J, Kline P (2013) Workplace heterogeneity and the rise of West German wage inequality. *Q J Econ* 128(3):967–1015
- Chen S, Khan S (2003) Semiparametric estimation of a heteroskedastic sample selection model. *Econ Theor* 19:1040–1064
- Chernozhukov V, Fernández-Val I, Luo S (2019) Distribution regression with sample selection, with an application to wage decompositions in the UK. [arXiv:1811.11603v2](https://arxiv.org/abs/1811.11603v2)
- Chernozhukov V, Fernández-Val I, Melly B (2013) Inference on counterfactual distributions. *Econometrica* 81(6):2205–2268
- Das M, Newey WK, Vella F (2003) Nonparametric estimation of sample selection models. *Rev Econ Stud* 70(1):33–58
- D'Haultfoeuille X, Maurel A, Zhang Y (2014) Extremal quantile regressions for selection models and the black-white wage gap. In: No, IZA Discussion Papers, p 8256
- Dustmann C, Fitzenberger B, Schönberg U, Spitz-Oener A (2014) From sick man of Europe to economic superstar: Germany's resurgent economy. *J Econ Perspect* 28(1):167–88
- Dustmann C, Ludsteck J, Schönberg U (2009) Revisiting the German wage structure. *Q J Econ* 124:843–881
- Fernandez-Val I, Peracchi F, van Vuuren A, Vella F (2019) Decomposing changes in the distribution of real hourly wages in the US. [arXiv:1901.00419v2](https://arxiv.org/abs/1901.00419v2)
- Fitzenberger B, Winker P (2007) Improving the computation of censored quantile regressions. *Comput Stat Data Anal* 52(1):88–108
- Fortin N, Lemieux T, Firpo S (2011) Decomposition methods in economics. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, volume 4A of handbooks in economics, chapter 1. North Holland, Amsterdam, pp 2–97
- Gartner H (2005) The imputation of wages above the contribution limit with the German IAB employment sample. In: FDZ Methodenreport, Institute for Employment Research (IAB), Nuremberg
- Heckman J (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161
- Heckman J (1990) Varieties of selection bias. *Am Econ Rev* 80(2):313
- Huber M, Melly B (2015) A test of the conditional independence assumption in sample selection models. *J Appl Econom* 30(7):1144–1168
- IAB (2017) IAB Labor Market Report. Institute for Employment Research (IAB), Nuremberg, 10/2017

- Klein R, Spady R (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61(2):387–421
- Lemieux T (2006) Increasing residual wage inequality: composition effects, noisy data, or rising demand for skill? *Am Econ Rev* 96(3):461–498
- Machado C (2017) Unobserved selection heterogeneity and the gender wage gap. *J Appl Econom* 32(7):1348–1366
- Machado JAF, Mata J (2005) Counterfactual decomposition of changes in wage distributions using quantile regression. *J Appl Econom* 20(4):445–465
- Melly B (2006) Public and private sector wage distributions controlling for endogenous sector choice. University of St. Gallen, Discussion Paper
- Möller J (2016) Lohnungleichheit - Gibt es eine Trendwende? *Wirtschaftsdienst* 96(1):38–44
- Paul M (2016) Is there a causal effect of working part-time on current and future wages? *Scand J Econ* 118:494–523
- Picchio M, Mussida C (2011) Gender wage gap: a semi-parametric approach with sample selection correction. *Labour Econ* 18:564–578
- Vella F (1998) Estimating models with sample selection bias: a survey. *J Human Resour* 127–169
- Vom Berge P, Burghardt A, Trenkle S (2013) Sample of integrated labour market biographies regional file 1975–2010 (siab-r 7510). Technical report, FDZ data report, 09/2013, Institute for Employment Research (IAB), Nuremberg

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.