**FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG**

Lehrstuhl für VWL, insbes. Arbeitsmarkt- und Regionalpolitik
Professor Dr. Claus Schnabel

**Diskussionspapiere**
**Discussion Papers**

NO. 111

# The Gender Pay Gap Revisited: Does Machine Learning offer New Insights?

STEPHANIE BRIEL AND MARINA TÖPFER

JANUARY 2020

ISSN 1615-5831

Editor: Prof. Dr. Claus Schnabel, Friedrich-Alexander-Universität Erlangen-Nürnberg

# The Gender Pay Gap Revisited: Does Machine Learning offer New Insights?*

Stephanie Briel[1] and Marina Töpfer[2♯]

[1] University of Hohenheim
[2] University of Erlangen-Nürnberg

**Abstract:** This paper analyses gender differences in pay at the mean as well as along the wage distribution. Using data from the German Socio-Economic Panel, we estimate the adjusted gender pay gap applying a machine learning method (post-double-LASSO procedure). Comparing results from this method to conventional models in the literature, we find that the size of the adjusted pay gap differs substantially depending on the approach used. The main reason is that the machine learning approach selects numerous interactions and second-order polynomials as well as different sets of covariates at various points of the wage distribution. This insight suggests that more flexible specifications are needed to estimate gender differences in pay more appropriately. We further show that estimates of all models are robust to remaining selection on unobservables.

**Zusammenfassung:** Dieser Beitrag untersucht die geschlechtsspezifische Lohnlücke am Mittelwert und entlang der Lohnverteilung. Für unsere Analyse nutzen wir Daten des Sozio-ökonomischen Panels. Wir schätzen die bereinigte Lohnlücke zwischen Männern und Frauen unter Verwendung einer Methode des Maschinellem Lernens (post-double-LASSO Ansatz). Die mit dieser Methode geschätzten bereinigten Lohnlücken unterscheiden sich substantiell von den Ergebnissen konventioneller Ansätze. Hauptgrund für diese Unterschiede ist, dass der Ansatz des Maschinellen Lernens eine Vielzahl von Interaktionen und Polynomen zweiter Ordnung sowie unterschiedliche Kontrollvariablen entlang der Lohnverteilung wählt. Dieses Ergebnis deutet daraufhin, dass flexiblere Modelspezifikationen benötigt werden, um die Geschlechterlohnlücke adäquat zu schätzen. Wir zeigen auch, dass die Schätzergebnisse robust gegenüber Selektion aufgrund unbeobachtbarer Merkmale sind.

**Keywords: Gender pay gap; Machine Learning; Selection on unobservables**
**JEL: J7, J16, J31**

---

* We thank Martin Biewen, Aderonke Osikominu, Winfried Pohlmeier, Mark Schelker, Claus Schnabel, Anthony Strittmatter and participants of the Annual Conference of the German Economic Association 2019 in Leipzig, the IAAE 2019 in Nicosia, the IIPF 2019 in Glasgow, the SSES Annual Congress 2019 in Geneva as well as participants of the Workshop on Causal Machine Learning 2020 in St. Gallen.

♯ Correspondence to: University of Erlangen-Nürnberg, School of Business and Economics, Chair of Labor and Regional Economics, Lange Gasse 20, 90403 Nuremberg. E-mail: marina.toepfer@fau.de.

# 1 Introduction

The Gender Pay Gap (GPG) persists worldwide despite political emphasis to close it (see for example Blau and Kahn, 2017; Goldin, 2014). In the European Union men's hourly wages are 16% higher than women's (Eurostat, 2018). The German wage gap lies with more than 20% well above the EU average (Eurostat, 2018). When analyzing the GPG, researchers typically use regression models that are based on Mincer-type wage equations controlling for individual, job or firm characteristics. Yet, there is an ongoing debate about the pivotal control variables for the estimation of the GPG (see Blau and Kahn, 2017). For identification of the 'true' GPG, the conditional independence assumption or unconfoundness has to hold. It is therefore necessary to control for all factors that are simultaneously correlated with wages and gender (see e.g. Fortin et al., 2011).

In this paper, we estimate the adjusted GPG, i.e. the pay penalty of being female, conditional on a set of covariates selected with a machine learning approach. We compare estimation results from conventional model specifications with model specifications based on variable selection performed by a machine learning algorithm.[1] In case of the conventional models, we estimate the adjusted gap using sets of covariates proposed by the literature. The analysis is conducted at the mean (using OLS) as well as at selected percentiles (using linear unconditional quantile regression or Recentered Influence Function (RIF)-OLS). For machine learning based model selection, we rely on post-double Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Belloni et al. (2014a,b).[2] Using this type of model selection ensures a model specification that controls for both major gender differences and important wage predictors. That is, it immunizes – at least theoretically – against omitted variable bias.[3] As we use unconditional quantile regression for estimations beyond the mean, we run a separate LASSO for each considered RIF. Besides using a fully unrestricted LASSO specification, we consider also a restricted LASSO specification that does not allow shrinkage of parameters of covariates that are considered pivotal by the literature. For example, following Mincer (1974), we do not allow for shrinkage of education or the second-order polynomial of full-time labor market experience in the restricted LASSO specification.

Next, we use the method proposed by Oster (2019) to evaluate how robust the estimated gender gaps are to remaining selection on unobservables. This method allows us

[1] We use supervised machine learning throughout the paper. In case of supervised machine learning, we have both an outcome variable and explanatory variables (inputs). The aim is to predict or estimate the outcome based on the explanatory variables. In case of unsupervised machine learning, there are only inputs and focus lies on identifying relations and structures of the data (see e.g. James et al., 2013, for details).

[2] We rely on the post-double-LASSO estimator as in a following step, we use the method of Oster (2019) to compare different model specifications. For this method it is necessary to rely on a linear model (see Oster, 2019).

[3] The latter holds, as long as the data set used contains all important variables, if not, it nevertheless leads to an estimate of the GPG that minimizes omitted variable bias given data restrictions.

to compare the data-driven and the conventional model specifications. Finally, we use standard decomposition techniques (Blinder, 1973; Oaxaca, 1973) to analyze the impact of the selected sets of controls on the raw gap as well as on its components (explained and unexplained). For estimation, we use the German Socio-Economic Panel Study (SOEP) providing a large nationally representative sample with detailed information on individual human capital and labor market characteristics.

Blau and Kahn (2017) provide an overview of what has been learned about gender differences in pay. In many studies, the observed (raw) GPGs are separated into an explained and unexplained part using Oaxaca-Blinder decompositions or further developments of this method. While the explained part stems from differences in the control variables, the unexplained part is related to gender differences in coefficients. Since the unexplained part reflects differences in pay of individuals with identical observable characteristics, some authors claim it to reflect discrimination (Goldin, 2014). Typically economists define labor-market discrimination as "pay differences that are not accounted for by productivity differences" (Blau and Kahn, 2017, p.830). Productivity differences either arise through human capital or other supply-side sources. Examples of general unobservable productivity are individual ability or motivation (Blau and Kahn, 2006).[4] In order to produce unbiased estimates of the components and the coefficient of the female dummy in an OLS regression, it is important to control for all relevant factors that are simultaneously correlated with wages and gender.

Key characteristics used as control variables for estimation of GPGs are observable labor market and human capital variables such as job market tenure, labor market experience, educational attainment as well as industrial and occupational controls (Blau and Kahn, 2017; Goldin, 2006; Mandel and Semyonov, 2014). Further, individual characteristics such as family background (Bailey et al., 2012; Fortin, 2008) and union status (Heinze and Wolf, 2010) as well as firm-specific characteristics like firm-size or presence of a works council may be important (Heinze and Wolf, 2010). The part of the GPG attributable to conventional human capital controls has decreased over time and conventional human capital controls are found to explain only little of today's GPG (Blau and Kahn, 2017). The largest part of the gap is driven by gender differences in sorting into occupations and industries (Blau and Kahn, 2017). Studies using models with conventional control variables usually reveal substantial unexplained parts of the GPG. As mentioned before, if gender is endogenous in the regression model (due to omitted controls), the estimates are biased. With the aim of decreasing omitted variables bias, a more recent part of the

[4]In contrast, the concept of statistical discrimination describes situations where, given that employers are uncertain about worker productivity or stability, they discriminate against minority groups like women based on real or perceived average differences (Blau and Kahn, 2017). While in case of statistical discrimination, discrimination is based on (more or less) rational beliefs about the average differences in characteristics between groups, Becker (1957)'s taste-based discrimination describes situations where discrimination is likely to be driven by prejudice (Coffman et al., 2018).

literature considers alternative driving forces behind gender differences in pay.

Some studies use character skills as important controls when estimating the GPG (Brenzel and Laible, 2016; Cattan, 2013; Fortin, 2008; Risse et al., 2018). Since character skills are found to be key determinants of social and economic success (Almlund et al., 2011) and are characterized by large gender-specific differences (Bertrand, 2011; Blau and Kahn, 2017; Weisberg et al., 2011), they are promising candidates for explaining the pay gap between men and women. In other words, as character skills may represent a source of unobserved human capital in many studies, they are potential omitted controls. The main part of the related literature focuses on the Big Five Personality Traits – openness, conscientiousness, extraversion, agreeableness,and neuroticism– as measures for character skills (Brenzel and Laible, 2016; Mueller and Plug, 2006). Some additionally control for locus of control (Nyhus and Pons, 2012), reciprocity (Heineck and Anger, 2010) or willingness to take a risk (Collischon, 2017). Mueller and Plug (2006) and Heineck and Anger (2010) find gender-specific differences in the returns to character skills, while most studies suggest that character skills are found to explain only a moderate part of the GPG (Blau and Kahn, 2017). Fortin (2008) shows that character skills have a higher impact on the GPG than traditional educational factors and are almost as important as experience or tenure. Women's higher agreeableness is found to widen the GPG (Braakmann, 2009; Nyhus and Pons, 2012; Risse et al., 2018), while conscientiousness narrows it (Risse et al., 2018). Other potentially relevant control variables proposed by the literature are: test scores (Blau and Kahn, 2017), motherhood (Juhn and McCue, 2017), gender difference in preferences for more flexible work arrangements (Goldin, 2014; Mas and Pallais, 2017), gender differences in the probability of leaving the job as well as the slow down in female promotions (Goldin, 2014).

Another point highlighted by the literature is that effects may not be homogeneous across different subgroups of the population. Focusing on the wage consequences of motherhood, Wilde et al. (2010) show that the wage gap associated with children is higher for high-skilled women. Juhn and McCue (2017) find changes in the marriage pay gap over the time being married. Gensowski (2018) shows that the effects of character skills differ between educational groups. Further, an employer has, in a first step, to learn about the character skills of the employees in order to reward them (employer learning). The latter implies that it is important to control for interactions between tenure with the current employer and the character skill measures (Heineck and Anger, 2010; Nyhus and Pons, 2012). Despite different model specifications, all of the mentioned studies end-up with a substantial residual part of the GPG. Overall, there is a multitude of potential drivers of gender differences in pay that may not be equally relevant at different points of the wage distribution. It is up to the researcher which of the (many) possible control variables to include for estimation of the GPG. So far, researchers generally use the same set of

variables across the wage distribution. In order to make selection of the variables not arbitrary but systematic, we propose to use supervised machine learning.

Supervised machine learning is usually used to make predictions about the outcome of interest given a set of potential controls (James et al., 2013). In economic applications, like the estimation of GPGs, focus lies on producing reliable parameter estimates rather than on making precise predictions. Although machine learning techniques, like LASSO, deliver coefficient estimates, the associated models hardly fulfill the usual properties of consistency and unbiasedness (Mullainathan and Spiess, 2017). At a first glance, machine learning techniques may therefore seem inappropriate for economic applications when causal inference is of main interest. We obtain interpretable and reliable estimates of gender differences in pay using the post-double-LASSO estimator, as Belloni et al. (2014b) show that the double-selection procedure leads to valid inference about a coefficient of interest (female dummy in our case). This result holds even under selection mistakes and the authors provide (substantial) simulation evidence that the procedure works across a wide variety of models. The post-double-LASSO estimator relies on two prediction steps, one for the outcome variable (log wages in our case) and one for the explanatory variable of main interest (female dummy in our case). case). After having collected a list of covariates based on economic reasoning, machine learning is used for model selection among the potential control variables. The underlying algorithm compares all possible model specifications and selects the criterion maximizing one (Athey, 2018). Using two selection steps also enhances efficiency by finding variables that are strongly predictive of the outcome and may remove residual variance (Belloni et al., 2014b).

The idea of this paper is to use machine learning to find the most relevant control variables among an extensive set of potential controls (more than 5,000 in our case). In addition, the data-driven approach allows to learn more about the functional form of the selected variables (Belloni et al., 2014a). Further, by running separate outcome LASSOs at different points of the wage distribution, we allow for different sets of control variables along the wage distribution.

This paper uses the post-double-LASSO estimator for estimation of the GPG. Closest related to this study is the work of Bach et al. (2018) focusing on heterogeneity in the US GPG using double-LASSO. Compared to our paper, they introduce heterogeneity not in terms of relations between different control variables but directly through multiple interactions with the female dummy. This approach yields individual-specific GPGs depending on the observed values of the control variables interacted with the female dummy. The main interest of their work lies in the distribution of the individual-specific GPGs, while we are interested in model selection and performance of the post-double-LASSO estimator compared to conventional models.

We find substantial differences in the adjusted GPGs at different points of the wage

distribution. In particular, depending on the model specification, the gap varies substantially at the bottom and top, while it is relatively stable at the mean and median of the wage distribution. The latter may be attributable to the fact that standard (augmented) Mincer-type wage models have been developed for estimation at the mean. In line with this, our results suggest that different sets of covariates should be used at different points of the wage distribution. Further, the selected models are more flexible than conventional models for estimation of GPGs, i.e. they contain numerous interactions and higher-order polynomials. While the model specifications based on the recent literature find evidence for glass ceiling and an increasing GPG along the distribution, models based on machine learning deliver only slight evidence for glass ceiling and suggest a U-shaped GPG along the distribution. We find that all estimates of the adjusted GPG are robust to remaining selection on unobservables. Decomposition results show that the data-driven models explain (substantially) larger shares of the GPG at the middle and top of the wage distribution. Gender differences in labor market characteristics such as experience and job tenure are main drivers of the gap.

The paper is organized as follows. Section 2 describes our estimation strategy. Section 3 presents the data set used for the empirical analysis. Empirical results are given in Section 4 and Section 5 repeats the estimation of the adjusted pay gap using a sample split. Finally, Section 6 concludes.

# 2 Estimation Method

In this Section, we outline the different estimation methods applied. We start with the standard OLS wage model. Next, we describe the machine learning approach applied for model selection and the method of Oster (2019). The latter is used in order to learn more about the consequences of remaining selection on unobservables in the conventional and data-driven model specifications. Finally, we shortly present the decomposition approach and the RIF-OLS or unconditional quantile regression model.

## 2.1 Standard Approach

For estimation of the adjusted GPG, we assume that the wage equation of individual $i$ with $i = 1, \ldots, N$ has the following linear form:

$$y_i = \beta \; d_i + x_i\gamma + u_i. \tag{1}$$

where $y_i$ is the log of hourly wages, $d_i$ is a gender dummy and equals one if the individual is female and zero otherwise. The $1 \times K$ vector $x_i$ contains the set of control variables plus a constant and $u_i$ is the error term. The coefficient of $d_i$ gives the adjusted GPG. For

unbiased estimation of the adjusted GPG[5] ($\hat{\beta}$), the conditional independence assumption has to hold. The latter implies that the set of control variables ($x_i$) has to contain all variables that are simultaneously correlated with $d_i$ and $y_i$, i.e. $\mathbb{E}(u_i|d_i, x_i) = 0$.

We compare various specifications of equation (1) that differ in the elements of the vector $x_i$. In a first step, we estimate a short and three different conventional wage models in line with the literature. In case of the short regression model, $x_i$ contains only a constant. Thus, the resulting estimate of the gap ($\hat{\beta}$) is the raw GPG. For the three conventional model specifications, the set of control variables $x_i$ varies across specifications and is gradually augmented. The richest conventional model specification contains, besides a constant, standard Mincer-type and character skill controls (see Section 3 for details). By construction, these model specifications are similar to model specifications used in the (recent) literature and are supposed to represent the conventional estimation approach, i.e. Mincer-type wage models estimated by OLS.

Concerning model specification, a number of questions arises, for instance with respect to the inclusion of character skills in the model. In most settings (if used) the complete available set of character skill controls is included linearly. However, it may not be appropriate to include the entire set of character skills as elements of $x_i$ linearly. Recent research considers two alternative model specifications. First, nonlinear functions of the character skill measures may be more adequate than a linear function (e.g. Heineck and Anger, 2010; Mueller and Plug, 2006). Second, the employer has first to learn about the character skills of the employee in order to reward him or her (employer learning). The latter implies the inclusion of interactions between tenure with the current employer and character skill measures as parts of $x_i$ (Heineck and Anger, 2010; Nyhus and Pons, 2012). Similar considerations may be reasonable for other control variables as well. Therefore, it may be appropriate to consider more flexible model specifications than usual for estimation of the (adjusted) GPG. That is, it may be important to include a variety of potential interactions and higher-order polynomials. Considering all possible two-way interactions and higher-order polynomials yields a high-dimensional set of potential controls. As it may be difficult for the researcher to decide which controls to use and in which functional form, we exploit a machine learning technique for model selection.

## 2.2 Model Selection using Machine Learning: The Post-Double-LASSO Estimator

In the following, we describe the post-double-LASSO estimator proposed by Belloni et al. (2014a,b) that we apply for model selection. The machine learning procedure uses the data at hand to identify pivotal control variables for the estimation of the GPG. Relaxing

[5] We refer to the adjusted GPG as GPG estimated conditional on control variables $x_i$.

the linearity assumption of equation (1), the partially linear wage model is given by:

$$y_i = \beta \; d_i + m(z_i) + \zeta_i. \tag{2}$$

Further, it holds for $d_i$, that

$$d_i = g(z_i) + \nu_i \tag{3}$$

where $z_i$ corresponds to the vector of control variables that need to be taken into account. $\zeta_i$ and $\nu_i$ are the respective error terms. It holds that $\mathbb{E}[\zeta_i|d_i, z_i] = 0$ and $\mathbb{E}[\nu_i|z_i = 0]$, with the functions $m(\cdot)$ and $g(\cdot)$ not being necessarily linear. Linear combinations of the control variables, i.e. interactions and higher-order polynomials, with $x_i = P(z_i)$, are used for their approximation, i.e.:

$$m(z_i) = x_{im}\gamma_m + r_{mi} \tag{4}$$

$$g(z_i) = x_{ig}\gamma_g + r_{gi} \tag{5}$$

where $x_{im}\gamma_m$ and $x_{ig}\gamma g$ are the linear approximations and $r_{im}$ and $r_{ig}$ are the corresponding approximation errors. As stated before, in the case of the conventional models, $x_i$ contains commonly used control variables. Based on model selection with the double-LASSO, the elements of $x_i$ are selected among the set of potential controls that contains, additionally to the variables commonly used, further controls as well as higher-order polynomials and two-way interactions of these controls (more details are given in Section 3). Theoretically, the set of potential controls can be huge. We denote the number elements in the set of potential controls by $P$.

As the proposed model selection procedure uses the LASSO, we have to assume approximate sparsity. This assumption implies that among the set of potential controls only a relatively small number of elements is needed for estimation of the model. The literature offers several statistical methods for model selection in case of approximately sparse regression models. Usually, these methods aim at predicting outcomes (Hastie et al., 2009). One of the most popular approaches for model selection in case of approximately sparse regression models is the LASSO introduced by Frank and Friedman (1993) and Tibshirani (1996). The LASSO chooses coefficients to minimize the sum of squared residuals plus an additional penalty term. The sum of absolute coefficients is included in the minimization problem in order to penalize the size of the model. Thus, the LASSO shrinks the regression coefficients by assigning a penalty to the sum of the absolute values of all coefficients (Hastie et al., 2009). In the following, we rely on a slightly modified version of the LASSO

that was first introduced by Belloni et al. (2012). This procedure defines the LASSO as:

$$\widehat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{P+1} x_{ij}\gamma_j)^2 + \underbrace{\lambda \sum_{j=2}^{P+1} |\gamma_j|\eta_j}_{\text{penalty term}} \tag{6}$$

where $P$ potential controls are considered, $\eta_j$ represents so-called 'penalty loadings' and the degree of shrinkage is reflected by the tuning parameter $\lambda$. The larger $\lambda$, the higher the degree of shrinkage, i.e. the more coefficients will be set equal to zero. In contrast to the standard LASSO, the absolute value of each coefficient is additionally multiplied by its penalty loading $\eta_j$.[6]

The inclusion of $\eta_j$ is associated with two main advantages over the standard LASSO. First, it ensures equivariance of the estimated coefficients to rescaling of the variables in the set of potential controls. Second, it allows to address non-normality in model errors, heteroskedasticity as well as clustering (Belloni et al., 2012). That is, the penalty loadings introduce self-normalization. Further, they are data-dependent and defined via an iterative procedure (see Belloni et al., 2012, 2014b, for more details). The penalty term in equation (6) has a kink at zero. Thus, it results in a sparse estimator with many coefficients set exactly to zero. Moreover, the LASSO is a convex optimization problem and highly efficient computational algorithms are available for its solution (Belloni et al., 2014a). If prediction of the outcome is of main interest, $\lambda$ is often chosen by cross-validation. However, as shown by Belloni et al. (2012), there are superior methods for the choice of $\lambda$ if prediction is not the main goal. In this paper, we rely on the so-called feasible LASSO where $\lambda$ is chosen dependent on the sample size and the number of potential controls. To be precise, $\lambda$ is a rescaled critical value and has the form: $\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \alpha/2P)$ where $c > 1$ is a constant and $(1 - \alpha)$ reflects a confidence level. As proposed by Belloni et al. (2014a,b), we use $c = 1.1$ and $\alpha = min(1/n, 0.05)$ (see Belloni et al., 2014b, for details).

The LASSO - in the current setting - is associated with two main challenges. First, the non-zero coefficients that are part of its solution are substantially biased towards zero. Second, the LASSO aims at prediction of outcome variables and not at drawing inference on model parameters. We can address the first concern by employing the post-LASSO estimator proposed by Belloni et al. (2012). Post-LASSO estimation proceeds in two steps. In a first step, the LASSO is used to determine important predictors for the outcome. Next, the coefficients of all covariates in the set of selected control variables (those variables with nonzero coefficients in the first step) are estimated via OLS.[7]

The second concern is more challenging in our context, i.e. that the LASSO aims at prediction and not at causal inference. Estimation of the GPG in the standard con-

[6] Please note that the intercept (j=1) is not included in the penalty term

[7] Belloni and Chernozhukov (2013) and Belloni et al. (2012) have shown that the post-LASSO is equal to or better than the LASSO in terms of rates of convergence and bias.

text (equation (1)) is based on a conditional-on-observables identification strategy. This strategy implies that it is important to control for all factors (within the set of potential controls) that are simultaneously correlated with $d_i$ (gender differences) and $y_i$ (wage predictors). A rather intuitive approach would be to apply the LASSO in equation (6) to the wage equation given in (1) and force $d_i$ to remain in the model by excluding its coefficient from the penalty term. This procedure, however, would (most likely) result in a set of selected controls that does not include control variables that are highly correlated with gender ($d_i$). Control variables highly correlated with gender would be dropped as their addition does not add much explanatory power given that the model already controls for gender. However, omitting variables that are strongly correlated with gender may lead to omitted variable bias in the estimation of the adjusted GPG (or $\hat{\beta}$) if these variables are (at least slightly) correlated with wages.

Hypothetically, we can rely on two different equations. We can use the LASSO for prediction of $y_i$ given the set of potential control variables (equation (4)) or prediction of $d_i$ given the set of potential controls (equation (5)). Relying on one equation is only sensible if there are no errors in variable selection. Applying the LASSO to the first regression ($y_i$ on $x_i$), variables with large coefficients in this specific equation are part of the set of selected controls. As the LASSO tends to miss variables with coefficients of only moderate size and if the neglected variables have large coefficients in the second regression ($d_i$ on $x_i$), our results will suffer from omitted variable bias (Belloni et al., 2014a). That is, when estimating the GPG and only including control variables selected by the LASSO for $y_i$ on $x_i$, the estimated GPG is likely to be inconsistent and biased.

Because of the potential drawbacks of model selection based on a single equation, it is important to consider a prediction for both $y_i$ (log hourly wages) and $d_i$ (gender). The corresponding procedure is called post-double-LASSO estimator and works in the following way: First, the LASSO is used to select the set of controls that are important to predict $y_i$. We call this the Wage-LASSO. The Wage-LASSO results in the set of selected wage controls $x_{im}^S$, where the index $m$ identifies the wage equation (4) and $S$ the set of selected controls. That is, $x_{im}^S$, includes all variables with non-zero coefficients in the Wage-LASSO. Second, the LASSO is used to select the set of controls that is relevant to predict $d_i$. We refer to this as the Gender-LASSO. The Gender-LASSO results in the set of selected gender controls $x_{ig}^S$ that contains all variables with non-zero coefficients in the Gender-LASSO. The indices $g$ and $S$ refer to the gender equation (5) and the set of selected controls, respectively. Finally, the adjusted GPG is estimated by OLS with the union of the two sets of selected controls, i.e. $x_i^S = x_{ig}^S + x_{im}^S$. This procedure ensures that any variables that have large effects in either of the two regressions are included in the model (Belloni et al., 2014b).

The post-double-LASSO estimator proposed by Belloni et al. (2014a,b) guards against

omitted variable bias of the estimated GPG by immunization against the above mentioned model-selection errors (Belloni and Chernozhukov, 2013). While the use of the Gender-LASSO ensures that we control for all important gender differences, the Wage-LASSO delivers important wage predictors and helps keeping the residual variance small. To sum up, the two selection steps of the post-double-LASSO increase the likelihood of orthogonality in estimation of the GPG compared to the conventional models. The post-double-LASSO helps to reduce the omitted variable bias and enables performance of uniform inference (for the adjusted GPG or the coefficient of $d_i$) after model selection, as long as the regression models of both selection steps are approximately sparse. Technically approximate sparsity requires $s^2 \ll N$. This rather strong assumption can be relaxed using sample splitting (see Section 5 for details on sample splitting).

If some variables are considered to be indispensable for estimation of the (adjusted) GPG, one may exclude the corresponding coefficients from the penalty term in equation (6). Such a set of variables is called amelioration set. In case of an non-empty amelioration set, model selection is restricted to a smaller set of potential controls ($P$ minus the amelioration set). The latter represents a 'restricted' LASSO, one does not allow for shrinkage of the coefficients of the controls in the amelioration set. A further regularity condition for the use of the post-double-LASSO estimator is that the number of variables in the amelioration set is allowed to be maximally as high as the number of elements in the larger of the two sets of selected controls $x_{ig}^S$ and $x_{im}^S$ (Belloni et al., 2014b). As one may argue that some controls of the standard Mincer-type wage model necessarily need to be included in the estimation of the conditional GPG, we estimate besides a fully flexible unrestricted LASSO also a more conservative restricted LASSO.

Note that the post-double-LASSO estimator ensures valid inference only with respect to the coefficient of the explanatory variable of main interest – the adjusted GPG in this paper (Belloni et al., 2014a,b). All other included control variables may be endogenous. Moreover, model selection is not stable and depends on the sample at hand (Mullainathan and Spiess, 2017).

### 2.3 Remaining Selection on Unobservables

The post-double-LASSO estimator provides model specifications that control for both important wage determinants and gender differences. The question is whether GPGs based on these models are less prone to omitted variable bias than conventionally estimated GPGs. In our setting, remaining selection on unobservables represents a major (potential) source for omitted variable bias.

To shed more light on this source of bias, we apply the method proposed by Oster (2019) representing a refinement of the method of Altonji et al. (2005). The method takes coefficient stability and changes in the degree of explained outcome variation when switch-

ing from the short (i.e. the model giving the raw GPG) to the intermediate regression model (i.e. the conventional or LASSO models) into account. We use this method to calculate the ratio of remaining selection on unobservables relative to selection on observables that is necessary to produce a zero GPG. High values of this ratio are usually interpreted as an indication that the model is less prone to omitted variable bias. In this paper, we consider model specifications that differ in their degree of selection on observables. If machine learning based model selection results in sets of control variables that control more appropriately for important confounders, the denominator of the ratio is higher than for conventional models, while the numerator is lower or at least not higher. Therefore, we assume that the ratio in case of the machine learning models is lower than for the conventional models.

In case of the short regression model, no control variables are included in equation (1), i.e. $d_i$ is apart from a constant the only right-hand-side variable. The intermediate regression model is the regression model with a non-empty set of control variables $x_i$ (plus a constant). In our setting, we compare the performance of five different intermediate models with respect to robustness against remaining selection on unobservables. That is, we consider five different sets of control variables, $x_i$. The intermediate models are the three conventional and the restricted and unrestricted LASSO models. We refer to the GPG in the short (intermediate) regression model as $\beta_{short}$ ($\beta_{inter}$). $R^2_{short}$ ($R^2_{inter}$) gives the $R^2$ corresponding to the short (intermediate) regression model.

The hypothetical full regression model represents the model that contains all factors (observables and unobservables) that are simultaneously correlated with gender and wages. The corresponding unbiased true GPG is given by $\beta_{full}$. Proportional selection, with proportionality parameter $\delta$, is the key assumption in this framework. This assumption implies that the degree of remaining selection on unobservables can be expressed as the degree of selection on observables multiplied by the proportionality parameter $\delta$. For instance, $\delta = 2$ implies that remaining selection on unobservables is two times as high as selection on observables. Relying on the additional assumption that the relative contribution of each variable in $x_i$ to $d_i$ is the same as their contribution to $y_i$ (equal relative contribution), the (easy-to-calculate) approximation of a bias-adjusted GPG ($\beta^*$) reads as:

$$\beta^* = \beta_{inter} - \delta[\beta_{short} - \beta_{inter}]\frac{R^2_{full} - R^2_{inter}}{R^2_{inter} - R^2_{short}} \tag{7}$$

The approximation is based on the coefficient movement from the short to the intermediate model ($\beta_{short} - \beta_{inter}$), the proportionality parameter ($\delta$) and the ratio of the shift in explained outcome variation when switching from the intermediate to the full model ($R^2_{full} - R^2_{inter}$) and the switch from the short to the intermediate model ($R^2_{inter} - R^2_{short}$). Equation (7) gives an adjusted version of the (adjusted) GPG from the intermediate re-

gression model ($\beta_{inter}$). To be precise, the gap is adjusted by the coefficient movement scaled by the proportionality parameter and the movements in explained outcome variation. Apart from $\delta$ and $R^2_{full}$ all terms in equation (7) are observable in standard regression outputs. Usually it is assumed that $\delta = 1$, implying that selection on unobservables is as high as selection on observables (following Altonji et al. (2005)). The most conservative assumption for $R^2_{full}$ is $R^2_{full} = 1$. However, as this assumption may be too restrictive and as we use RIF-OLS for the estimation beyond the mean, where the dependent variable is binary, assuming $R^2_{full} = 1$ is not the appropriate choice. Therefore, we follow Oster (2019) and assume $R^2_{full} = 1.3 \times R^2_{inter}$.[8]

In this paper, we do not focus on the bias-adjusted version of the GPG but are interested in how large remaining selection on unobservables (relative to selection on observables) has to be in order to produce a zero GPG. Thus, instead of calculating a $\beta^*$ based on assumptions about $\delta$, we search for the value of $\delta$ that, given assumptions on $R^2_{full}$, results in a zero bias-adjusted GPG ($\beta^* = 0$) and refer to this as $\delta^*$. Setting $\beta^* = 0$ and solving equation (7) for $\delta^*$ gives the following (easy-to-calculate) approximation of the proportionality parameter resulting in a zero GPG:

$$\delta^* = \frac{\beta_{inter}}{[\beta_{short} - \beta_{inter}]} \times \frac{R^2_{inter} - R^2_{short}}{R^2_{full} - R^2_{inter}} \tag{8}$$

As the assumption of equal relative contribution is rather restrictive, we rely on the less restrictive version of (7) proposed by Oster (2019) and use the $\delta^*$ derived in this version. Given the intuition behind the restrictive (equation (7)) and the less restrictive version is the same, we skip the formal representation of the much more complicated less restrictive expressions for $\beta^*$ and $\delta^*$ (see Oster, 2019, for details). Instead of assuming equal relative contributions, we assume that the bias arising from remaining selection on unobservables does not result in a change of the sign of correlation between gender and the index of observables.

In general, $\delta^*$ is interpreted as $\delta^* = \frac{\text{Degree Remaining of Selection on Unobservables}}{\text{Degree of Selection on Observables}}$, i.e. the ratio of the remaining degree of selection on unobservables and selection on observables necessary to result in a zero GPG. For $\delta^* \geq 1$ the estimated GPG is considered to be robust to remaining selection on unobservables (Oster, 2019).

In our case, large values of the ratio can occur for two reasons. First, it may be the case that high remaining selection on unobservables is necessary to result in a zero GPG (high values in the numerator). In this case higher values of $\delta^*$ would reflect higher robustness to remaining selection on unobservables. Second, $\delta^*$ may be high due to small selection on observables (low values in the denominator). In the latter case, higher values

[8] Despite restrictive assumptions, Oster (2019) shows that (7) is an easy-to-calculate and reasonable approximation for a more precise and less restrictive version of $\beta^*$ (see Oster, 2019, for details).

of $\delta^*$ would only reflect inferior assessment of selection on observables. If machine learning based model selection works as expected and results in a set of control variables that contains all important wages predictors and gender differences selection on observables is higher (or at least not lower) than in case of the conventional models. As long as the set of potential controls contains all variables included in the conventional models, remaining selection on unobservables is unlikely to be higher for the data-driven models compared to the conventional ones. Consequently, we expect lower values for $\delta^*$ in case of the two post-double-LASSO estimates than for the conventional models.

### 2.4 Decomposition

After having obtained the set of selected controls by the double-LASSO procedure (Section 2.2) and investigated robustness with respect to remaining selection on unobservables (Section 2.3), we estimate the wage equations separately for men and women and decompose the raw or unadjusted GPG following Blinder (1973) and Oaxaca (1973). The raw gap is, as in the standard two-fold Oaxaca-Blinder decomposition, decomposed in an endowments and a coefficients component using men as the non-discriminatory wage structure:

$$\bar{Y}_M - \bar{Y}_F = \underbrace{(\bar{X}_M - \bar{X}_F)\hat{\beta}_M}_{\text{Endowments Effect}} + \underbrace{\bar{X}_F(\hat{\beta}_M - \hat{\beta}_F)}_{\text{Coefficients Effects}} \tag{9}$$

where $\bar{Y}_G$ is the average log hourly wage of $G$, with $G = M, F$ and where $M$ identifies men and $F$ women. Analogously, $\hat{\beta}_M$ and $\hat{\beta}_F$ are the male and female coefficient vectors, respectively. The endowments effect is often referred to as 'explained' and the coefficients effect as 'unexplained' component of the gap. As we are interested in understanding differences between conventional and data-driven model outcomes, we conduct a detailed decomposition of equation (9). In detailed decompositions, the contribution of each covariate or subset of covariates to the raw gap is estimated.

Oaxaca-Blinder decompositions suffer from two main problems. First, the decomposition is not unique, i.e. the estimated components may change depending of the choice of the non-discriminatory wage structure (men or women in our case, see e.g. Neumark, 1988; Oaxaca and Ransom, 1994). Therefore, we also report the decomposition results using women as non-discriminatory wage structure (see Appendix C). Second, the assignment of categorical variables to the unexplained part of the GPG is not invariant to the choice of the left-out category (Oaxaca and Ransom, 1999). This problem can be solved by imposing a zero-sum restriction on the coefficients of the single categories and by re-expressing the effects as deviations from the grand mean. This procedure is referred to as deviation contrast transformation (Gardeazabal and Ugidos, 2004; Yun, 2005). As our data-driven model specifications are likely to include a huge number of interactions among

different types of categorical variables, it may be hard (or even impossible) to classify the different interacted categorical variables into meaningful groups. Thus, (for simplicity) we do not represent the corresponding coefficient estimates as deviations from the grand mean. As changing the left-out category leads to a change in the quantitative contribution of a dummy variable to the coefficients effect only (Gardeazabal and Ugidos, 2004), we will rely exclusively on the endowments effect (explained part) in the detailed decomposition. Oaxaca and Ransom (1999) show that even if the contribution of individual dummy variables to the wage decomposition is not identified, the aggregate unexplained component is identified.[9] The latter implies that we can rely on the aggregate unexplained component.

Despite the choice of the appropriate set of control variables for estimation of the GPG, there may also be differences in the GPGs across the wage distribution (Albrecht et al., 2003; Arulampalam et al., 2007). We therefore extend the analysis beyond the mean using unconditional quantile regressions. Unconditional quantile regressions allow estimation with OLS. The latter, we exploit inter alia for the method of Oster (2019). For estimation along the wage distribution, we run every step of our estimation procedure at each of the selected percentiles allowing for quantile-specific selected sets of controls as well as quantile-specific coefficient (Firpo et al., 2009) and decomposition estimates (Fortin et al., 2011).

### 2.5 Unconditional Quantile Regression: RIF-OLS

In matrix notation, the standard wage model (represented in equation (1)) has the following form:

$$Y = D\beta + X\gamma + u \tag{10}$$

where $Y$ and $D$ are a $N \times 1$ vectors of log hourly wages and gender dummies, respectively, X is a $N \times K$ matrix of observable characteristics and $u$ is a $N \times 1$ vector of disturbances. We use the unconditional quantile regression model in order to estimate the effect of explanatory variables, $X$, and the gender dummy, $D$, on the unconditional quantile, $Q_\tau$, of the log hourly wage $Y$ (Firpo et al., 2009).

In unconditional quantile regressions, the dependent variable is the Recentered Influence Function (RIF) that is a transformation of $Y$ such that its expectation equals the actual sample quantile. Consider the aggregate RIF at quantile $\tau$:

$$RIF(Y; Q_\tau) = Q_\tau + \frac{\tau - \mathbb{1}\{Y \leq Q_\tau\}}{f_Y(Q_\tau)} \tag{11}$$

where $f_Y(Q_\tau)$ is the density of the marginal distribution of $Y$ computed at $Q_\tau$. $\mathbb{1}\{Y \leq Q_\tau\}$ is an indicator function that takes value one if the condition in $\{\cdot\}$ is true and zero

[9] Gelbach (2002) argues that it is not an identification but a conceptual problem.

otherwise. The following properties of the $RIF(\cdot)$ hold. The mean of the $RIF(\cdot)$ is equal to the actual quantile; $E[RIF(Y;Q_\tau)] = Q_\tau$. For the mean of its expectation conditional on covariates $X$, we have: $E_X[E[RIF(Y;Q_\tau)|X]] = Q_\tau$. Firpo et al. (2009) suggest a linear relation between the conditional expectation of the RIF, $E[RIF(Y;Q_\tau)|X]$, and the explanatory variables, $X$ and $D$, defined as unconditional quantile regression. The unconditional quantile regression model can then be estimated as a linear probability model and reads as:

$$RIF(Y;Q_\tau) = D\beta_\tau + X\gamma_\tau + u_\tau \tag{12}$$

where $D$ is a vector of gender dummies and $X$ is a matrix of regressors with $K$ explanatory variables including the constant. The corresponding quantile-specific coefficient vectors are $\beta_\tau$ and $\gamma_\tau$, respectively. The quantile-specific error term is represented by $u_\tau$.

Given linearity in the linear unconditional quantile or the RIF-OLS model, the coefficient vector $\beta_\tau$ can be estimated by OLS and interpreted by the unconditional mean interpretation.

As mentioned before, we run separate Wage-LASSOs at selected percentiles. In the Wage-LASSO at $\tau$, the corresponding $RIF(\cdot)$ becomes the outcome variable in equation (2). That is, we may obtain different sets of selected controls at different points of the wage distribution. To be precise, the quantile-specific set of selected controls is the union of the Gender-LASSO (equation (3)) and the corresponding quantile specific Wage-LASSO. Thus, controls selected by the Gender-LASSO are the same in all models, i.e. at the mean as well as along the distribution.

To take additional uncertainty of the estimation procedure, introduced by the calculation of the RIFs, into account, we compute standard errors through bootstrapping with 500 replications for the estimation beyond the mean. We apply a clustered bootstrap procedure, where we re-sample individuals to account for potential correlations over time. In each bootstrap replication we re-run the complete procedure except the model selection based on the double-LASSO due to computational intensity of the LASSO. Thus, each bootstrap replication relies on the same sets of selected controls.

# 3 Data and Descriptive Statistics

For our empirical analysis, we use data from the German Socio-Economic Panel Study (SOEP), a representative annual household panel covering more than 11,000 households in Germany (Wagner et al., 2008). The annual surveys consist of a set of household and individual questionnaires specifically designed for the different household members. As the SOEP includes both high quality individual level and important job characteristics, it is particularly well suited for our analysis. We use information from the survey years

2005, 2009 and 2013 (the years, where most character skill measures are included) and restrict the analysis to full-time employees.

The dependent variable of our analysis is the natural logarithm of hourly gross wages. We calculate this variable based on monthly gross wages and weekly working hours (agreed in contract). We use the discounted three-year average of hourly wages in order to compensate for one-time reporting deviations. As we use pooled data over different survey years, calculating the RIFs on the pooled sample ignores potential changes in the wage distribution over time (DiNardo et al., 1996). Panel A of Figure 1 shows that the wage distribution in our sample indeed varies over time.[10] Therefore, we calculate the RIF based on log hourly wages adjusted for the survey years. To be precise, we run a regression of the log hourly wage on the full set of survey-year dummies and instead of the log hourly wage we use the residual from this regression for calculation of the RIFs. Panel B of Figure 1 shows that based on this survey-year adjustment the wage distributions do no longer vary substantially over the survey years.[11] Alternative approaches such as Inverse Probability Weighting (IPW) do yield much more distorted wage distribution over time (see Panel C of Figure 1).[12]

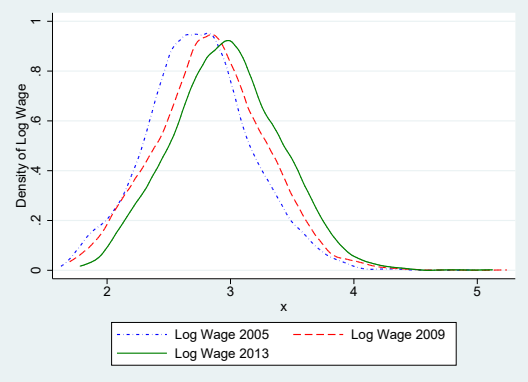
We exploit the extensive information provided by the SOEP and extract control variables like years of education, actual labor market experience and job tenure as well as job characteristics like type of contract (limitation yes or no), firm size, union status or presence of a works council. Additionally, we include occupational and industrial or sectoral dummies. For their classification, we use the ISCO88 (1-Digit) in case of occupations and NACE (2-Digits) for industries and sectors. Further, the SOEP questionnaires contain questions covering the Big Five Personality Traits (openness, conscientiousness, extraversion, agreeableness and neuroticism), locus of control, reciprocity and willingness to take a risk (see Appendix A for details on the character skill measures and definitions). Since character skills have only recently been considered as potential drivers of gender pay differentials, their presence is one of the main advantages of the SOEP compared to other data sources. In order to account for potential endogeneity between character skills and earnings, we restrict our sample to individuals aged between 30 to 60 years. Recent research has shown that character skills are stable for adults in working age, while they change mainly for young or elderly individuals (Cobb-Clark and Schurer, 2012, 2013). Since character skills might change over the life cycle, following Nyhus and Pons (2005), we control for age by regressing each character skill measure on age as well as on the second-order polynomial of age. The residuals from these regressions reflect the character skill measures

[10] The means of the distributions are statistically significantly different at a 1% significance level.

[11] In this case, the means of the distributions are not statistically significantly different at a 10% significance level.

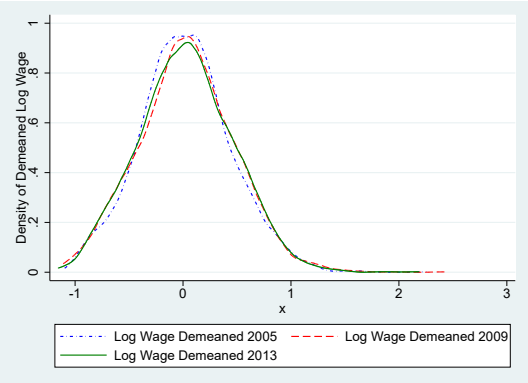
[12] As in the case with no adjustment, the means of the distributions are statistically significantly different at a 1% significance level.

Figure 1: Distribution of Log Hourly Wages in 2005, 2009 and 2013 by Adjustment Method



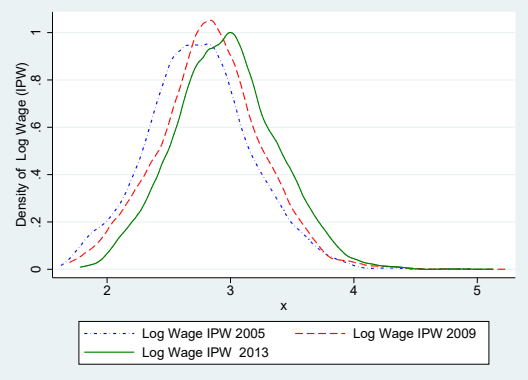

Panel A: Distribution of Log Hourly Wages in 2005, 2009 and 2013 (No Adjustment)

*Notes:* Wage distribution in 2005 (dash-dotted line), 2009 (dotted line) and 2013 (solid line). No sample weights used. Source: SOEP v33.




Panel B: Distribution of Log Hourly Wages in 2005, 2009 and 2013 (Survey-Year Adjustment)

*Notes:* Wage distribution in 2005 (dash-dotted line), 2009 (dotted line) and 2013 (solid line). No sample weights used. Source: SOEP v33.




Panel C: Distribution of Log Hourly Wages in 2005, 2009 and 2013 (Inverse Probability Weighting)

*Notes:* Wage distribution in 2005 (dash-dotted line), 2009 (dotted line) and 2013 (solid line). No sample weights used. Source: SOEP v33.

free of age effects (Heineck and Anger, 2010). In our analysis, we use the age-adjusted skill measures as (potential) control variables.

In the following, we define the control variables of the three conventional models. We start with a model specification that replicates the 'full' specification of Blau and Kahn (2017). This specification includes controls for years of education, age, actual labor market experience[13], migration background, urban residence, dummies for federal state, survey year, occupations and industries as well as union status. In line with Blau and Kahn (2017) we refer to this specification as the 'full specification'. In a next step, we augment this specification with the nine character skill measures and end up with the 'augmented specification'. This model controls additionally for the Big Five Personality Traits – openness, conscientiousness, extraversion, agreeableness and neuroticism –, locus of control, reciprocity and willingness to take a risk. The set of character skill controls is motivated by the literature and by data availability (Collischon, 2017; Heineck and Anger, 2010; Mueller and Plug, 2006; Nyhus and Pons, 2012). Finally, we add experience squared, tenure with the current employer, age squared, type of contract, firm size, presence of a works council and marital status and refer to this as the 'baseline specification'. We expect these model specifications to yield estimates of the adjusted GPG in line with the recent literature and refer to these three models as 'conventional models'.

The main contribution of our paper is the comparison of model specifications proposed by the post-double-LASSO procedure and the 'conventional models'. Remember that the post-double-LASSO estimator selects among the set of potential controls (important) wage predictors and variables with (pronounced) gender differences. Therefore, the careful identification of the set of potential controls among which the data-driven procedure chooses is pivotal for the analysis. Besides the above mentioned variables used in the conventional model specifications, we include information about religion of the individual, i.e. dummies for being Catholic, Protestant, Muslim or belonging to another religion, regions of origin (with base category Germany for non-migrants). The potential set of controls is extended further to dummies for the highest level of educational attainment, German and Math grades in last record, whether or not a Mathematics, Informatics, Natural Science or Technology (MINT) subject was studied, parental educational attainment, asset income, labor market experience as self-employed or part-time employee and past periods of unemployment. Further, we add controls for life and work satisfaction, years of parental leave, number of children and a dummy variable that equals one if children below the age of six are present in the household and zero otherwise. Additionally, as SOEP data is self-reported, we add controls for interview conditions using average sun hours on the

[13]Blau and Kahn (2017) use PSID data and control for potential labor market experience.

interview day in the federal state (based on data from the German weather service[14]) and a dummy identifying whether the interviewer and interviewee have the same sex.

Our final sample consists of 8,489 positive wage observations of full-time employed individuals aged between 30 and 60 years (3,161 or 37% women and 5,328 or 63% men). Table 1 reports descriptive statistics for log hourly wages and selected variables separately for men and women. The last column shows the gender-specific difference and whether this difference is statistically significant. Panel A of Table 1 represents log hourly wages, Panel B selected variables of the conventional models and Panel C selected further potential control variables used in the machine learning models. In our sample, women earn significantly less than men. The raw GPG in hourly wages amounts to 16.8% (log approximation). In line with the widely discussed reversal of the gender education gap (Blau and Kahn, 2017), we find that women have on average 4.5 more months of education. Men have on average about 4.5 years more labor market experience and their average job tenure exceeds those of females by almost two years. Moreover, men are more likely to have a permanent contract and to be a union member. Among full-time employed individuals men are more likely to be married. On average, women are more likely to work in small firms. The above mentioned variables differ significantly between men and women at the 1% significance level, except permanent contract and small firm that differ at the 10% and 5% level, respectively. Both men and women are equally likely to work in medium or large firms and in firms with a works council and do not significantly differ with respect to their answers on migration background and urban residence.

Panel C of Table 1 shows that we can think of various other potential control variables that are significantly different between men and women. For example, women study substantially less often a MINT subject, have significantly longer periods of part-time employment, unemployment and parental leave. Further, full-time female employees have significantly fewer children and are less like to have young children compared to their male colleagues.

Figure 2 shows gender-specific differences in industrial sorting. While in *Health and Social Work* the female share is more than four times as high as the male share, male shares exceed female shares in industries like *Basic Metals and Fabricated Metal Products*, *Construction* or *Transport, Storage and Communication.* Moreover, females are more likely to work in the *Education* sector. Women and men differ not only with respect to sorting into industries, but also with respect to sorting into occupations. Figure 3 shows that most women belong to *Skilled Workers*, which includes inter alia nurses and administrative secretaries. Here, 25% of all men can be assigned to *Craftsmen.* These

[14] *Source:* `https://www.dwd.de/DE/leistungen/klimadatendeutschland/klarchivtagmonat.html?nn=16102` (accessed October,2019). We identify the individual interview condition using indicator variables of the interviewee and region and match it with the sun hours in that region at the date of the interview.

Table 1: Descriptive Statistics by Gender, Selected Variables

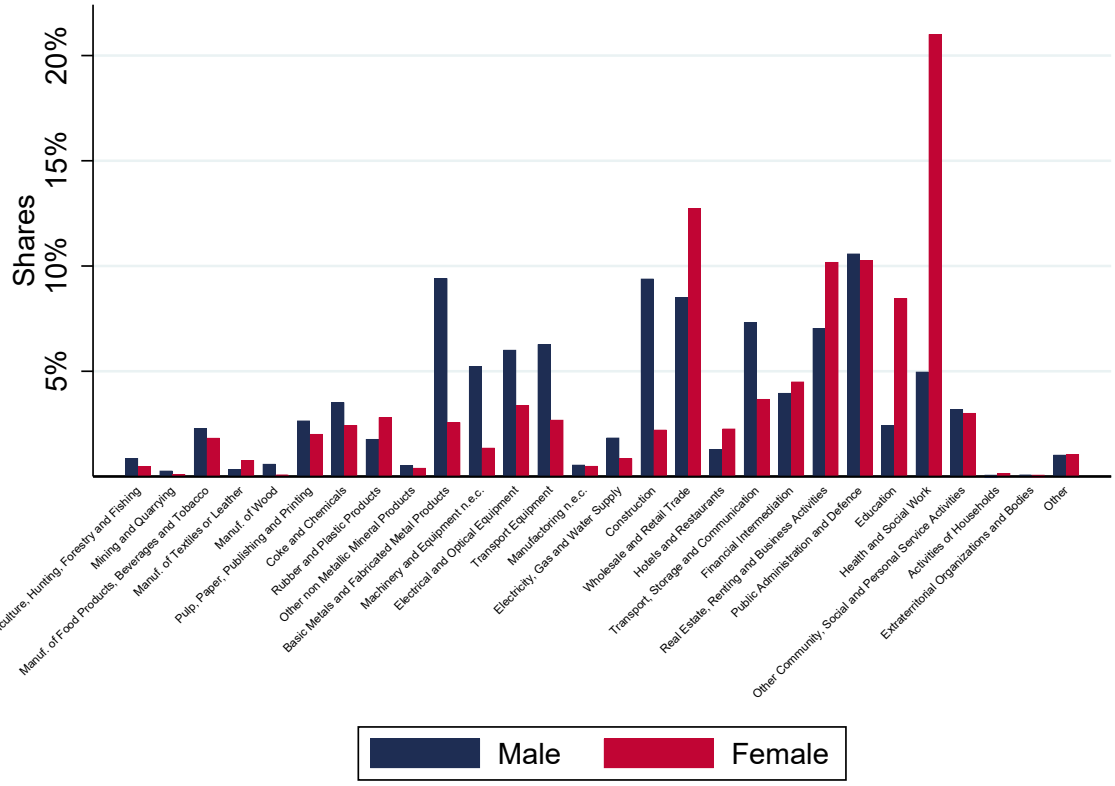
| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Men | | Women | | Difference |
| Variable | Mean | Std.Dev. | Mean | Std.Dev. | Men - Women |
| *Panel A: Dependent Variable* | | | | | |
| Log Hourly Gross Wage | 2.860 | 0.402 | 2.693 | 0.414 | 0.168*** |
| *Panel B: Selected Conventional Control Variables* | | | | | |
| Education (Years) | 12.32 | 2.536 | 12.70 | 2.543 | -0.379*** |
| Labor Market Experience (Years) | 22.30 | 9.412 | 17.76 | 9.805 | 4.546*** |
| Tenure (Years) | 14.49 | 10.45 | 12.72 | 9.537 | 1.771*** |
| Permanent Contract (Dummy) | 0.937 | 0.242 | 0.920 | 0.271 | 0.017* |
| Small Firm (Dummy) | 0.152 | 0.359 | 0.188 | 0.391 | -0.037** |
| Medium Firm (Dummy) | 0.295 | 0.456 | 0.274 | 0.446 | 0.021 |
| Large Firm (Dummy) | 0.553 | 0.497 | 0.538 | 0.499 | 0.016 |
| Works Council (Dummy) | 0.612 | 0.487 | 0.594 | 0.491 | 0.017 |
| Union Member (Dummy) | 0.278 | 0.448 | 0.179 | 0.383 | 0.099*** |
| Married (Dummy) | 0.641 | 0.480 | 0.508 | 0.500 | 0.133*** |
| Migration Background (Dummy) | 0.158 | 0.365 | 0.162 | 0.368 | - 0.004 |
| Urban Residence (Dummy) | 0.703 | 0.457 | 0.685 | 0.464 | 0.018 |
| *Panel C: Selected Further Potential Control Variables* | | | | | |
| Studied MINT Subject (Dummy) | 0.0800 | 0.271 | 0.0324 | 0.177 | 0.048*** |
| Part-time Experience (Years) | 0.517 | 1.781 | 3.956 | 5.600 | -3.439*** |
| Period of Unemployment (Years) | 0.479 | 1.202 | 0.687 | 1.556 | -0.207*** |
| Parental Leave (Years) | 0.0027 | 0.0703 | 0.0301 | 0.202 | -0.027*** |
| Number of Children | 0.770 | 0.992 | 0.471 | 0.758 | 0.299*** |
| Children < 7 Years (Dummy) | 0.0195 | 0.138 | 0.00380 | 0.0616 | 0.016*** |
| Observations | 5,328 | | 3,161 | | |

*Notes:* Calculations use SOEP sample weights. 'Small Firm' equals one if the firm has at most 19 employees, zero otherwise. 'Medium Firm' equals one if the firm has between 20 and 199 employees, zero otherwise. 'Large Firm' equals one if the firm has at least 200 employees, zero otherwise. 'Number of Children' and 'Children < 7 Years' refer to children in the household. Reported differences are based on a regression of the selected variables on a male dummy. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. Robust standard errors (clustered at the individual level) are used. *Source:* SOEP v33.

descriptive results are in line with the literature (e.g. Blau and Kahn, 2017) suggesting that sorting in industries and occupations may be an important driver of gender differences in pay.

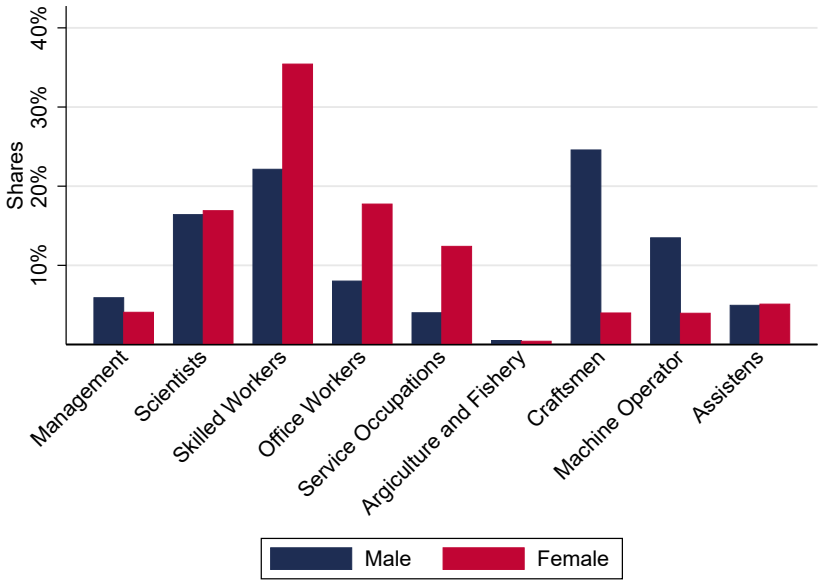
Figure 4 shows gender-specific differences in the nine considered character skills measures. In line with the recent literature, we find pronounced gender differences in individual character skills (Weisberg et al., 2011). In particular, female scores are higher for openness, conscientiousness, extraversion, agreeableness and neuroticism. On average female scores for conscientiousness are 0.11 standard deviations higher. In contrast, women are less willing to take a risk and less prone to negative reciprocity. Differences in these character skills are statistically significant at a 1% significance level. We find no significant differences between men and women for external locus of control and positive reciprocity, i.e. the tendency to believe that personal successes or failures result from external factors beyond the individual's control and to reward kind actions with kind behavior, respectively.

Figure 2: Gender-Specific Differences in Industries



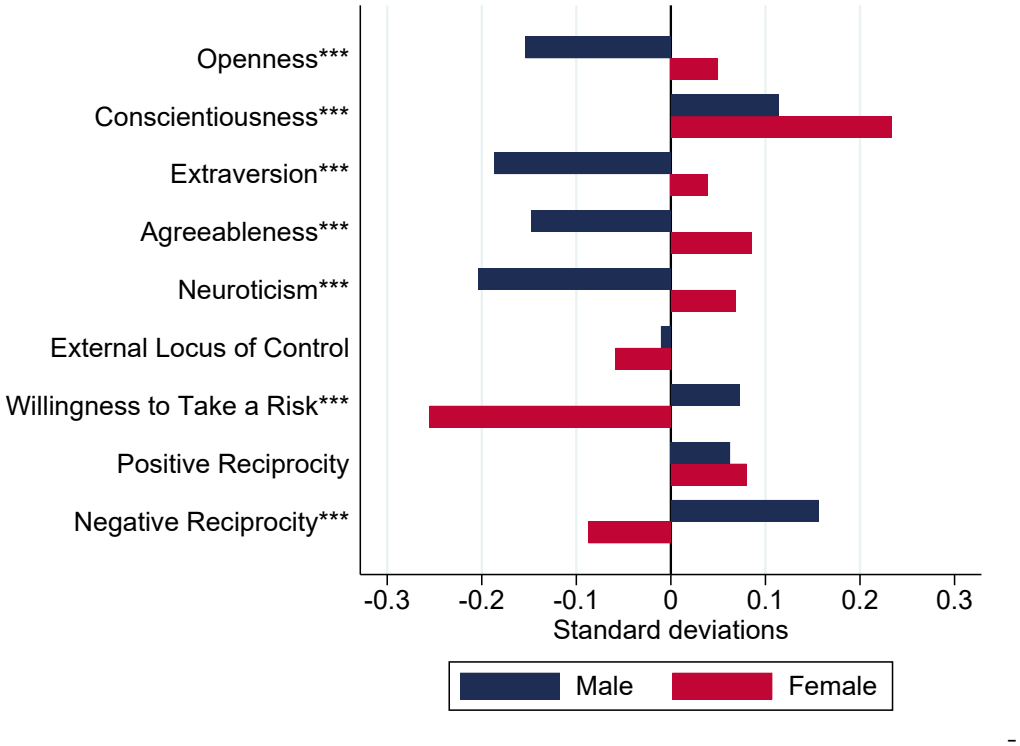

*Notes:* Calculations use SOEP sample weights. Male and female means of industry dummies (based on NACE 2-Digits) are reported. *Source:* SOEP v33.

Figure 3: Gender-Specific Differences in Occupations




*Notes:* Calculations use SOEP sample weights. Male and female means of occupational dummies (based on ISCO88 1-Digit) are reported. *Source:* SOEP v33.

Figure 4: Gender-Specific Differences in Character Skills




*Notes:* Calculations use SOEP sample weights. Reported differences are based on a regression of the standardized character skill measure on a male dummy. *** behind the name of the character skill indicate significant differences at the 1%-significance-level. Robust standard errors (clustered at the individual level) are used. *Source:* SOEP v33.

# 4 Results

In this Section, we present the sets of controls selected by double-LASSO. Next, we compare estimation results of the (adjusted) GPGs obtained from the different model specifications. Further, we compare the estimated GPGs with respect to robustness to remaining selection on unobservables evaluating the capability of the corresponding regression models to control for the observables in our data set. Finally, we compare aggregate and detailed decomposition results of the conventional and machine learning based models.

## 4.1 Selected Control Variables for the Post-Double-LASSO Estimator

As stated, the set of potential controls among which double-LASSO selects contains all control variables that are part of the baseline specification. Recall that our baseline specification controls for quadratic polynomials of experience and age, tenure with the current employer and years of education as well as for marital status, migration background, urban residence, type of contract (limitation yes or no), firm size, union status, presence of a works council and for the federal state as well as for the survey years. We further include additional potential controls like religion, migration background, region of origin, school-leaving degree, last grades in Math and German, whether a MINT subject was studied, experience as self-employed and part-time worker as well as periods of unemployment and parental leave. We extend the set of potential controls further by background information such as parents' education, life and work satisfaction, asset income, number of children, young children and interview characteristics like whether the interviewer has the same sex as the interviewee and sunshine hours at the interview day in the federal state.

After having collected these variables commonly used in and/or suggested for estimation of the GPG, we end-up with a set of 111 raw potential controls (see Table 2). For all non-binary potential control variables (23 variables) we construct second-order polynomials. In a next step we interact all 134 (111 + 23) potential controls with each other. The latter results in $\binom{134}{2} + 134 = 9,045$ theoretical potential control variables. Among these 9,045 variables, in line with Belloni et al. (2014a), we exclude those with extremely small standard deviations (3,143 variables or 35%).[15] Next, we check whether any variables in our set of potential controls are highly correlated. Again relying on Belloni et al. (2014a), we exclude one variable among each pair of variables with a bivariate correlation coefficient higher than 0.99. The latter leads to exclusion of further 81 variables and a final set of 5,821 potential controls.

[15] For all theoretical potential control variables, we divide the standard deviation by the range between the 99th and 1th percentile. Whenever this ratio is smaller than 0.1 (the 10th percentile of the ratio for all potential controls) the corresponding variable is excluded from the set of theoretical potential controls. Variables with no difference between the 99th and 1th percentile are also dropped.

Table 2: Set of Potential Controls

| | | |
|---|---|---|
| Set of raw controls | 111 | |
| Set of higher-order polynomials | 23 | |
| Set of interaction terms | 8,911 | |
| Theoretical set of control variables | 9,045 | |
| Excluded: small standard deviations | 3,143 | 35% |
| Excluded: highly correlated | 81 | 1% |
| Final set of potential controls | 5,821 | |

In case of the restricted LASSO, we force the control variables of the baseline model, except the character skill measures and industrial and occupational dummies, to be part of the model. This set of controls is the so-called amelioration set in our restricted LASSO specification and is the same for all points of the wage distribution. Overall, the amelioration set contains 32 controls, i.e. the restricted double-LASSO selects among 5,789 (5,821 - 32) potential control variables. For the unrestricted LASSO, we relax the constraint that the variables included in the amelioration set are forced to be part of the wage equation. In this case, the LASSOs (Gender and Wage) select among the full set of potential controls (i.e. 5,821 potential control variables).

As different factors may matter at different points of the wage distribution, we use the Wage-LASSO for model selection at the mean as well as at selected percentiles. Thus, as we use unconditional quantile regression or RIF-OLS, we run a separate Wage-LASSO for model selection at the mean as well as for the lower (RIF at the 10th percentile), middle (RIF at the 50th percentile) and upper part (RIF at the 90th percentile) of the wage distribution.

Table 3: Number of Selected Controls

| | Mean | | 10th Percentile | |
|---|---|---|---|---|
| LASSO | Restricted | Unrestricted | Restricted | Unrestricted |
| Selected Set of Controls: | 98 | 173 | 89 | 146 |
| $s^2$ | 9,604 | 29,929 | 7,921 | 21,316 |
| | 50th Percentile | | 90th Percentile | |
| LASSO | Restricted | Unrestricted | Restricted | Unrestricted |
| Selected Set of Controls: | 86 | 150 | 97 | 133 |
| $s^2$ | 7,396 | 22,500 | 9,409 | 17,689 |

*Notes:* Approximate Sparsity: $s^2 \ll N$, with $N = 8,489$. The amelioration set includes 32 control variables.

Tables B.1 - B.4 in Appendix B show detailed lists of the sets of selected controls.

Tables B.1 and B.3 contain all variables selected by the restricted and unrestricted Gender-LASSO, respectively. Tables B.2 and B.4 consist of different sets of selected controls chosen by the different Wage-LASSOs. Each of the latter two tables contains the lists for estimation at the mean, the 10th, 50th and 90th percentile.

Both Gender-LASSOs detect important gender differences with respect to family-related variables like time of parental leave or number of children and marital status. Further, typically male-dominated occupations like *Technicians* or *Craftsmen*, and female-dominated sectors like *Education* or *Health and Social Work* are selected. All nine character skill measures besides positive reciprocity are selected at least as parts of interactions. Character skill measures are often selected as interactions with human capital and labor market controls as well as with sectors and occupations. The selected interactions of character skills and educational controls support the findings of Gensowski (2018) suggesting that the effects of character skills differ across educational groups. Moreover, the Gender-LASSOs detect gender differences with respect to human capital measures in a flexible form. Numerous interactions with human capital and for example occupations, industries or labor market characteristics are chosen.

Compared to the baseline specification both Wage-LASSOs at the mean select not only years of education and age as human capital controls. Type of education and grades seem to be also important wage predictors. As in the Gender-LASSOs, human capital controls enter the model very flexibly and are especially often interacted with occupations and sectors or as second-order polynomials. Concerning labor market controls, except for years of parental leave, all are selected at least as parts of interactions. Thus, not only full-time labor market experience and tenure do matter, but also past periods of unemployment and part-time employment. As the restricted LASSO already contains federal state dummies as part of the amelioration set they are never chosen, while in case of the unrestricted LASSO they appear interacted with other demographic and firm characteristics. Again, as in the Gender-LASSOs, family-related controls like number of children and martial status are found to be important. Risk aversion enters at the mean in both LASSO models not only linearly but also interacted with firm characteristics (permanent contract). Overall, not only risk aversion but also external locus of control, negative reciprocity, openness, agreeableness and neuroticism are chosen. Besides individual factors, job-related characteristics like firm size and type of contract are found to be important for wage prediction at the mean. Regarding occupational and industrial controls, the Wage-LASSOs do not select occupations like *Management* or *Skilled Workers*. In case of industries the female-dominated fields *Education*, *Health and Social Work* and *Wholesale and Retail Trade* as well as male-dominated fields like *Transport Equipment* or *Machinery and Equipment* are selected. The latter are particularly often included as interactions with human capital and labor market controls suggesting that gender differences in industrial

sorting are important for wage prediction.

Along the wage distribution the composition of the sets of selected controls changes. We observe that human capital variables are often chosen interacted with demographic controls like being Catholic or number of children as well as with certain sectors (e.g. *Construction*). At the 90th percentile academic education and second-order polynomials of years of education (as parts of interactions) are more often selected than at other parts of the distribution. Despite several interactions of human capital variables with demographic and sectoral controls, we find further interactions with job-related characteristics like firm size and type of contract. For labor market controls, in case of the restricted LASSO – where experience, experience squared and tenure are forced to be part of the models – at the 10th percentile mainly past periods of unemployment are chosen. The unrestricted LASSO selects experience and tenure mostly interacted with job-related characteristics. At the 50th percentile the restricted LASSO selects experience, tenure, and past periods of unemployment interacted with sectors and occupations, while the unrestricted LASSO additionally picks part-time experience. Labor market controls enter the model interacted with human capital variables and demographics (number of children) particularly at the median. At the 90th percentile experience dominates the other labor market controls. This result suggests that the level of experience is an important wage predictor at the top and that lower levels of experience are particularly punished at the top of the distribution. At the 10th and 50th percentile among the set of potential demographic controls mainly federal states interacted with other demographic controls, human capital and job-related characteristics are selected. Many interactions with educational measures are chosen. At the top and bottom, interactions of having children and education are selected relatively often. The result is in line with Wilde et al. (2010) who show that the wage gap associated with children differs between educational groups. Several interactions with character skills are chosen. Risk aversion seems to be particularly relevant at the bottom of the distribution, while at the median and top character skills like negative reciprocity and external locus of control are found to be important wage predictors.

For estimation of the GPG at the mean as well as along the distribution we combine the variables selected by the Gender-LASSO and the respective Wage-LASSO (mean, 10th, 50th or 90th percentile). In these combined sets individual characteristics like education, years of labor market experience but also years of part-time experience and past periods of unemployment are, at least as parts of interactions, chosen. Moreover, number of children is as part of interactions always selected. The main interacted controls are marital status, educational attainment, years of labor market experience. However, presence of small children in the household is never chosen. This result suggests that small children is an appropriate exclusion restriction in sample-selection models, while the number of children is not. The latter is in line with results of Castagnetti et al. (2018) testing the validity

of these instruments in employment selection. Furthermore, the fact that the number of children is usually not included in conventional wage models may be a potential source of endogeneity or omitted variable bais.

Among the considered character skill measures positive reciprocity is the only character skill that is never, neither as part of an interaction nor as a direct control, chosen. Concerning employer learning, the only selected interaction of a character skill measure with job tenure occurs in case of openness (restricted and unrestricted Gender-LASSO). Even if we never force occupational and industrial controls to be part of the model, the LASSO selects at all points and in both the restricted and unrestricted specification numerous occupational and industrial variables. Consequently, occupational and industrial sorting, as suggested by the literature (e.g. Blau and Kahn, 2017), plays an important role for gender differences in pay.

All in all, compared to the conventional models data-driven model selection suggests that more flexible model specifications (with interactions and second-order polynomials) are required for estimation of gender differences in pay. As these controls may be hard to detect for a (human) researcher, the double-LASSO offers a convenient way for model selection. In particular, machine learning suggests that different sets of control variables are relevant at different points of the wage distribution. This result is in line with the literature finding that different factors are important for different subgroups of the population (Gensowski, 2018; Juhn and McCue, 2017; Wilde et al., 2010). The sets of selected controls contain numerous interactions of dummy variables. Thus, as mentioned in Section 2.4, it is not feasible (or at least very difficult) to classify meaningful groups among these variables. Therefore, we do not represent the corresponding coefficient estimates as deviations from the grand mean in Section 4.4.

Table 3 shows the number of elements in the different sets of selected controls. As the regularity condition ($s^2 \ll N$) is not fully met, we repeat the analysis using sample splitting in Section 5. In case of sample splitting, the assumption of approximate sparsity becomes: $s \ll N$.

### 4.2 Regression Outcome

In the following, we discuss the estimation results obtained from the different model specifications at the mean as well as beyond. Table 4 compares the estimated GPGs for the short and conventional wage models. Panel A shows the results at the mean, Panel B, C and D at the 10th, 50th and 90th percentile, respectively. Column (1) shows the raw or unadjusted GPG obtained from the short regression model.

The unadjusted gap is equal to 16.8% (log approximation) at the mean. Over the distribution, we find a U-shaped pattern of raw gender differences in pay. That is, we find particularly pronounced gender differences in pay at the bottom and top. The full

Table 4: Unadjusted and Adjusted Gender Pay Gap (GPG) at the Mean and Selected Percentiles, Short Model and Conventional Models

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | | Specification: | |
| | Short | Full | Augmented | Baseline |
| *Panel A: Mean* | | | | |
| Unadjusted Gap | -0.168*** | | | |
| | (0.017) | | | |
| Adjusted GPG | | -0.109*** | -0.101*** | -0.099*** |
| | | (0.014) | (0.015) | (0.013) |
| $R^2$ | 0.037 | 0.549 | 0.559 | 0.618 |
| *Panel B: 10th Percentile* | | | | |
| Unadjusted Gap | -0.255*** | | | |
| | (0.040) | | | |
| Adjusted GPG | | -0.107*** | -0.103*** | -0.104*** |
| | | (0.035) | (0.037) | (0.037) |
| $R^2$ | 0.021 | 0.168 | 0.242 | 0.247 |
| *Panel C: 50th Percentile* | | | | |
| Unadjusted Gap | -0.165*** | | | |
| | (0.021) | | | |
| Adjusted GPG | | -0.125*** | -0.113*** | -0.099*** |
| | | (0.018) | (0.019) | (0.018) |
| $R^2$ | 0.029 | 0.260 | 0.363 | 0.372 |
| *Panel D: 90th Percentile* | | | | |
| Unadjusted Gap | -0.198*** | | | |
| | (0.029) | | | |
| Adjusted GPG | | -0.200*** | -0.176*** | -0.170*** |
| | | (0.033) | (0.034) | (0.034) |
| $R^2$ | 0.014 | 0.176 | 0.255 | 0.259 |
| Observations | 8,489 | 8,489 | 8,489 | 8,489 |

*Notes:* SOEP sample weights used. Figures show the unadjusted and adjusted gender wage gaps obtained from a linear regression of the log wage on a dummy for gender (female = 1) and different sets of covariates estimated at the individual level. In case of the short model, the set of covariates includes only a constant. The full specification includes education, age, labor market experience, migration background, urban residence as well as federal state, survey-year, occupational and industrial dummies. The augmented specification controls additionally for the Big Five Personality Traits (openness, conscientiousness, extroversion, agreeableness and neuroticism), the locus of controls, willingness to take a risk, negative and positive reciprocity. The baseline specification includes additionally age squared, labor market experience squared, job tenure and dummies for marital status, contract type, works council and firm size (large and medium). Robust and bootstrapped standard errors (500 replications) clustered at the individual level in parentheses for the mean and the selected percentiles, respectively. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. *Source:* SOEP v33.

model based on Blau and Kahn (2017) suggests an increasing adjusted gap over the distribution (column (2)). Adding character skills to the model corrects the gap downwards at all points (column (3)), while the increasing pattern over the distribution persists. In contrast, our proposed baseline model suggests, like the short model, a particularly high GPG at the top of the distribution (column (4)). Across all model specifications the adjusted GPG is most pronounced at the top (Panel D). The results of the conventional models suggest that there is a glass ceiling over Germany (difference in GPG at the 10th or 50th and 90th percentile of more than two percentage points). In contrast, only the full model finds slight evidence for a sticky floor (difference in GPG between the 10th and 50th percentile of at least two percentage points). The short model suggests both glass ceiling and sticky floors.

Table 5 shows the adjusted GPG for the restricted and unrestricted LASSO specifications (column (1) and (2), respectively). Again, Panel A shows the results at the mean, Panel B, C and D at the 10th, 50th and 90th percentile, respectively. In both LASSO specifications, the estimated gap at the mean and median remains roughly stable. Moreover, the mean and median estimates of the adjusted gap change only slightly compared to the augmented and baseline model specifications. In the restricted LASSO, the gap at the median does not change compared to the baseline model (column (1)) and remains at 9.9%. The unrestricted LASSO estimates the gap in the middle of the distribution to be equal to 10%. At the top and bottom, the RIF-OLS models based on the LASSO specifications show that estimates of the adjusted GPG differ substantially compared to the conventional models. In both LASSO specifications, the gap at the top (amounting to 14%) is – as in the conventional models – most pronounced, though three to six percentage points lower. In the unrestricted LASSO specification, the gap at the bottom and top of the wage distribution are equally pronounced (14%), while in the restricted LASSO model the gap at the bottom is two percentage points lower. Both data-driven models suggest glass ceiling as well as sticky floors.

In order to check whether the quantile-specific selected sets of controls deliver different results compared to the (restricted and unrestricted) mean model specifications, we estimate the adjusted GPG at the top and bottom using the mean model specifications. The latter is motivated by the fact that the literature generally uses the mean specification at all points of the wage distribution. The estimated (adjusted) GPGs at the bottom and top differ up to two percentage points compared to using the selected set of controls by the double LASSO at the corresponding percentiles.[16] As a consequence, we obtain changed relations of top-to-bottom or bottom-to-top GPGs. However, in our case the insights on glass ceiling and sticky floors do not change. All in all, this result implies that using percentile-specific sets of control variables may be more appropriate for estimation

[16]Results are available from the authors upon request.

of gender differences in pay.

The results of both LASSO specifications underline the finding of the short model of a U-shaped pattern of the unadjusted gap over the wage distribution and contradict the finding of a strictly increasing adjusted pay gap along the wage distribution. The latter was suggested by the full and augmented specification based on the literature. This implies that the policy conclusions drawn from the estimates obtained from the post-double-LASSO estimators and the conventional models may be different. Beyond finding evidence for glass ceiling, using the LASSO specifications, we find also evidence for sticky floors. Further, the gap at the bottom is markedly corrected upwards (two to four percentage points), while the gap at the top is substantially corrected downwards (three to six percentage points) in the LASSO specifications. That is, conventional models may lead to an underestimation of the gap at the bottom and an overestimation of the gap at the top. Misled policy measures may be the consequence.

### 4.3 Robustness to Remaining Selection on Unobservables

In the next step, we apply the method of Oster (2019) in order to make statements about robustness to remaining selection on unboservables of the estimated (adjusted) GPGs. Based on the assumption that selection on observables is informative about selection on unobservables, we calculate the value of the proportionality parameter necessary to produce a zero GPG ($\delta^*$). The latter allows us to determine the degree of remaining selection on unobservables (relative to selection on observables) that would be required to result in a zero GPG. In line with Oster (2019), we consider the adjusted GPG to be robust to remaining selection on unobservables whenever $\delta^* \geq 1$ (see Section 2.3 or Oster, 2019, for details).

We report the values of $\delta^*$ for the conventional and data-driven models in Table 6. Panel A shows the results at the mean, Panel B, C and D at the 10th, 50th and 90th percentile, respectively. The five columns refer to the different model specifications. Thus, again we compare the conventional to the data-driven models. In case of the LASSO specifications, model selection is based on the double-selection procedure and the adjusted gaps are estimated by the corresponding post-double-LASSO estimator. Regardless of the model specification used, the values for $\delta^*$ are highest at the top and lowest at the bottom of the distribution. This result suggests that estimated GPGs at the top are more robust to remaining selection on unobservables. Across models, all reported values of $\delta^*$ exceed the threshold of 1. To be precise, $\delta^*$ ranges between 1.1 and 4.9. For example, in the baseline model, shown in column (3), the proportionality parameter ranges between 1.1 and 3.4. The number 3.4 corresponds to the 90th percentile and suggests that the estimated GPG at the 90th percentile would turn to zero if remaining selection on unobservables is 3.4 times as high as the selection on observables. The latter implies either that the

Table 5: Adjusted Gender Pay Gap (GPG) at the Mean and Selected Percentiles, Restricted and Unrestricted LASSO Models

| | (1) | (2) |
|---|---|---|
| | Specification: | |
| | Restricted LASSO | Unrestricted LASSO |
| *Panel A: At the Mean* | | |
| Adjusted GPG | -0.106*** | -0.104*** |
| | (0.014) | (0.014) |
| $R^2$ | 0.638 | 0.633 |
| *Panel B: 10th Percentile* | | |
| Adjusted GPG | -0.121*** | -0.142*** |
| | (0.039) | (0.041) |
| $R^2$ | 0.337 | 0.314 |
| *Panel C: 50th Percentile* | | |
| Adjusted GPG | -0.099*** | -0.100*** |
| | (0.020) | (0.020) |
| $R^2$ | 0.430 | 0.434 |
| *Panel D: 90th Percentile* | | |
| Adjusted GPG | -0.140*** | -0.142*** |
| | (0.037) | (0.037) |
| $R^2$ | 0.354 | 0.352 |
| Observations | 8,489 | 8,489 |

*Notes:* SOEP sample weights used. This table shows adjusted gender wage gaps obtained from a linear regression of the log wage on a dummy for gender (female = 1) and different sets of covariates estimated at the individual level. In case of the restricted LASSO, the parameters of variables in the amelioration set are excluded from shrinkage. The amelioration set includes quadratic polynomials of age and labor market experience, job tenure, years of education, marital status, migration background, urban residence, type of contract (limitation yes or no), firm size, union status, presence of a works council as well as dummies for the federal state and survey year. The unrestricted LASSO chooses without restriction among the set of potential controls (5,821 control variables). The detailed list of covariates for the restricted and unrestricted LASSO specifications, respectively, are shown in Appendix B (Tables B.1-B.4). Robust and bootstrapped standard errors (500 replications) clustered at the individual level in parentheses for the mean and the selected percentiles, respectively. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. *Source:* SOEP v33.

Table 6: Remaining Selection on Unobservables at the Mean and Selected Percentiles

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Conventional Models | | | LASSO Specifications | |
| | Full | Augmented | Baseline | Restricted | Unrestricted |
| *Panel A: Mean* | | | | | |
| Proportionality Parameter ($\delta^*$) | 2.562 | 1.987 | 1.903 | 1.643 | 1.673 |
| *Panel B: 10th Percentile* | | | | | |
| Proportionality Parameter ($\delta^*$) | 1.256 | 1.137 | 1.109 | 1.420 | 1.137 |
| *Panel C: 50th Percentile* | | | | | |
| Proportionality Parameter ($\delta^*$) | 2.749 | 1.957 | 1.916 | 1.581 | 1.558 |
| *Panel D: 90th Percentile* | | | | | |
| Proportionality Parameter ($\delta^*$) | 4.907 | 3.731 | 3.380 | 2.111 | 2.024 |
| Observations | 8,489 | 8,489 | 8,489 | 8,489 | 8,489 |

*Notes:* Calculations use SOEP sample weights. Outcome variable is the natural logarithm of gross hourly wages. $\delta^*$ is calculated under the assumption that $R^2_{full} = 1.3 \times R^2_{inter}$. *Source:* SOEP v33.

estimated GPG is extremely robust to remaining selection on unobservables or that it is not adequately controlled for selection on observables, i.e. that not the right set of observables has been chosen. The results for the two LASSO specifications are shown in columns (4) and (5) of Table 6, respectively. Besides the bottom of the wage distribution, the estimated proportionality parameters are lower in the LASSO specifications. Connecting the lower proportionality parameters with the fact that our LASSO specifications explicitly search for important gender differences (Gender-LASSO) we may conclude that the LASSO models control more appropriately for selection on observables.

Overall, we find that all of the calculated proportionality parameters exceed one. Thus, the adjusted GPGs are robust to remaining selection on unobservables. This finding may suggest that at least parts of the remaining gap are due to discrimination. As the proportionality parameters are highest at the top, discrimination may be particularly relevant in the upper part of the wage distribution.

### 4.4 Decomposition Results

The decomposition analysis allows us to identify the part of the raw GPG attributable to gender differences in observable characteristics as well as the part caused by gender differences in prices of these characteristics. We follow the standard case of the Oaxaca (1973) and Blinder (1973) decomposition and use men as non discriminatory wage structure or reference category. As the decomposition results may change on the basis of the selected non-discriminated group, we show for completeness in Appendix C the decomposition

results using women as reference category.

Remember that the choice of the omitted reference category in case of categorical variables is particularly relevant for the estimation of the (detailed) unexplained component (see for example Gardeazabal and Ugidos, 2004). Since the post-double-LASSO estimators result in sets of selected controls that contain numerous interactions of different categorical variables and we are not able to classify them into meaningful groups, representing the coefficient estimates as deviation from the grand mean is hardly feasible. Therefore, the focus lies on the analysis of the explained component. We compare the estimated explained component across the different model specifications considered (conventional and LASSO) on aggregate and in detail. For the detailed decomposition, we look at gender differences in observable characteristics attributable to human capital, labor market, occupational or industrial sorting as well as character skills. We are aware that the post-double-LASSO estimator ensures valid inference only with respect to the adjusted GPG in our setting. The variables in the selected sets of controls may still be endogenous and model selection is not stable as it depends on the underlying sample (Mullainathan and Spiess, 2017). Nevertheless, we inspect the contributions of the different sets of control variables to the explained part of the GPG across models.

Figure 5 shows the estimated explained component of the aggregate decomposition from the conventional and LASSO models along the wage distribution. Figure 6 represents the part of the explained component that can be attributed to human capital characteristics. Figure 7, Figure 8 and Figure 9 report the corresponding estimates for labor market, industrial and occupational controls and character skills, respectively. Table C.1 in Appendix C shows the results of the detailed decomposition at the mean using the conventional model specifications, while Table C.2 represents the corresponding results of the LASSO specifications.

The results in Figure 5 show that based on the conventional wage models a substantial part of the GPG remains unexplained at all parts of the distribution. In contrast, the data-driven models and in particular the unrestricted LASSO specification explain almost the entire GPG from the 50th percentile onwards. For lower percentiles the unrestricted LASSO specification explains only a small part of the differential and substantially less than the conventional models, while the restricted LASSO explains a fraction comparable to that of the conventional wage models. This result can be interpreted in two ways. First, by relying on conventional wage models and restricted data-driven models, we overestimate the explained part at lower percentiles and underestimate it for upper percentiles. Second, the LASSO models are most appropriate for upper parts of the wage distribution. Yet, as we explicitly run the LASSO algorithm at the 10th, 50th and 90th percentile, we expect to reduce omitted variable bias and hence to more appropriately model gender differences in pay at the selected percentiles.

The explained component attributable to human capital characteristics differs substantially between conventional and data-driven models (Figure 6). The effect of the pure human capital controls is negligible in all models and at all points of the wage distribution. At a first glance, this result is surprising as the (pure) human capital controls were almost entirely selected by the Gender-LASSO (cfr. Tables B.1 and B.3 in Appendix B). However, the result is in line with the literature suggesting that classical wage controls such as human capital characteristics are less important in explaining gender differences in pay (Blau and Kahn, 2017). As the double-LASSO selects (beyond the pure human capital controls) various interactions of human capital characteristics and other categories, we represent also the part of the gap that can be attributed to human capital characteristics as well as to their interactions with variables form other sets of potential controls (*All HC restricted LASSO* and *All HC unrestricted LASSO*). In this case, both LASSO models suggest that the part explained by human capitals controls (and their interactions) is decreasing along the wage distribution. The restricted LASSO provides up to the median similar results as the conventional models, while it suggests that female employees outperform their male colleagues in human capital at the top (negative explained component). In contrast, the unrestricted LASSO remains positive throughout the wage distribution and explains markedly more than the other models at all points of the wage distribution.

Gender differences in observable labor market characteristics explain a non-negligible fraction of the gap at all points of the wage distribution using the conventional specifications (Figure 7). Pure labor market controls explain a substantially higher fraction of the GPG for the upper part of the wage distribution using the LASSO models. Contrary, at the bottom, they explain substantially less. Considering pure and interacted human capital controls, the restricted LASSO provides results comparable to the conventional models. The unrestricted LASSO suggests a strictly increasing contribution of gender differences in labor market characteristics along the distribution. Based on the assumption that the LASSO models successfully reduce omitted variable bias, this result implies again that we overestimate the explained component at the bottom but underestimate it at the top using conventional wage models.

Figure 8 shows that gender differences in occupational and industrial sorting do not help to explain the gap at all points of the distribution using the conventional wage models. The LASSO specifications suggest that occupational and industrial sorting lowers the GPG at the bottom but raises it at the top. In this case, a substantial part at the top can be explained by differences in industrial and occupational sorting between men and women. At the bottom, the LASSO models suggest that gender differences in sorting actually decrease the gap. The latter suggests once more that we overestimate the explained component at the bottom but underestimate it at the top. Adding interaction effects to the set of pure occupational and industrial controls brings the estimated effect back to the

level of the conventional models. Thus, we do not find that industrial and occupational sorting explains the GPG except when looking at the pure controls selected by both double-LASSOs.

In terms of gender differences in character skills, we do not find a substantial contributions for most points of the distribution and across all model specifications (Figure 9). This finding is in line with the literature suggesting that character skills explain only a small fraction of the gap (Blau and Kahn, 2017). The only exception is the top, where the unrestricted LASSO model suggests that character skills contribute to a reduction of the GPG (negative component). Considering interactions with character skills additionally to the pure character skill measures in the LASSO specifications slightly corrects the contribution of character skills upwards at the bottom and top. In the latter case, the component is adjusted back to the level of the conventional models, i.e. no significant effect. In the previous case, approximately five percentage points of the 26% raw gap at the 10th percentile can be attributed to character skills and their interactions.

All in all, the decomposition analysis suggests that conventional wage models overestimate the explained part of the gap for lower income earners, while they underestimate the explained component for top income earners. The (aggregate) explained part differs substantially based on the model specification used (conventional or machine learning). Detailed decompositions have shown that the estimates across the conventional and machine learning models are similar for human capital characteristics and character skills but differ substantially for labor market controls as well as occupational and industrial sorting. Taking the more flexible specifications of the LASSO models into account, i.e. interactions, we find generally larger effects of the data-driven compared to the conventional models. This result suggests that depending on the model selection, different conclusions may be drawn. Consequently, model selection is pivotal for detecting potential drivers of GPGs in the labor market.

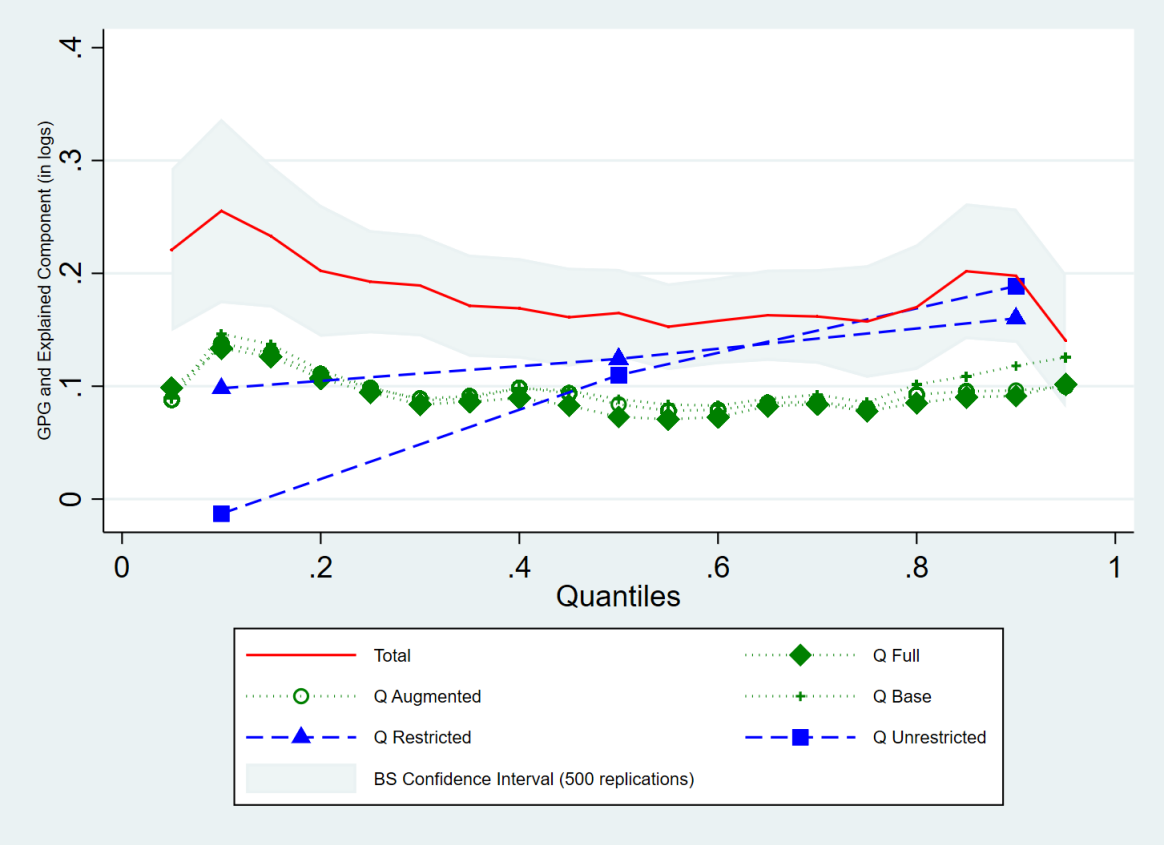
# 5 Robustness Check: Sample Split

In this Section, we repeat the estimation of the adjusted GPG for the LASSO specifications using a sample-split procedure as the assumption of approximate sparsity was not fully met (cfr. Section 4).[17] In case of sample splitting, the regularity condition becomes $s \ll N$ (instead of $s^2 \ll N$, Belloni et al., 2012; Chernozhukov et al., 2018). For sample splitting we divide the sample randomly in two samples of equal size, i.e. in a training and a test sample.

While we use the training sample for model selection, the test sample is used for the

[17]For completeness, we have also run the estimation on the entire period 2005-2013 using forward-filling of the character skill measures. The main insights do not change. Results are available from the authors upon request.

Figure 5: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Aggregate Decomposition (Explained Component) – Male-Reference Category



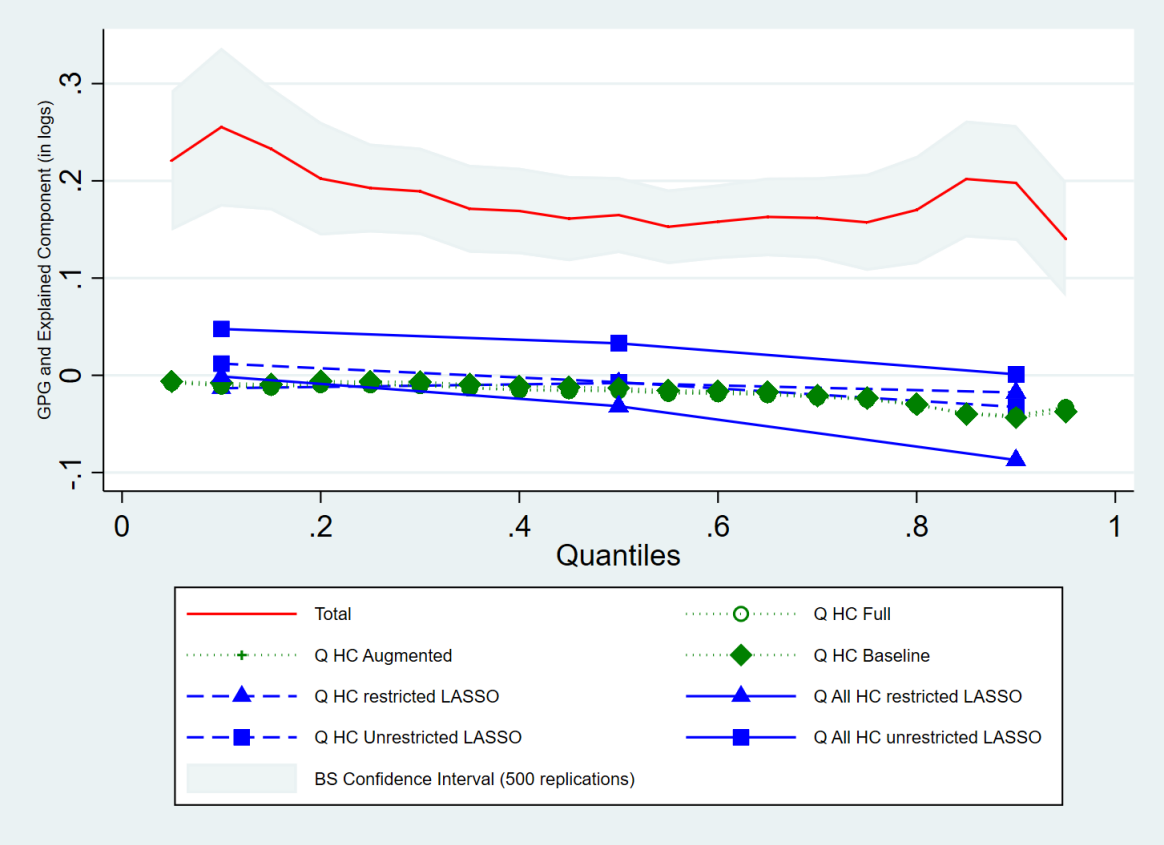

*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as its explained component ($Q$) obtained from a standard Oaxaca-Blinder decomposition of the conventional models (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). The raw gap and the explained component of the conventional models (Full, Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

post-LASSO estimation. Changing the ordering of the two samples results in two different post-double-LASSO estimators. The average of these two different estimators gives the post-double-LASSO estimator based on sample splitting (Chernozhukov et al., 2018).

Table 7 shows the estimated adjusted GPG at the mean and selected percentiles using the restricted and unrestricted LASSO specifications based on a sample split. The results of the LASSO specifications represented in Section 4 and the adjusted gap based on a the restricted-LASSO specification using sample-split differ at all percentiles and the mean by less than one percentage point (cfr. Table 5). The difference in the unrestricted LASSO is slightly more pronounced amounting to two percentage points at most at the bottom of the wage distribution.

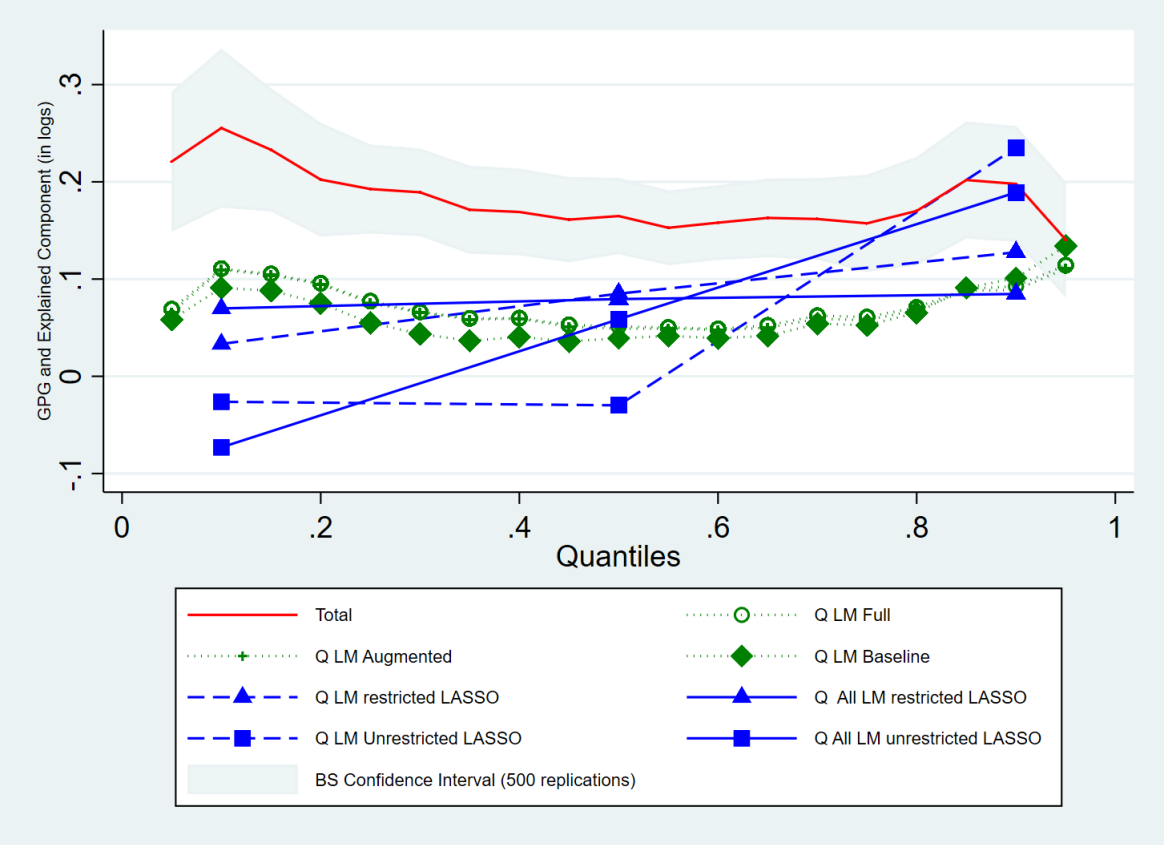
All in all, as in the estimation without sample splitting, the results based on sample

Figure 6: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Detailed Decomposition attributable to Human Capital Characteristics (Explained Component) – Male-Reference Category



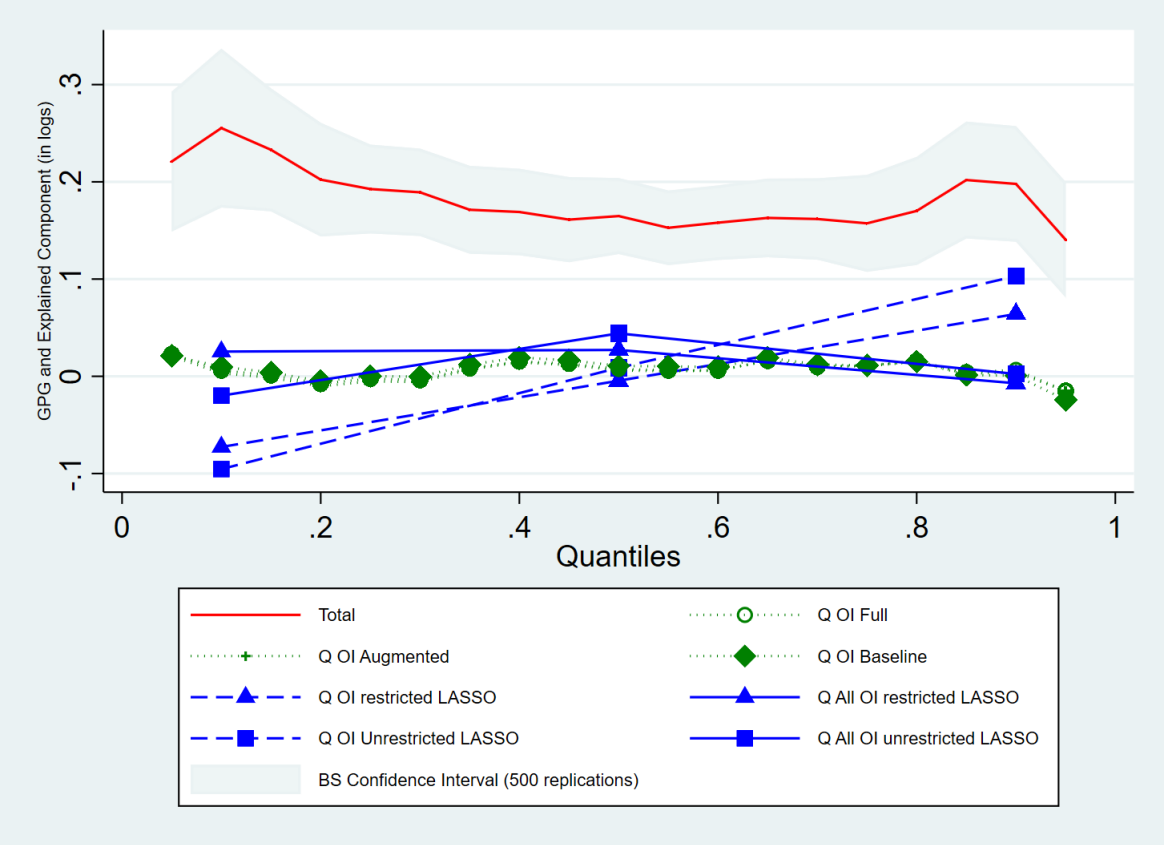

*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as the part of the explained component ($Q$) attributable to human capital characteristics ($HC$) obtained from a standard Oaxaca-Blinder decomposition of the conventional models (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). $Q$ $HC$ includes pure human capital controls, while $Q$ $HC$ $All$ includes also interactions of human capital characteristics with variables from other categories. In case of the Full and Augmented model, $HC$ includes years of education and age. In the Baseline model $HC$ includes additionally age squared. In the Restricted LASSO model, $HC$ includes years of education and quadratic polynomials of age (amelioration set) as well as further controls presented in Appendix B. $HC$ in the Unrestricted LASSO model includes the controls specified in Appendix B. The raw gap and the disaggregate explained component of the conventional models (Full, Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

Figure 7: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Detailed Decomposition attributable to Labor Market Characteristics (Explained Component) – Male-Reference Category



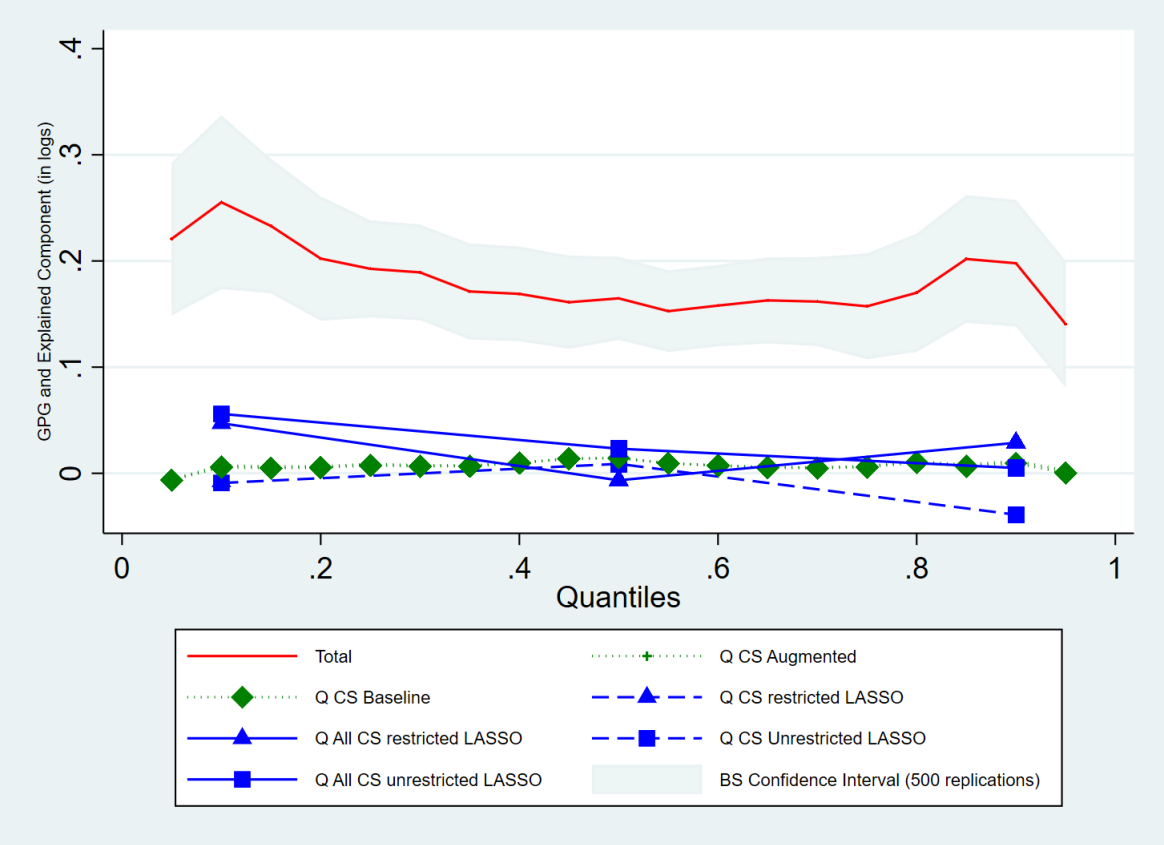

*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as the part of the explained component ($Q$) attributable to labor market characteristics ($LM$) obtained from a standard Oaxaca-Blinder decomposition of the conventional models (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). $Q\ LM$ includes pure labor market controls, while $Q\ LM\ All$ includes also interactions with labor market characteristics with variables from other categories. In case of the Full and Augmented model, $LM$ includes labor market experience. In the Baseline model $LM$ includes additionally labor market experience squared and job tenure. In the Restricted LASSO model, $LM$ includes quadratic polynomials of labor market experience and job tenure (amelioration set) as well as further controls presented in Appendix B. $LM$ in the Unrestricted LASSO model includes the controls specified in Appendix B. The raw gap and the disaggregate explained component of the conventional models (Full, Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

Figure 8: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Detailed Decomposition attributable to Industrial and Occupational Sorting (Explained Component) – Male-Reference Category




*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as the part of the explained component ($Q$) attributable to occupational and sectoral dummies ($OI$) obtained from a standard Oaxaca-Blinder decomposition in the conventional models (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). *Q OI* includes pure industrial and occupational controls only, while *Q OI All* includes also interactions with occupational or industrial dummies with variables from other categories. In case of the Full, Augmented and Baseline model, $OI$ include occupational and sectoral dummies. In the Restricted and Unrestricted LASSO models, $OI$ includes controls presented in Appendix B. The raw gap and the disaggregate explained component of the conventional models (Full, Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

Figure 9: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Detailed Decomposition attributable to Character Skills (Explained Component) – Male-Reference Category




*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as the part of the explained component ($Q$) attributable to character skills ($CS$) obtained from a standard Oaxaca-Blinder decomposition in the conventional models (augmented and baseline) (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). *Q CS* includes pure character skill controls, while *Q CS All* includes also interactions with character skill measure with variables from other categories. In case of the Augmented and Baseline models, $CS$ includes the Big Five Personality Traits, locus of control, reciprocity and willingness to take a risk. In the Restricted and Unrestricted LASSO models, $CS$ includes the controls specified in Appendix B. The raw gap and the disaggregate explained component of the conventional models (Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

splitting suggest again a U-shaped pattern over the distribution. Similarly, the restricted LASSO model suggests that the GPG is particularly pronounced at the top. The results differ slightly at the bottom. The unrestricted LASSO with sample splitting finds that the GPG is most pronounced at the bottom, while the unrestricted LASSO without sample splitting finds equally pronounced gaps at the bottom and top. However, in all cases the results of the sample split range within the 95% confidence interval of the corresponding result in Table 5. That is, our results presented in Section 4 are robust to sample splitting.

Table 7: Adjusted Gender Pay Gap (GPG) at the Mean and Selected Percentiles, Sample Split

| | (1)<br>Restricted LASSO | (2)<br>Unrestricted LASSO |
|---|---|---|
| *Panel A: At the Mean* | | |
| Adjusted Gap | -0.106*** | -0.107*** |
| | (0.015) | (0.015) |
| *Panel B: At the 10th Percentile* | | |
| Adjusted GPG | -0.130*** | -0.160*** |
| | (0.037) | (0.039) |
| *Panel C: At the 50th Percentile* | | |
| Adjusted GPG | -0.109*** | -0.111*** |
| | (0.019) | (0.020) |
| *Panel D: At the 90th Percentile* | | |
| Adjusted GPG | -0.156*** | -0.146*** |
| | (0.037) | (0.036) |
| Observations | 8,489 | 8,489 |

*Notes:* SOEP sample weights used. This table shows average adjusted gender wage gaps obtained from a linear regression of the log wage on a dummy for gender (female = 1) and sets of covariates selected by the restricted and unrestricted LASSO, respectively, on a training and test sample estimated at the individual level. Bootstrapped standard errors (500 replications, clustered at the individual level) in parentheses. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. *Source:* SOEP v33.

# 6 Conclusion

This paper analyzes the German GPG at the mean as well as along the wage distribution using linear unconditional quantile regression. We apply a machine learning technique (post-double-LASSO estimator) in order to find the appropriate set of control variables (given data restrictions) for estimation of gender differences in pay. To the best of our knowledge, this is the first paper that compares conventionally estimated adjusted GPGs to estimates of the latter based on the post-double-LASSO estimator proposed by Belloni et al. (2014a,b) at the mean as well as beyond. Additionally, we conduct a decomposition

analysis and evaluate the estimated models with respect to their robustness to remaining selection on unobservables using the method of Oster (2019). This method has not yet been applied in the context of the GPG.

We use machine learning for identification of the relevant set of controls for estimation of the (adjusted) GPG. Careful identification of the set of control variables is important in order to guard against omitted variable bias. We set-up two LASSO specifications (restricted and unrestricted) and three conventional wage models inspired by the GPG literature. The results suggest substantial differences in the estimated GPGs depending on the method used for model selection (data-driven vs conventional). For instance, the LASSO specifications suggest a U-shaped pattern of the pay gap along the distribution, while the conventional models suggest most pronounced GPGs at the top. The sets of selected controls for the data-driven models differ across the wage distribution and contain more interactions and higher-order polynomials. That is, the machine learning based model specifications are more flexible than conventional model specifications for estimation of GPGs and differ along the wage distribution.

Apart from the careful model selection, we check how robust our estimates are to remaining selection on unobservables (Altonji et al., 2005; Oster, 2019). Therefore, we calculate the degree of remaining selection on unobservables relative to selection on observables necessary to result in a zero GPG. The estimated GPGs in both the conventional and LASSO models are robust to remaining selection on unobservables. In the LASSO specifications, we find lower values for the proportionality parameters necessary to produce a zero GPG. The latter may be due to the fact that our LASSO specifications explicitly search for important gender differences (Gender-LASSO) and wage predictors (Wage-LASSO). Having the latter in mind, we conclude that the LASSO models control more appropriately for selection on observables.

Further, we decompose the GPG into an explained and an unexplained part. Gender differences in human capital characteristics explain only a negligible fraction of the differential, while gender differences in labor market characteristics are main drivers of the GPG. This result holds across all model specifications as well as at all points of the distribution. We find, however, differences between the conventional and machine learning models. For instance, the machine learning models, compared to conventional wage models, correct the explained component upwards at the top but downwards at the bottom of the wage distribution.

All in all, our results suggest the usage of different control variables at different points of the wage distribution. The post-double-LASSO estimator helps to identify the corresponding set of controls from a large set of potential regressors (5,821 variables in our case). The large number of chosen interactions and polynomials may be hard to detect for researchers. Therefore, machine learning represents a systematic and helpful tool for

model selection.

We find that the estimated gaps at the mean and median are relatively stable over all model specifications (conventional and LASSO). Thus, the conventional methods may be particularly suited for estimation at the mean and median – and less for the top and bottom of the distribution. As stated, the LASSO specifications suggest that the estimated GPGs at the top are upward-biased, while those at the bottom are downward-biased in conventional wage models. This result may be relevant for policy implications. For example, instead of fighting mainly the top GPG, our results suggest that the gaps in both tails are especially pronounced and should be considered equally in political measures to close the gap.

## References


Albrecht, J., A. Björklund, and S. Vroman (2003): "Is there a Glass Ceiling in Sweden?" *Journal of Labor Economics*, 21, 145–177.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011): "Personality Psychology and Economics," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, S. Machin, and L. Wössmann, Amsterdam: Elsevier Science, vol. 4, 1–181.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151–184.

Arulampalam, W., A. Booth, and M. Bryan (2007): "Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wages Distribution," *Industrial and Labor Relations Review*, 60, 163–186.

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

Bach, P., V. Chernozhukov, and M. Spindler (2018): "Closing the US gender wage gap requires understanding its heterogeneity," *arXiv preprint arXiv:1812.04345*.

Bailey, M., B. Hershbein, and A. Miller (2012): "The opt-in revolution? Contraception and the gender gap in wages," *American Economic Journal: Applied Economics*, 225–254.

Becker, G. S. (1957): *The economics of discrimination*, Chicago: University of Chicago Press.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A. and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28, 29–50.

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650.

Bertrand, M. (2011): *New perspectives on gender*, Elsevier, vol. 4, 1543–1590.

Blau, F. D. and L. Kahn (2006): "The US Gender Pay Gap in the 1990s: Slowing Convergence," *Industrial Labor Relations Review*, 45–66.

Blau, F. D. and L. M. Kahn (2017): "The gender wage gap: Extent, trends, and explanations," *Journal of Economic Literature*, 55, 789–865.

Blinder, A. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Braakmann, N. (2009): "The role of psychological traits for the gender gap in full-time employment and wages: evidence from Germany," *SOEP Paper No. 162*.

Brenzel, H. and M.-C. Laible (2016): "Does personality matter? The impact of the Big Five on the migrant and gender wage gaps," IAB-Discussion Paper 26/2016, Nürnberg.

Castagnetti, C., L. Rosti, and M. Töpfer (2018): "Discriminate Me – if You Can! The Disappearance of the Gender Pay Gap among Public-Contest Selected Employees," FAU-Discussion Paper 103, Chair of Labor and Regional Economics.

Cattan, S. (2013): "Psychological traits and the gender wage gap," *The Institute for Fiscal Studies*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Cobb-Clark, D. A. and S. Schurer (2012): "The stability of big-five personality traits," *Economics Letters*, 115, 11–15.

——— (2013): "Two economists' musings on the stability of locus of control," *The Economic Journal*, 123.

Coffman, K. B., C. L. Exley, and M. Niederle (2018): *When gender discrimination is not about gender*, Harvard Business School.

Collischon, M. (2017): "The Returns to Personality Traits across the Wage Distribution," *SOEPpapers 912*.

Cotton, J. (1988): "On the Decomposition of Wage Differentials," *The Review of Economics and Statistics*, 70, 236–43.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1044.

Eurostat (2018): "Gender Pay Gap Statistics," http://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_pay_gap_statistics, accessed: 2018-10-26.

Firpo, S., N. M. Fortin, and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, 77, 953–973.

Fortin, N., T. Lemieux, and S. Firpo (2011): "Decomposition Methods in Economics," in *Handbook of labor economics*, Elsevier, vol. 4(A), 1–102.

Fortin, N. M. (2008): "The gender wage gap among young adults in the united states the importance of money versus people," *Journal of Human Resources*, 43, 884–918.

Frank, L. E. and J. H. Friedman (1993): "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Gardeazabal, J. and A. Ugidos (2004): "More on Identification in Detailed Wage Decompositions," *Review of Economics and Statistics*, 86, 1034–1036.

Gelbach, J. B. (2002): "Identified Heterogeneity in Detailed Wage Decompositions," University of Maryland at College Park.

Gensowski, M. (2018): "Personality, IQ, and lifetime earnings," *Labour Economics*, 51, 170–183.

Gerlitz, J.-Y. and J. Schupp (2005): "Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP," DIW Research Notes # 26, German Institute for Economic Research (DIW), Berlin.

Goldin, C. (2006): "The Quiet Revolution that Transformed Women's Employment, Education, and Family," *American Economic Review*, 96, 1–21.

——— (2014): "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104, 1091–1119.

Hastie, T., R. Tibshirani, and J. Friedman (2009): "Linear Methods for Regression," in *The elements of statistical learning*, Springer, 43–100.

Heineck, G. and S. Anger (2010): "The returns to cognitive abilities and personality traits in Germany," *Labour economics*, 17, 535–546.

Heinze, A. and E. Wolf (2010): "The intra-firm gender wage gap: a new view on wage differentials based on linked employer–employee data," *Journal of Population Economics*, 23, 851–879.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An introduction to statistical learning*, Springer.

Juhn, C. and K. McCue (2017): "Specialization then and now: Marriage, children, and the gender earnings gap across cohorts," *Journal of Economic Perspectives*, 31, 183–204.

Mandel, H. and M. Semyonov (2014): "Gender Pay Gap and Employment Sector: Sources of Earnings Disparities in the United States, 1970–2010," *Demography*, 51, 1597–1618.

Mas, A. and A. Pallais (2017): "Valuing alternative work arrangements," *American Economic Review*, 107, 3722–59.

Mincer, J. A. (1974): *Schooling, Experience, and Earnings*, NBER Books, Columbia University Press.

Mueller, G. and E. Plug (2006): "Estimating the effect of personality on male and female earnings," *ILR Review*, 60, 3–22.

Mullainathan, S. and J. Spiess (2017): "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 31, 87–106.

Neumark, D. (1988): "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," *Journal of Human Resources*, 23, 279–295.

Nyhus, E. K. and E. Pons (2005): "The effects of personality on earnings," *Journal of Economic Psychology*, 26, 363–384.

——— (2012): "Personality and the gender wage gap," *Applied Economics*, 44, 105–118.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

Oaxaca, R. and M. R. Ransom (1994): "On Discrimination and the Decomposition of Wage Differentials," *Journal of Econometrics*, 61, 5–21.

——— (1999): "Identification in Detailed Wage Decompositions," *Review of Economics and Statistics*, 81, 154–157.

Oster, E. (2019): "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business and Economic Statistics*, 37, 187–204.

Risse, L., L. Farrell, and T. R. Fry (2018): "Personality and pay: do gender gaps in confidence explain gender gaps in wages?" *Oxford Economic Papers*, 70, 919–949.

Rotter, J. B. (1966): "Generalized expectancies for internal versus external control of reinforcement," *Psychological monographs: General and applied*, 80, 1.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Wagner, G. G., J. Göbel, P. Krause, R. Pischner, and I. Sieber (2008): "Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland–Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender)," *AStA Wirtschafts-und Sozialstatistisches Archiv*, 2, 301–328.

Weisberg, Y. J., C. G. DeYoung, and J. B. Hirsh (2011): "Gender differences in personality across the ten aspects of the Big Five," *Frontiers in Psychology*, 2, 178.

Wilde, E. T., L. Batchelder, and D. T. Ellwood (2010): "The mommy track divides: The impact of childbearing on wages of women of differing skill levels," Tech. rep., National Bureau of Economic Research.

Yun, M.-S. (2005): "A Simple Solution to the Identification Problem in Detailed Wage Decompositions," *Economic Inquiry*, 43, 766–772.

# Appendix

# A Details on Character Skill Measures

## A.1 Measures for Character Skills

The Big Five Personality Traits are a widely used approach to measure personality. Personality is organized hierarchically, with five higher-level factors and multiple lower-level facets associated with each higher-level factor (Almlund et al., 2011). The five higher-level factors – the so-called Big Five Personality Traits – are (1) openness, (2) conscientiousness, (3) extraversion, (4) agreeableness and (5) neuroticism. Conscientiousness reflects the tendency of an individual to be hardworking, organized and responsible. Openness in turn refers to a person's openness towards new cultural, intellectual or aesthetic experiences. Extraversion measures the degree of outward orientation of an individual. Agreeableness describes the tendency of an individual to cooperate with others in an unselfish way. Finally, neuroticism refers to chronic levels of emotional instability and the tendency to suffer from psychological distress (Almlund et al., 2011).

Since 2005, the SOEP individual questionnaire contains 15 items on the Big Five, three for each higher-level factor. Respondents are asked to rate 15 statements using a seven-point Likert scale, where a one refers to complete disagreement and a seven to complete agreement (Gerlitz and Schupp, 2005). Table A.1 shows in the first column, the 15 different statements and in the second column the corresponding Big Five factor. Some statements are positively related with the factor of interest, while others are negatively related. Positive relations are indicated by (+) and negative relations by (–). So far, the questions have been part of the questionnaires in 2005, 2009 and 2013.

The concept of Locus of Control goes back to Rotter (1966). It captures individual believes about the relation between own behavior and its consequences. If an individual believes that the events in his or her life are caused by his or her individual actions, it has an internal locus of control. Thus, individuals with an internal locus of control believe that they are responsible for the things that happen in their life, while individuals with an external locus of control hold others responsible for their life events (Almlund et al., 2011). As in the case of the Big Five the questions concerning Locus of Control are included in the SOEP individual questionnaire in the years 2005, 2010, and 2015. Respondents are asked to rate eight statements on the same seven-point Likert scale as in case of the Big Five. Table A.2 shows the eight different statements and assigns them either to external or internal Locus of Control. Each Locus of Control measure is based on four statements.

Reciprocity measures the way in which individuals react to other people's behavior. Positive reciprocity corresponds to a tendency to reward kind actions, while negative reciprocity is present if the individual tends to punish other individuals for their unkind

actions (Almlund et al., 2011). The six corresponding items and their assignment to either positive or negative reciprocity are shown in Table A.3. The statements, which where included in the individual questionnaires in 2005, 2010 and 2015, had to be rated, again, on the same seven-point Likert scale as in case of the Big Five and Locus of Control.

The questions to measure Willingness to take a risk are included in the SOEP individual questionnaire in 2004, 2009, and 2014. Respondents are asked to answer the six questions given in Table A.4 based on a 11-point Likert scale, where zero refers to 'risk averse' and ten to 'fully prepared to take risks'.

In total, we have thus ten factors for which we construct indices for our analysis. We build the different indices as the averages of the corresponding items (see Section A.2 for detailed list of the survey questions).[18] Finally, we standardize all measures to allow for effect comparison. The items corresponding to Internal Locus of Control are associated with a Cronbach's $\alpha$ of only 0.28. Therefore, we exclude the index of Internal Locus of Control from the analysis. The Cronbach's $\alpha$ values of the other factors range from 0.52 to 0.82. Our result for the Cronbach's $\alpha$ are similar to those of other studies using SOEP Data that also exclude internal locus of control from the analysis (see Collischon, 2017; Heineck and Anger, 2010).

[18]Further, we use a factor analysis to check for each character skill measure whether the factor loadings are in line with the classifications presented in Tables A.1 - A.4. In case of all measures the choice of the corresponding items is confirmed.

## A.2 Survey Questions Related to Character Skill Measures

Table A.1: SOEP Survey Questions Corresponding to the Big Five Personality Traits

| Content | Factor |
|---|---|
| I am communicative, talkative (+) | Extraversion |
| I am outgoing, sociable(+) | Extraversion |
| I am reserved (–) | Extraversion |
| | |
| I do a thorough job (+) | Conscientiousness |
| I tend to be lazy (–) | Conscientiousness |
| I carry out duties efficiently and effectively (+) | Conscientiousness |
| | |
| I come up with new ideas (+) | Openness |
| I value aesthetic and artistic experiences (+) | Openness |
| I have a lively imagination (+) | Openness |
| | |
| I am sometimes somewhat rude to others (–) | Agreeableness |
| I can forgive others (+) | Agreeableness |
| I am considerate and kind to others (+) | Agreeableness |
| | |
| I worry a lot (+) | Neuroticism |
| I get nervous easily (+) | Neuroticism |
| I am relaxed, I handle stress well (–) | Neuroticism |

Source: SOEP Questionnaire.
*Notes:* (+) indicates a positive relation. (–) indicates a negative relation.

Table A.2: SOEP Survey Questions Corresponding to Locus of Control

| Content | Factor |
|---|---|
| What a person achieves in life is above all a question of fate or luck | External Locus of Control |
| I frequently have the experience that other people have a controlling influence over my life | External Locus of Control |
| The opportunities that I have in life are determined by the social conditions | External Locus of Control |
| I have little control over the things that happen in my life | External Locus of Control |
| Inborn abilities are more important than any efforts one can make | External Locus of Control |
| One has to work hard in order to succeed | Internal Locus of Control |
| If I run up against difficulties in life, I often doubt my own abilities | Internal Locus of Control |
| How my life goes depends on me | Internal Locus of Control |

Source: SOEP Questionnaire.

Table A.3: SOEP Survey Questions Corresponding to Reciprocity

| Content | Factor |
|---|---|
| If someone does me a favor, I am prepared to return it | Positive Reciprocity |
| I go out of my way to help somebody who has been kind to me before | Positive Reciprocity |
| I am ready to undergo personal costs to help somebody who helped me before | Positive Reciprocity |
| If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost | Negative Reciprocity |
| If somebody puts me in a difficult position, I will do the same to him/her | Negative Reciprocity |
| If somebody offends me, I will offend him/her back | Negative Reciprocity |

Source: SOEP Questionnaire.

Table A.4: SOEP Survey Questions Corresponding to Willingness to take a Risk

People can behave differently in different situations. How would you rate your willingness to take risks in the following areas?

- while driving?
- in financial matters?
- during leisure and sport?
- in your occupation?
- with your health?
- your faith in other people?

Source: SOEP Questionnaire.

# B Detailed Set of Selected Controls (LASSO Specifications)

In the following we show the detailed sets of selected controls of the restricted (Section B.1) and unrestricted LASSO specifications (Section B.2). We divide the variables in seven sets of controls and present the selected variables in the following order: Human Capital, Labor Market, Demographic, Character Skill, Occupational and Industrial, Firm as well as Survey Year Controls.

The set of control variables 'Human Capital' may include – depending on the choice of the LASSO – educational attainment, degree, grade in last record and a control for Mathematics, Informatics, Natural Science or Technology (MINT) subject and age. We refer to labor market experience, job tenure, years of parental leave, past periods of unemployment or as part-time as 'Labor Market' controls. We define as 'Demographic' controls: migration status, urban residence, residence $< 10$ km to city centre, federal state, being married or cohabiting and whether the partner ID was actually reported and information on parents' education. Further, dummies for different religion, the number of children and young children as well as asset flows and regions of origin can be chosen in this set. 'Character Skill' controls include the Big Five Personality Traits (openness, conscientiousness, agreeableness and neuroticism) as well as risk aversion, positive and negative reciprocity and external locus of control. Additionally, we attribute work or life satisfaction controls to the set of character skills. 'Occupational and Industrial' controls are based on ISCO88 (1-digit) and NACE (2-digit), respectively. 'Firm' controls include dummies for works council, civil servant, firm size (small, medium or large), having a permanent contract, and being a union member. 'Survey Year' dummies may include 2009 or 2013. In each of these seven sets of variables, we include the above mentioned pure controls as well as interactions within a set of variables. For example, marital status and urban residence belong both to the set of demographic controls and thus also interactions among them are attributed to this set. These sets of controls are highlighted as 'Pure Sets of Controls' in Tables B.1 - B.4, while interactions across different sets of controls are highlighted as 'Interactions of Different Sets of Controls'.

## B.1 Restricted LASSO

Table B.1: Set of Selected Controls, Restricted Gender-LASSO

*Pure Sets of Controls*

*Human Capital Controls*

Studied MINT Subject X German Grade in Last Record, Highest Degree *Fachhochschule* X Studied MINT Subject, Math Grade in Last Record X Highest Degree High School

*Labor Market Controls*

Parental Leave (Years), Labor Market Experience (Years) X Part-time (Years)

*Demographic Controls*

Partner ID Reported X Number of Children, Married or Cohabiting X Number of Children, Interviewer Same Sex X Saxony-Anhalt

*Occupational and Industrial Controls*

Technicians, Craftsmen, Technicians X Educational Sector , Technicians X Health and Social Work, Service Occupations X Health and Social Work, Craftsmen X Construction

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Labor Market Experience (Years) X German Grade in Last Record

*Human Capital Interacted with Demographic Controls*

Interviewer Same Sex X German Grade in Last Record

*Human Capital Interacted with Character Skill Controls*

Conscientiousness X Math Grade in Last Record, Conscientiousness X Education (Years), Extraversion X Education (Years), Neurocitism X Education (Years), Risk Aversion X Education (Years), Negative Reciprocity X Education (Years), Math Grade in Last Record X Neurocitism

*Human Capital Interacted with Occupational and Industrial Controls*

Transport, Storage and Communication X Highest Degree High School, Education X Math Grade in Last Record, Technicians X Highest Degree *Realschule*, Office Workers X Highest Degree *Realschule*, Craftsmen X Highest Degree *Realschule*, Machine Operator X German Grade in Last Record, Machine Operator X Math Grade in Last Record, Math Grade in Last Record X Construction, Education (Years) squared X Machine Operator, German Grade in Last Record squared X Machinery and Equipment, Math Grade in Last Record squared X Basic Metals and Fabricated Metal Production,

*Labor Market Interacted with Demographic Controls*

Number of Children X Part-time (Years), Number of Children X Labor Market Experience (Years), Married or Cohabiting X Part-time (Years)

*Labor Market Interacted with Character Skill Controls*

Openness X Job Tenure (Years)

*Labor Market Interacted with Occupational and Industrial Controls*

Health and Social Work X Labor Market Experience (Years), Office Workers X Labor Market Experience (Years), Office Workers X Education (Years), Machine Operator X Labor Market Experience (Years))

*Demographic Interacted with Character Skill Controls*

Partner ID Reported X Risk Aversion, Saxony X Neurocitism, Mother holds High School Degree X Conscientiousness, Close to City Centre ($< 10km$) X Conscientiousness

*Demographics Interacted with Occupational and Industrial Controls*

Health and Social Work X Sun Hours at Interview Day, Father with High School Degree X Health and Social Work, Mother holds High School Degree X Office Workers, Close to City Centre ($< 10km$) X Transport, Storage and Communication, Interviewer Same Sex X Machine Operator, Real Estate, Renting and Business Activities X Baden-Wuerttemberg

*Demographic Interacted with Firm Controls*

Permanent Contract X German Grade in Last Record

*Character Skill Interacted with Occupational and Industrial Controls*

Education X Agreeableness, Agreeableness squared X Educational Sector, External Locus of Control squared X Craftsmen, Negative Reciprocity squared X Educational Sector

*Occupational and Industrial Interacted with Firm Controls*

Health and Social Work X Small Firm, Craftsmen X Small Firm, Civil Servant X Negative Reciprocity, Technicians X Civil Servant

---

Table shows the selected set of controls from the restricted Gender-LASSO. In the restricted LASSO specification, additional to the variables in this table, the variables of the amelioration set are included. The amelioration set contains second-order polynomials of age (human capital) and labor market experience (labor market characteristic) as well as education (human capital), job tenure (labor market characteristic), federal and wave dummies, information on urban residence, marital status, partner id (demographics), union membership, contract type and firm size (firm controls). The set of control variables 'Human Capital' may include additionally grade in last record (German and Maths), a dummy for MINT subject and the school-leaving degree: *Fachhochschule* (University of Applied Sciences), *Abitur* (high-school degree) and *Realschule* (type of secondary school in Germany). Additional Labor market controls include years of parental leave, past periods of unemployment or as part-time as labor market controls. Demographic controls may additionally include residence $<$ 10 km to city centre, federal state, religion, information on parents' education, number of children, presents of small children in household as well as sun hours at interview day and the sex of the interviewer. Character skills include openness, conscientiousness, extraversion, agreeableness, neuroticism (Big Five Personality Traits), risk aversion, positive and negative reciprocity, external locus of control as well as life and work satisfaction controls. Occupational and industrial dummies are defined according to ISCO88 (1-digit) and NACE (2-digit). Note that the set of potential controls includes also two-way interactions of these control variables as well as quadratic polynomials of non-dummies.

Table B.2: Set of Selected Controls, Restricted Wage-LASSO

**At the mean:**

*Pure Sets of Controls*

*Labor Market Controls:*

Labor Market Experience (Years) X Unemployment (Years)

*Character Skill Controls:*

Risk Aversion, Negative Reciprocity squared X Openness squared

*Occupational and Industrial Controls:*

Service Occupations, Technicians X Transport Equipment

*Firm Controls:*

Large Firm X Civil Servant

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Demographic Controls:*

Number of Children X Education (Years), Household holds Assets X Education (Years)

*Human Capital Interacted with Firm Controls:*

Permanent Contract X Studied MINT Subject

*Human Capital Interacted with Occupational and Industrial Controls:*

Wholesale and Retail Trade X Education (Years), Education X Highest Degree *Abitur*, Coke and Chemicals X Education (Years), German Grade in Last Record squared X Office Workers, Education (Years) squared X Assistants, Public Administration X Highest Degree *Abitur*, Office Workers X Eduction (Years), Craftsmen X Highest Degree *Realschule*, Education (Years) squared X Machine Operator, , Scientists X Age (Years), Service Occupations X Age (Years)

*Labor Market Interacted with Demographic Controls:*

Married or Cohabiting X Unemployment (Years), Household holds Assets X Labor Market Experience (Years)

*Labor Market Interacted with Character Skill Controls:*

Work Satisfaction squared X Unemployment (Years)

*Labor Market Interacted with Occupational and Industrial Controls:*

Machinery and Equipment X Labor Market Experience (Years), Office Workers X Unemployment (Years)

*Demographic Interacted with Character Skill Controls:*

Lives in Urban Area X Agreeableness

*Demographic Interacted with Occupational and Industrial Controls:*

Migration Background X Scientists, Father with High School Degree X Machinery and Equipment, Household holds Assets X Coke and Chemicals

*Demographic Interacted with Firm Controls:*

Household holds Assets X Permanent Contract, Lives in Urban Area X Civil Servant

*Character Skill Interacted with Occupational and Industrial Controls:*

Openness squared X Machine Operator, Neurocitism squared X Assistants

*Character Skill Interacted with Firm Controls:*

Permanent Contract X External Locus of Control, Permanent Contract X Risk Aversion, Permanent Contract X Work Satisfaction

*Occupational and Industrial Interacted with Firm Controls:*

Transport Equipment X Large Firm, Financial Intermediation X Permanent Contract, Service Occupations X Permanent Contract, Assistants X Permanent Contract

**At the 10th Percentile**

*Pure Sets of Controls*

*Labor Market Controls:*

Unemployment (Years)

*Occupational and Industrial Controls*

Educational Sector, Technicians, Service Occupations, Service Occupations X Public Administration and Defense, Technicians X Educational Sector

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Math Grade in Last Record X Unemployment (Years), Highest Degree *Realschule* X Unemployment (Years)

*Human Capital Interacted with Demographic Controls*

Academic Education X Catholic, Highest Degree *Abitur* X Catholic, Baden-Wuerttemberg X Academic Education, Bavaria X Highest Degree *Abitur*

*Human Capital Interacted with Occupational and Industrial Controls*

Educational Sector X Highest Degree *Realschule*, Craftsmen X Highest Degree High School, Math Grade in Last Record X Educational Sector, Technicians X Age (Years)

*Human Capital Interacted with Firm Controls*

Highest Degree *Realschule* X Civil Servant

*Labor Market Interacted with Occupational and Industrial Controls*

Service Occupations X Unemployment (Years)

*Labor Market Interacted with Firm Controls*

Part-time (Years) squared X Works Council

*Demographics Interacted with Occupational and Industrial Controls*

Saxony-Anhalt X Scientists, Mother holds *Realschul*-Degree X Scientists, Baden-Wuerttemberg X Real Estate

*Demographic Interacted with Firm Controls*

Lives in Urban Area X Small Firm, North-Rhine Westphalia X Works Council, Saxony X Large Firm, Saxony X Works Council, Saxony-Anhalt X Civil Servant

*Character Skill Interacted with Occupational and Industrial Controls*

Service Occupations X Risk Aversion, Risk Aversion squared X Service Occupations

*Occupational and Industrial Interacted with Firm Controls*

Service Occupations X Small Firm, Technicians X Civil Servants

---

**At the 50th Percentile**

*Pure Sets of Controls*

*Demographic Controls*

Number of Children

*Character Skill Controls*

External Locus of Control, Risk Aversion

*Occupational and Industrial Controls*

Scientists, Service Occupations, Technicians X Transport Equipment, Service Occupation X Health and Social Work

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Occupational and Industrial Controls*

Wholesale and Retail Trade X German Grade in Last Record, Assistants X Education (Years), Machine Operator X Education (Years) squared

*Labor Market Interacted with Occupational and Industrial Controls*

Electricity Gas and Water Supply X Job Tenure (Years), Manufacturing of Textiles or Leather, or of Wood X Job Tenure (Years), Scientists X Part-time (Years), Office Workers X Unemployment (Years), Assistants X Labor Market Experience (Years)

*Demographics Interacted with Occupational and Industrial Controls*

Financial Intermediation X Sun Hours at Interview Day, Father with High School Degree X Machinery and Equipment

*Demographics Interacted with Firm Controls*

Household holds Assets X Permanent Contract

*Character Skill Interacted with Firm Controls*

Permanent Contract X Risk Aversion

*Occupational and Industrial Interacted with Firm Controls*

Electricity Gas and Water Supply X Works Council, Wholesale and Retail Trade X Large Firm, Financial Intermediation X Permanent Contract, Service Occupations X Permanent Contract, Service Occupations X Large Firm, Assistants X Permanent Contract, Assistants X Works Council, Assistants X Civil Servant

---

**At the 90th Percentile**

*Pure Sets of Controls*

*Labor Market Controls*

Labor Market Experience (Years) X Part-time (Years)

*Demographic Controls*

Household holds Assets X Lives in Urban Area

*Occupational and Industrial Controls*

Machine Operator X Basic Metals and Fabricated Metal Production

*Firm Controls*

Large Firm X Civil Servant, Works Council X Civil Servant

---

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Demographic Controls*

Studied MINT Subject X Number of Children, Academic Education X Number of Children, Berlin X Academic Education, Education (Years) squared X Household holds Assets

*Human Capital Interacted with Character Skill Controls*

Academic Education X Life Satisfaction

*Human Capital Interacted with Occupational and Industrial Controls*

Electrical and Optical Equipment X Academic Education, Public Administration and Defence X Academic Education, Public Administration and Defence X Highest Degree *Abitur*, Machine Operator X Education (Years) squared

*Human Capital Interacted with Firm Controls*

Civil Servant X Education (Years), Academic Education X Civil Servant, Education (Years) squared X Civil Servant, Education (Years) squared X Large Firm, Education (Years) squared X Large Firm

*Labor Market with Demographic Controls*

Saxony X Labor Market Experience (Years), Household holds Assets X Labor Market Experience (Years)

*Labor Market with Character Skill Controls*

External Locus of Control X Labor Market Experience (Years), Risk Aversion X Labor Market Experience (Years)

*Labor Market Interacted with Occupational and Industrial Controls*

Craftsmen X Labor Market Experience (Years)

*Labor Market Interacted with Firm Controls*

Civil Servant X Job Tenure (Years)

*Demographic Interacted with Character Skill Controls*

Household holds Assets X External Locus of Control

*Demographic Interacted with Occupational and Industrial Controls*

Machine Operator X Partner ID Reported, Married or Cohabiting X Public Administration and Defence, Married or Cohabiting X Craftsmen, Married or Cohabiting X Machine Operator, Married or Cohabiting X Assistants, Migration Background X Basic Metals and Fabricated Metal Production

*Demographic Interacted with Firm Controls*

Civil Servant X Partner ID Reported

*Character Skill Interacted with Occupational and Sectoral Controls*

Agreeableness squared X Machine Operator

*Occupational and Industrial Interacted with Firm Controls*

Public Administration and Defence X Large Firm, Scientists X Large Firm, Craftsmen X Large Firm, Assistants X Large Firm, Assistants X Works Council

Table shows restricted Wage-LASSO at the mean, 10th, 50th, and 90th percentile, respectively. In the restricted LASSO specification, additional to the variables in this table, the variables of the amelioration set are included. The amelioration set contains second-order polynomials of age (human capital) and labor market experience (labor market characteristic) as well as education (human capital), job tenure (labor market characteristic), federal and wave dummies, information on urban residence, marital status, partner id (demographics), union membership, contract type and firm size (firm controls). The set of control variables 'Human Capital' may include additionally grade in last record (German and Maths), a dummy for MINT subject and the school-leaving degree: *Fachhochschule* (University of Applied Sciences), *Abitur* (high-school degree) and *Realschule* (type of secondary school in Germany). Additional Labor market controls include years of parental leave, past periods of unemployment or as part-time as labor market controls. Demographic controls may additionally include residence $< 10$ km to city centre, federal state, religion, information on parents' education, number of children, presents of small children in household as well as sun hours at interview day and the sex of the interviewer. Character skills include openness, conscientiousness, extraversion, agreeableness, neuroticism (Big Five Personality Traits), risk aversion, positive and negative reciprocity, external locus of control as well as life and work satisfaction controls. Occupational and industrial dummies are defined according to ISCO88 (1-digit) and NACE (2-digit). Note that the set of potential controls includes also two-way interactions of these control variables as well as quadratic polynomials of non-dummies.

## B.2 Unrestricted LASSO

Table B.3: Set of Selected Controls, Unrestricted Gender-LASSO

*Pure Sets of Controls*

*Human Capital Controls*

German Grade in Last Record, Studied MINT Subject X German Grade in Last Record, Highest Degree *Fachhochschule* X Studied MINT Subject, Highest Degree *Realschule* X Academic Education

*Labor Market Controls*

Part-time (Years), Parental Leave (Years), Labor Market Experience (Years) X Part-time (Years)

*Character Skill Controls:*

Risk Aversion

*Demographic Controls*

Brandenburg X Number of Children, Married or Cohabiting X Number of Children, Interviewer Same Sex X Saxony-Anhalt, Married or Cohabiting X Catholic

*Occupational and Industrial Controls*

Technicians, Craftsmen, Technicians X Educational Sector, Technicians X Health and Social Work, Service Occupations X Health and Social Work, Craftsmen X Construction, Office Workers X Real Estate, Renting and Business Activities

*Firm Controls*

Union Member X Large Firm

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Labor Market Experience (Years) X German Grade in Last Record, Maths Grade in Last Record X Part-time (Years) Highest Degree High School X Labor Market Experience (Years), Highest Degree *Realschule* X Part-time (Years)

*Human Capital Interacted with Demographic Controls*

Interviewer Same Sex X German Grade in Last Record, Highest Degree *Fachhochschule* X Married or Cohabiting

*Human Capital Interacted with Character Skill Controls*

Conscientiousness X Education (Years), Extraversion X Education (Years), Neurocitism X Education (Years), Risk Aversion X Education (Years), Negative Reciprocity X Education (Years), Math Grade in Last Record X Neurocitism

*Human Capital Interacted with Occupational and Industrial Controls*

German Grade in Last Record squared X Machinery and Equipment, Maths Grade in Last Record squared X Basic Metals and Fabricated Metal Production, Education (Years) squared X Machine Operator, German Grade in Last Record X Basic Metals and Fabricated Metal Production, Construction X Highest Degree High School, Transport, Storage and Communication X Highest Degree High School, Educational Sector X Math Grade in Last Record, Health and Social Work X Highest Degree *Realschule*, Technicians X Highest Degree *Realschule*, Office Workers X Highest Degree *Realschule*, Office Workers X Education (Years), Craftsmen X Highest Degree *Realschule*, Machine Operator X German Grade in Last Record, Machine Operator X Math Grade in Last Record, Math Grade in Last Record X Construction

*Human Capital Interacted with Firm Controls*

German Grade in Last Record X Union Member

*Labor Market Interacted with Demographic Controls*

Number of Children X Part-time (Years), Number of Children X Labor Market Experience (Years)

*Labor Market Interacted with Character Skill Controls*

Openness X Job Tenure (Years)

*Labor Market Interacted with Occupational and Industrial Controls*

Health and Social Work X Labor Market Experience (Years), Health and Social Work X Job Tenure (Years), Office Workers X Education (Years), Machine Operator X Labor Market Experience (Years), Wholesale and Retail Trade X Part-time (Years), Technicians X Unemployment (Years)

*Demographic Interacted with Character Skill Controls*

Close to City Centre ($< 10km$) X Conscientiousness

*Demographics Interacted with Occupational and Industrial Controls*

Health and Social Work X Sun Hours at Interview Day, Renting and Business Activities, Father with High School Degree X Health and Social Work, Migration Background X Health and Social Work, Lives in Urban Area X Machine Operators, Baden-Wuerttemberg X Real Estate, Renting and Business Activities

*Demographic Interacted with Firm Controls*

Partner ID Reported X Union Member, Interviewer Same Sex X Large Firm, Interviewer Same Sex X Union Member, Civil Servant X Negative Reciprocity

*Character Skill Interacted with Occupational and Industrial Controls*

Educational Sector X Agreeableness, Agreeableness squared X Educational Sector, External Locus of Control squared X Craftsmen, Negative Reciprocity squared X Educational Sector, Negative Reciprocity squared X Health and Social Work

*Occupational and Industrial Interacted with Firm Controls*

Health and Social Work X Small Firm, Craftsmen X Small Firm, Office Workers X Small Firm, Technicians X Civil Servant

Table shows the selected set of controls from the unrestricted Gender-LASSO. In case of the unrestricted LASSO specification no variables are forced to be part of the model. Thus, the LASSO selects among the full set of potential controls. The set of control variables 'Human Capital' may include – depending on the choice of the LASSO – educational attainment, age, degree, grade in last record and a control for Mathematics, Informatics, Natural Science or Technology (MINT) subject. We refer to labor market experience, job tenure, years of parental leave, past periods of unemployment or as part-time as labor market controls. We define as demographic controls: migration status, urban residence, residence $< 10$ km to city centre, federal state, being married or cohabiting an whether the partner ID was actually reported and information on parents' education, number of children, presents of small children in household, region of origin, dummies for different religions as well as sun hours at interview day and the sex of interviewer. Character skill controls include the Big Five Personality Traits (openness, conscientiousness, agreeableness and neuroticism) as well as risk aversion, positive and negative reciprocity and external locus of control. The attribute work or life satisfaction controls to the set of character skills. Occupational and industrial controls are based in ISCO88 (1-digit) and NACE (2-digit). Firm controls include dummies for works council, civil servant, firm size (medium or large), having a permanent contract, being a union member. Survey years dummies may include 2009 or 2013. *Fachhochschule* is a University of Applied Sciences, *Abitur* a high-school degree and *Realschule* a type of secondary school in Germany.

Table B.4: Set of Selected Controls, Unrestricted LASSO at the Mean

**At the mean**

*Pure Sets of Controls*

*Human Capital Controls*

Education (Years), Studied MINT Subject X German Grade in Last Record,

*Labor Market Controls*

Part-time (Years), Unemployment (Years)

*Demographic Controls*

Mother holds High School Degree X Baden-Wuerttemberg, Interviewer Same Sex X Berlin, Interviewer Same Sex X Brandenburg, Number of Children X Bavaria, Hesse X Lives in Urban Area

*Character Skill Controls*

External Locus of Control, Risk Aversion

*Occupational and Industrial Controls*

Wholesale and Retail Trade, Scientists, Office Workers, Service Occupations, Technicians X Transport Equipment, Technicians X Health and Social Work

*Firm Controls*

Small Firm, Civil Servant X Large Firm,

*Survey Year Controls*

2013

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Job Tenure (Years) X German Grade in Last Record, Highest Degree *Realschule* X Unemployment (Years)

*Human Capital Interacted with Character Skill Controls*

Education (Years) X Life Satisfaction, Agreeableness X Education (Years), Risk Aversion X German Grade in Last Record

*Human Capital Interacted with Occupational and Industrial Controls*

Education (Years) squared X Machine Operator, Education (Years) squared X Coke and Chemicals, Coke and Chemicals X Education (Years), Machinery and Equipment X Education (Years), Public Administration and Defence X Highest Degree *Abitur*, Craftsmen X Highest Degree *Realschule*, Assistants X Education (Years), Assistants X Age (Years), Service Occupations X Age (Years), Scientists X Age (Years)

*Human Capital Interacted with Survey Year Controls*

2013 X German Grade in Last Record, 2013 X Education (Years)

*Human Capital Interacted with Demographic Controls*

Number of Children X Academic Education, Mecklenburg-Vorpommern X German Grade in Last Record, Saxony X Math Grade in Last Record, Saxony-Anhalt X German Grade in Last Record, Saxony-Anhalt X Math Grade in Last Record, Highest Degree *Fachhochschule* X Partner ID Reported, Highest Degree *Fachhochschule* X Married or Cohabiting, Highest Degree *Realschule* X Saxony, Education (Years) X Household holds Assets, Maths Grade in last Record X Brandenburg, German Grade in last Record X Saxony, Grade Grade in last Record squared X Saxony, Education (Years) squared X Lives in Urban Area, Education (Years) squared X North Rhine-Westphalia, Education (Years) squared X Bavaria, Age (Years) X Saxony, , Age (Years) X Life Satisfaction

*Human Capital Interacted with Firm Controls*

Permanent Contract X Studied MINT Subject, Highest Degree *Realschule* X Small Firm, Education (Years) X Large Firm, Works Council X Education (Years), Maths Grade in last Record X Small Firm, Education (Years) squared X Permanent Contract

*Labor Market Interacted with Demographic Controls*

Number of Children X Labor Market Experience (Years), Partner ID Reported X Unemployment (Years), Catholic X Labor Market Experience (Years), Household holds Assets X Labor Market Experience (Years), Catholic X Labor Market Experience (Years)

*Labor Market Interacted with Character Skill Controls*

Work Satisfaction squared X Unemployment (Years)

*Labor Market Interacted with Occupational and Industrial Controls*

Machine Operator X Part-time (Years), Coke and Chemicals X Job Tenure (Years), Machinery and Equipment X Labor Market Experience (Years), Machinery and Equipment X Job Tenure (Years), Office Workers X Unemployment (Years)

*Labor Market Interacted with Firm Controls*

Large Firm X Job Tenure (Years), Works Council X Labor Market Experience (Years), Works Council X Job Tenure (Years)

*Labor Market Interacted with Survey Year Controls*

2013 X Labor Market Experience (Years), 2009 X Job Tenure (Years)

*Demographics Interacted with Occupational and Industrial Controls*

Financial Intermediation X Partner ID Reported, Lives in Urban Area X Coke and Chemicals Household holds Assets X Coke and Chemicals Lives in Urban Area X Financial Intermediation

*Demographic Interacted with Character Skill Controls*

Partner ID Reported X Neurocitism, Lives in Urban Area X Agreeableness

*Demographic Interacted with Firm Controls*

Lives in Urban Area X Civil Servant, Married or Cohabiting X Large Firm, Baden-Wuerttemberg X Permanent Contract, Schleswig - Holstein X Works Council, Saxony-Anhalt X Small Firm, Household holds Assets X Permanent Contract, Household holds Assets X Works Council

*Demographic Interacted with Survey Year Controls*

2009 X Lives in Urban Area, 2009 X Bavaria, 2009 X Household holds Assets

*Character Skill Interacted with Occupational and Industrial Controls*

Openness squared X Machine Operator

*Character Skill Interacted with Firm Controls*

Permanent Contract X External Locus of Control, Permanent Contract X Neurocitism, Permanent Contract X Risk Aversion, Lives in Urban Area X Agreeableness

*Occupational and Industrial Interacted with Firm Controls*

Coke and Chemicals X Large Firm, Basic Metals and Fabricated Metal Production X Large Firm, Machinery and Equipment X Permanent Contract, Transport Equipment X Large Firm, Service Occupations X Small Firm, Financial Intermediation X Permanent Contract, Assistants X Permanent Contract

*Firm Interacted with Survey Year Controls*

2009 X Permanent Contract, 2009 X Works Council, 2013 X Works Council

---

**At the 10th Percentile**

*Pure Sets of Controls*

*Labor Market Controls*

Unemployment (Years)

*Demographic Controls*

North-Rhine Westphalia X Protestant, Hesse X Lives in Urban Area, Bavaria X Number of Children, Bavaria X Protestant, Bavaria X Married or Cohabiting, Catholic X Number of Children, Mother holds High School Degree X Father holds *Abitur*, Household holds Assets X Lower-Saxony, Household holds Assets, Household holds Assets X Bavaria, Household holds Assets X Father holds *Realschul*-Degree, Close to City Centre ($< 10km$) X Protestant, Interviewer Same Sex X Schleswig - Holstein

*Occupational and Industrial Controls*

Service Occupations, Service Occupations X Public Administration and Defence, Technicians X Educational Sector

*Firm Controls*

Works Council, Large Firm X Permanent Contract

*Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Math Grade in Last Record X Unemployment (Years), Highest Degree *Realschule* X Unemployment (Years)

*Human Capital Interacted with Demographic Controls*

Father holds *Abitur* X Academic Education, Saxony X Math Grade in Last Record, Saxony X Highest Degree *Realschule*, Math Grade in Last Record squared X Hesse, Math Grade in Last Record squared X Household holds Assets, Education (Years) squared X Urban Area, Education (Years) squared X Bavaria

*Human Capital Interacted with Character Skills Controls*

Highest Degree *Realschule* X Life Satisfaction

*Human Capital Interacted with Occupational and Industrial Controls*

Financial Intermediation X Highest Degree *Abitur*, Public Administration and Defence X Highest Degree *Realschule*, Educational Sector X Education (Years), Educational Sector X Highest Degree *Realschule*, Scientists X Education (Years), Technicians X German Grade in Last Record, Technicians X Math Grade in Last Record, Craftsmen X Highest Degree High School, Math Grade in Last Record X Educational Sector, , Assistants X Age (Years)

*Human Capital Interacted with Character Skill Controls*

Risk Aversion X German Grade in Last Record

*Human Capital Interacted with Firm Controls*

Civil Servant X Education (Years), Small Firm X Highest Degree *Realschule*, Medium Firm X Studied MINT Subject, Medium Firm X Highest Degree *Fachhochschule*, Medium Firm X Highest Degree *Abitur*, Large Firm X Education (Years), Works Council X Education (Years), Works Council X German Grade in Last Record

*Labor Market Interacted with Demographic Controls*

Number of Children X Job Tenure (Years), Schleswig-Holstein X Labor Market Experience (Years), Lower-Saxony X Job Tenure (Years), North-Rhine Westphalia X Labor Market Experience (Years), Baden-Württemberg X Job Tenure (Years), Bavaria X Labor Market Experience (Years), Saxony-Anhalt X Unemployment (Years)

*Labor Market Interacted with Firm Controls*

Large Firm X Job Tenure (Years)

*Demographic Interacted with Character Skill Controls*

Risk Aversion squared X Saxony-Anhalt

*Demographic Interacted with Firm Controls*

Large Firm X Protestant, Baden-Württemberg X Permanent Contract, Saxony X Small Firm, Saxony-Anhalt X Small Firm

*Character Skill Interacted with Occupational and Industrial Controls*

Service Occupations X Risk Aversion, Risk Aversion squared X Service Occupations

*Occupational and Industrial Interacted with Firm Controls*

Basic Metals and Fabricated Metal Production X Large Firm, Scientists X Permanent Contract, Technicians X Permanent Contract, Service Occupations X Small Firm, Technicians X Civil Servants

*Demographics Interacted with Occupational and Industrial Controls*

Migration Background X Craftsmen, Lives in Urban Area X Coke and Chemicals, Lower-Saxony X Craftsmen, Bavaria X Office Workers, Bavaria X Craftsmen, Household holds Assets X Financial Intermediation

---

**At the 50th Percentile**

*Pure Sets of Controls*

*Human Capital Controls*

Education (Years), Studied MINT Subject X German Grade in Last Record squared

*Labor Market Controls*

Part-time (Years), Unemployment (Years)

*Demographic Controls*

Hesse X Lives in Urban Area, Saxony X Married or Cohabiting, Mother holds High School Degree X Baden-Württemberg, Catholic X Married or Cohabiting

*Character Skill Controls*

External Locus of Control, Risk Aversion, External Locus of Control squared X Life Satisfaction

*Occupational and Industrial Controls*

Scientists, Service Occupations, Technicians X Transport and Equipment, Health and Social Work X Service Occupations

*Firm Controls*

Small Firm, Works Council X Large Firm

*Panel B: Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Maths Grade in Last Record X Unemployment (Years), Education (Years) X Job Tenure (Years), Highest Degree High School X Part-time (Years), Highest Degree *Realschule* X Unemployment (Years)

*Human Capital Interacted with Demographic Controls*

Highest Degree *Fachhochschule* X Partner ID Reported, North-Rhine Westphalia X Academic Education, North-Rhine Westphalia X Highest Degree *Fachhochschule*, Bavaria X Academic Education, Mecklenburg-West Pomerania X Highest Degree *Realschule*, Saxony X Maths Grade in Last Record, Saxony X Highest Degree *Realschule*, Saxony-Anhalt X German Grade in Last Record, Household holds Assets X Education (Years), Maths Grade in Last Record X Brandenburg, German Grade in Last Record squared X Saxony, Maths Grade in Last Record squared X Brandenburg, Education (Years) squared X Lives in Urban Area, Highest Degree *Fachhochschule* X Married or Cohabiting

*Human Capital Interacted with Character Skill Controls*

External Locus of Control X Education (Years), Conscientiousness squared X Highest Degree *Fachhochschule*

*Human Capital Interacted with Occupational and Industrial Controls*

Coke and Chemicals X Education (Years), Wholesale and Retail Trade X German Grade in Last Record, Craftsmen X Highest Degree *Realschule*, Assistants X Education (Years), Maths Grade in Last Record X Assistants, German Grade in Last Record squared X Coke and Chemicals, German Grade in Last Record squared X Electricity Gas and Water Supply, Education (Years) squared X Machine Operators, Assistants X Age (Years)

*Human Capital Interacted with Firm Controls*

Small Firm X Maths Grade in Last Record, Education (Years) squared X Permanent Contract, Education (Years) squared X Works Council, Large Firm X Education (Years)

*Labor Market Interacted with Demographic Controls*

Hesse X Labor Market Experience(Years), Labor market Experience (Years) X Number of Children

*Labor Market Interacted with Character Skill Controls*

Negative Reciprocity squared X Part-time (Years)

*Labor Market Interacted with Occupational and Sectoral Controls*

Machinery and Equipment X Labor Market Experience (Years), Electricity Gas and Water Supply X Job Tenure (Years), Wholesale and Retail Trade X Part-time (Years), Manufacturing of Textiles or Leather or of Wood X Job Tenure (Years), Technicians X Labor Market Experience (Years), Office Workers X Part-time (Years), Office Workers X Unemployment (Years), Service Occupations X Part-time (Years), Machine Operator X Part-time (Years)

*Labor Market Interacted with Firm Controls*

Large Firm X Labor Market Experience (Years), Works Council X Labor Market Experience (Years), Works Council X Job Tenure (Years)

*Demographic with Character Skill Controls*

Life Satisfaction squared X Saxony, Agreeableness squared X Saxony

*Demographic Interacted with Occupational and Industrial Controls*

Financial Intermediation X Partner ID Reported, Financial Intermediation X Sun Hours at Interview Day, Lives in Urban Area X Financial Intermediation, North-Rhine Westphalia X Scientists, Saxony X Health and Social Work, Saxony X Technicians, Father with High School Degree X Machinery and Equipment, Interviewer Same Sex X Assistants

*Character Skill with Firm Controls*

Permanent Contract X Risk Aversion

*Character Skill with Occupational and Industrial Controls*

Openness squared X Machine Operator, Risk Aversion squared X Assistants

*Occupational and Industrial Interacted with Firm Controls*

Assistants X Civil Servant, Coke and Chemicals X Large Firm, Basic Metals and Fabricated Metal Production X Large Firm, Transport Equipment X Large Firm, Financial Intermediation X Permanent Contract, Assistants X Permanent Contract

---

**At the 90th Percentile**

*Pure Sets of Controls*

*Demographic Controls*

Household holds Assets X Lives in Urban Area, Married or Cohabiting X Saxony

*Occupational and Industrial Controls*

Craftsmen X Basic Metals and Fabricated Metal Production, Machine Operator X Basic Metals and Fabricated Metal Production, Technicians X Educational Sector, Technicians X Health and Social Work

*Firm Controls*

Civil Servant, Large Firm X Civil Servant

*Panel B: Interactions of Different Sets of Controls*

*Human Capital Interacted with Labor Market Controls*

Studied MINT Subject X Labor Market Experience (Years), Highest Degree *Realschule* X Part-time (Years)

*Human Capital Interacted with Demographic Controls*

Studied MINT Subject X Number of Children, Academic Education X Number of Children, Academic Education X Catholic, North-Rhine Westphalia X Academic Education, Hesse X Highest Degree *Abitur*, Maths Grade in Last Record X Mecklenburg-West Pomerania, German Grade in Last Record squared X Mecklenburg-West Pomerania, Maths Grade in Last Record squared X East European Background, Maths Grade in Last Record squared X Brandenburg, Education (Years) squared X Number of Children, Education (Years) squared X Married or Cohabiting, Education (Years) squared X Lives in Urban Area, Education (Years) squared X Household holds Assets

*Human Capital Interacted with Character Skill Controls*

Academic Education X Life Satisfaction

*Human Capital Interacted with Occupational and Industrial Controls*

Electrical and Optical Equipment X Academic Education, Construction X Highest Degree *Realschule*, Public Administration and Defence X Academic Education, Public Administration and Defence X Highest Degree *Abitur*, Craftsmen X Highest Degree *Realschule*, Craftsmen X Age (Years)

*Human Capital Interacted with Firm Controls*

Small Firm X Highest Degree *Realschule*, Large Firm X Academic Education, Education (Years) squared X Permanent Contract, Education (Years) squared X Large Firm, Education (Years) squared X Works Council

*Labor Market Interacted with Demographic Controls*

Saxony X Labor Market Experience, Saxony-Anhalt X Job Tenure (Years)

*Labor Market Interacted with Character Skill Controls*

External Locus of Control X Labor Market Experience (Years), Risk Aversion X Labor Market Experience (Years)

*Labor Market Interacted with Occupational and Industrial Controls*

Office Workers X Unemployment (Years), Assistants X Labor Market Experience

*Labor Market Interacted with Firm Controls*

Permanent Contract X Unemployment (Years)

*Demographic Interacted with Character Skill Controls*

Partner ID Reported X External Locus of Control, Assets X External Locus of Control, Agreeableness squred X Saxony-Anhalt, Negative Reciprocity squared X Saxony-Anhalt

*Demographic Interacted with Occupational and Industrial Controls*

Public Administration and Defense X Partner ID Reported, Machine Operator X Number of Children, Machine Operator X Partner ID Reported, Migration Background X Basic Metals and Fabricated Metal Production, Berlin X Technicians, Father with High School Degree X Machine Operator, Interviewer Same Sex X Machine Operator,

*Demographic Interacted with Firm Controls*

Saxony-Anhalt X Permanent Contract

*Character Skill Interacted with Occupational and Industrial Controls*

Agreeableness squared X Machine Operator

*Character Skill Interacted with Firm Controls*

Permanent Contract X External Locus of Control

*Occupational and Industrial Interacted with Firm Controls*

Wholesale and Retail Trade X Union Member, Public Administration and Defense X Large Firm, Scientists X Large Firm, Technicians X Small Firm, Service Occupations X Large Firm, Assistants X Large Firm

Table shows unrestricted Wage-LASSO at the mean, 10th, 50th, and 90th percentile, respectively. In case of the unrestricted LASSO specification no variables are forced to be part of the model. Thus, the LASSO selects among the full set of potential controls. The set of control variables 'Human Capital' may includes – depending on the choice of the LASSO – educational attainment, age, degree, grade in last record and a control for Mathematics, Informatics, Natural Science or Technology (MINT) subject. We refer to labor market experience, job tenure, years of parental leave, past periods of unemployment or as part-time as labor market controls. We define as demographic controls: migration status, urban residence, residence $<$ 10 km to city centre, federal state, being married or cohabiting and whether the partner ID was actually reported and information on parents' education, number of children, presents of small children in household, region of origin, dummies for different religions as well as sun hours at interview day and the sex of interviewer. Character skill controls include the Big Five Personality Traits (openness, conscientiousness, agreeableness and neuroticism) as well as risk aversion, positive and negative reciprocity and external locus of control. The attribute work or life satisfaction controls to the set of character skills. Occupational and industrial controls are based in ISCO88 (1-digit) and NACE (2-digit). Firm controls include dummies for works council, civil servant, firm size (medium or large), having a permanent contract, being a union member. Survey years dummies may include 2009 or 2013. *Fachhochschule* is a University of Applied Sciences, *Abitur* a high-school degree and *Realschule* a type of secondary school in Germany.

# C Further Decomposition Results

In this Section, we present the decomposition results using the conventional and machine learning models at the mean. Further, we show how the estimated aggregate explained components change when using female as non-discriminatory wage structure along the wage distribution.

## C.1 Decomposition Results at the Mean

We present both the aggregate and detailed explained component for the case of the mean.[19] In case of the detailed decomposition, we look at gender differences in observable characteristics attributable to human capital, labor market controls, occupational or industrial sorting, firm and demographic characteristics as well as character skills and time. Table C.1 shows the results of the decomposition at the mean using the conventional model specifications. On aggregate, gender differences in observable characteristics explain between 48% and 55% of the total or aggregate wage gap using the conventional model specifications. Gender differences in observed human capital controls such as education narrow the gap in all specifications. Observable labor market characteristics are found to widen the gap significantly in all three specifications. Differences in observable character skills, if added, lead to a widening of the gap. Occupational and industrial sorting does not affect the GPG at the mean. Average differences between men and women in

[19] Decomposition results at the mean using women as non-discriminatory wage structure are available from the author upon request.

demographic background characteristics such as being married also widen the GPG in all three specifications. Firm characteristics including inter alia dummies for firm size and works council increase the gap only in the first two specifications. In the baseline model specification (column (5) and (6)), these characteristics do not affect the GPG. Time or survey year dummies correct the gap slightly downwards in all three specifications.

Table C.2 shows the decomposition of the restricted and unrestricted LASSO specification. Both specifications explain a substantial part of the GPG. In particular, while the restricted LASSO model explains – as the conventional wage models – half of the gap, the unrestricted LASSO explains more than 60% of the pay gap. That is, using the restricted LASSO allows to explain 20% more of the gap. Moreover, while the restricted LASSO model yields similar results as the conventional wage models, the results of the unrestricted LASSO model differs markedly. In particular, in the unrestricted LASSO differences in terms of human capital and labor market characteristics between men and women as well as occupational and industrial sorting do no longer substantially affect the GPG. The same holds for demographic and character skill controls. Human capital controls interacted with demographic controls explain a part of the GPG that is in size comparable the effect of human capital in the conventional models. These results suggest as in the case along the wage distribution (cfr. Section 4.4) that different procedures for model selection (conventional, restricted or unrestricted machine learning) yield different results.

Table C.1: Decomposition Results at the Mean, Conventional Models

| | (1) | (2) | (3) |
|---|---|---|---|
| | Specification: | | |
| | Full | Augmented | Baseline |
| Difference | | 0.168*** | |
| | | (0.017) | |
| Explained Component (Total) | 0.081*** | 0.086*** | 0.092*** |
| | (0.017) | (0.017) | (0.017) |
| *Detailed Decomposition:* | | | |
| Human Capital | -0.019** | -0.018** | -0.017** |
| | (0.008) | (0.008) | (0.007) |
| Labor Market | 0.072*** | 0.070*** | 0.061*** |
| | (0.012) | (0.012) | (0.011) |
| Occupations and Industries | 0.006 | 0.005 | 0.007 |
| | (0.012) | (0.011) | (0.010) |
| Firm | 0.007** | 0.007** | 0.013** |
| | (0.003) | (0.003) | (0.005) |
| Demographics | 0.023*** | 0.023*** | 0.029*** |
| | (0.006) | (0.006) | (0.005) |
| Character Skills | | 0.008* | 0.007* |
| | | (0.005) | (0.004) |
| Time | -0.008*** | -0.008*** | -0.008*** |
| | (0.002) | (0.002) | (0.002) |
| Observations | 8,489 | 8,489 | 8,489 |

*Notes:* Calculations use SOEP sample weights. Robust standard errors clustered at the individual level in parentheses. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. The set of controls Human Capital includes years of education and age in the Full and Augmented specification. In the Baseline model, Human Capital additionally includes age squared. Labor Market includes in the Full and Augmented model labor market experience, while it additionally contains job tenure as well as labor market experience squared in the Baseline model. Demographics include in the Full and Augmented specification dummies for urban residence, migration background as well as federal-state dummies. In the Baseline model, this category includes additionally a dummy for being married. Occupations and Industries includes occupational and sectoral dummies. Firm includes in the Full and Augmented specification dummies for union membership. In case of the Baseline specification, the set contains additionally dummies for having a permanent contract, presence of a works council, firm size (medium and large, where medium refers to firm between 20 and 199 employees and large refers to firms with at least 200 employees). Time includes controls for the survey year. Character Skills include the Big Five Personality Traits, locus of control, reciprocity and willingness to take a risk. *Source:* SOEP v33.

Table C.2: Decomposition Results at the Mean, Restricted and Unrestricted LASSO Models

| | (1) | (2) |
|---|---|---|
| | Specification: | |
| | Restricted LASSO | Unrestricted LASSO |
| Difference | | 0.168*** |
| | | (0.017) |
| Explained Component (Total) | 0.090*** | 0.106*** |
| | (0.020) | (0.025) |
| *Detailed Decomposition:* | | |
| Human Capital (HC) | -0.020*** | -0.007 |
| | (0.006) | (0.004) |
| *Interactions:* | | |
| HCXLM | -0.015 | 0.027 |
| | (0.016) | (0.049) |
| OIXHC | 0.013 | 0.003 |
| | (0.016) | (0.020) |
| HCXFirm | 0.014* | 0.003 |
| | (0.008) | (0.005) |
| HCXDemo | 0.015** | 0.016*** |
| | (0.006) | (0.005) |
| HCXCS | -0.000 | -0.012 |
| | (0.011) | (0.017) |
| HCXTime | | -0.002 |
| | | (0.003) |
| Labor Market (LM) | 0.054*** | 0.021 |
| | (0.019) | (0.044) |
| *Interactions:* | | |
| LMXOI | 0.012 | 0.004 |
| | (0.009) | (0.013) |
| LMXFim | | 0.009** |
| | | (0.004) ) |
| LMXDemo | 0.018 * | 0.010 |
| | (0.013) | (0.009) |
| LMXCS | -0.001 | -0.001 |
| | (0.001) | (0.001) |
| LMXTime | | -0.001 |
| | | (0.001) |
| Sectors and Industries (OI) | -0.022 | 0.027 |
| | (0.015) | (0.019) |

| | | |
|---|---|---|
| *Interactions:* | | |
| OIXFirm | 0.008 | 0.018** |
| | (0.006) | (0.007) |
| OIXDemo | -0.008 | -0.011 |
| | (0.008) | (0.008) |
| OIXCSc | 0.001 | -0.026*** |
| | (0.004) | (0.008) |
| Firm | 0.007 | 0.006* |
| | (0.005) | (0.003) |
| *Interactions:* | | |
| FimXDemo | 0.003 | 0.003* |
| | (0.002) | (0.002) |
| FirmXCS | 0.004 | 0.010 |
| | (0.006) | (0.008) |
| FirmXTime | | 0.001 |
| | | (0.002) |
| Demographics (Demo) | 0.017* | 0.009 |
| | (0.009) | (0.006) |
| *Interactions:* | | |
| DemoXCS | 0.004 | 0.005 |
| | (0.004) | (0.003) |
| DemoXTime | | -0.000 |
| | | (0.001) |
| Character Skills (CS) | -0.005 | 0.002 |
| | (0.013) | (0.019) |
| Time | -0.009*** | -0.007* |
| | (0.002) | (0.004) |
| Observations | 8,489 | 8,489 |

*Notes:* Calculations use SOEP sample weights. Robust standard errors clustered at the individual level in parentheses. *, ** and *** denote significance at the 10%-, 5%- and 1%-level, respectively. The set of controls Human Capital includes in the Restricted LASSO years of education and quadratic polynomials of age (amelioration set). Further human capital controls selected by the double robust restricted LASSO are presented in Appendix B. In the Unrestricted LASSO models, Human Capital includes the variables specified in Appendix B. Labor Market contains job tenure as well as quadratic polynomials of labor market experience in the Restricted LASSO model (amelioration set). The remaining controls as well as the control attributable to labor market characteristics in the Unrestricted LASSO specification are shown in Appendix B. Demographics include dummies for urban residence, migration background, federal-state dummies and a dummy for being married. Interactions with demographic control variables and the definition of Demographics in the Unrestricted LASSO specification is shown in Appendix B. Occupations and Industries includes the control variables described in Appendix B. Firm includes in the Restricted LASSO model dummies for union membership, having a permanent contract, presence of a works council, firm size (medium and large, where medium refers to firm between 20 and 199 employees and large refers to firms with at least 200 employees). The latter is the amelioration set. Interactions with firm dummies and the definition of Firm for the Unrestricted LASSO is shown in Appendix B. Selected Character Skills for the Restricted and Unrestricted LASSO models are shown in Appendix B. Time includes survey year dummies in the Restricted LASSO model (amelioration set). Interactions with the survey year dummies and the selected controls of Time for the Unrestricted LASSO is shown in Appendix B. *Source:* SOEP v33.

## C.2 Decomposition Results using Women as Non-Discriminatory Wage Structure

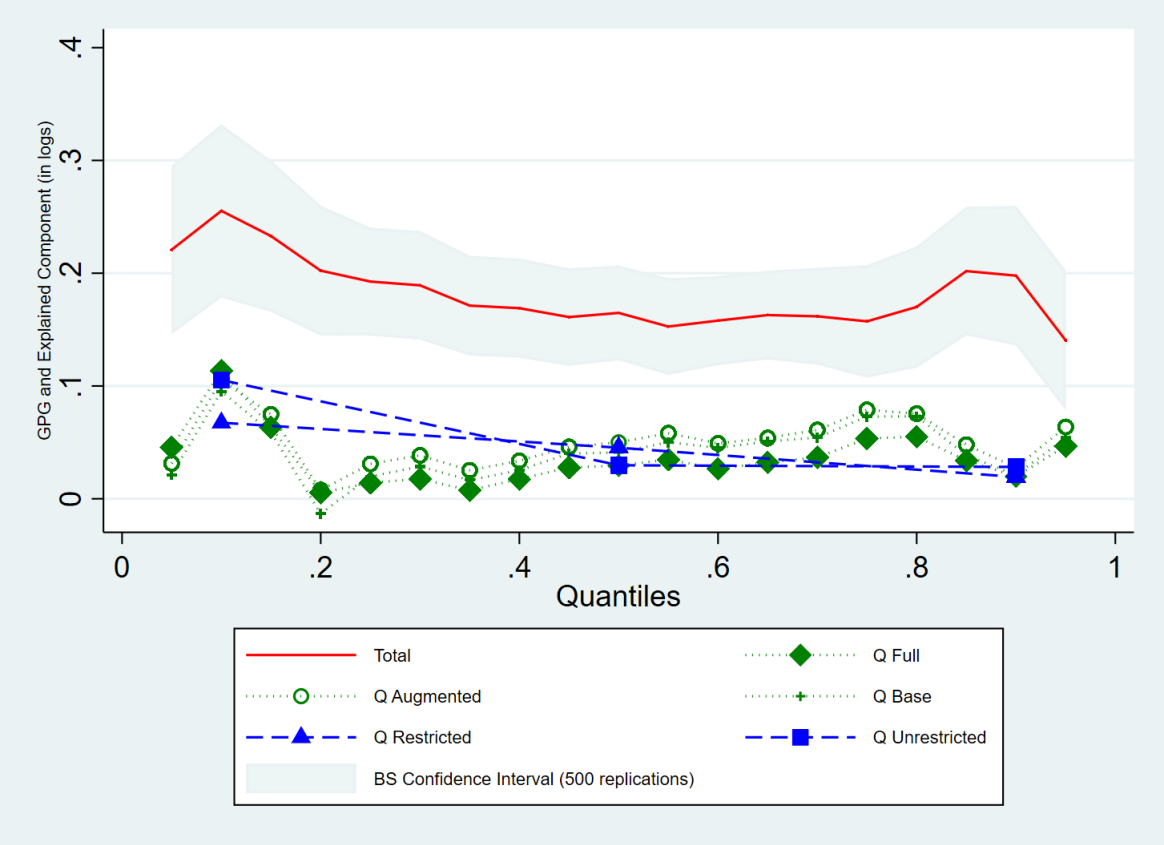
It is well-known in the literature that the estimated components of the GPG in Oaxaca-Blinder decompositions may change substantially depending on the reference category chosen (see e.g. Cotton, 1988, for theoretical details and an application). Therefore,

we repeat the decomposition analysis using women as non-discriminatory wage structure or reference category. We present the raw gap and the aggregate explained component along the wage distribution. Remember that the post-double-LASSO estimator ensures valid inference only with respect to the coefficient of the explanatory variable of main interest (i.e. the adjusted GPG in our case) and that all other included control variables may be endogenous. Further, the variables selected by the double-LASSO procedure may change depending on the sample at hand (Mullainathan and Spiess, 2017). Therefore, meaningful interpretation of coefficient estimates of variables that are part of the set of selected controls is, according to theory, not advisable. Therefore, we skip the presentation of the detailed decomposition here. The results of the explained part attributable to human capital, labor market characteristics, industrial and occupational sorting as well as to character skills at different points of the distribution with women as reference category are available from the authors upon request.

Figure C.1 shows the aggregate decomposition with explained components and raw gaps. While the contribution of all observable characteristics (explained component) is relatively stable across the distribution using men as reference category in the conventional models (cfr. Figure 5), the explained part varies substantially when women are the reference category (see Figure C.1; variation between zero and ten percentage points). We also find that the choice of the reference category matters in the data-driven models. The latter explain almost the entire wage gap at the top in the case of the male reference group, but explain almost nothing at the top when females are the reference. Moreover, the explained part is decreasing from the bottom to the top in the LASSO models. That is, while gender differences in observable characteristics explain about 10 percentage points at the 10th percentile (raw gap of 26%), they explain only about two percentage points of the 90th percentile GPG (raw gap of 20%).

To sum up, we find substantial differences in the explained components when using women as non-discriminatory wage structure compared to using men as reference category. However, the main implication persists. Depending on the procedure for model selection (conventional or data-driven) the results may differ and thus the same may apply for policy conclusions.

Figure C.1: Oaxaca-Blinder Decomposition of the Gender Pay Gap (GPG) in Log Hourly Wages along the Distribution: Aggregate Decomposition (Explained Component) – Female Reference Category




*Notes:* Figure shows the raw or unadjusted GPG along the wage distribution (solid line) as well as its explained component ($Q$) obtained from a standard Oaxaca-Blinder decomposition of the conventional models (dotted line) and the restricted and unrestricted LASSO specifications (dashed line). The raw gap and the explained component of the conventional models (Full, Augmented and Baseline) are estimated at each 5th quantile $\tau$ with $\tau \in [0.5, 0.95]$, while the explained component of the LASSO specifications (Restricted and Unrestricted) due to computational intensity is estimated only at the $0.1, 0.5$ and $0.9$ quantile, respectively. SOEP sample weights used. *Source:* SOEP v33.

**In der Diskussionspapierreihe sind kürzlich erschienen:**

**Recently published Discussion Papers:**

| | | | |
|---|---|---|---|
| 111 | Briel, S., Töpfer, M. | The Gender Pay Gap Revisited: Does Machine Learning offer New Insights? | 01/2020 |
| 110 | Kölling, A., Schnabel, C. | Owners, external managers, and industrial relations in German establishments | 11/2019 |
| 109 | Hinz, T., Lechmann, D. | The role of job satisfaction and local labor market conditions for the dissolution of worker-job matches | 07/2019 |
| 108 | Prümer, S., Schnabel, C. | Questioning the stereotype of the “malingering bureaucrat”: Absence from work in the public and private sector in Germany | 05/2019 |
| 107 | Prümer, S. | Ist der Staat der bessere Arbeitgeber? Arbeitsqualität im Öffentlichen und Privaten Sektor in Deutschland | 05/2019 |
| 106 | Bossler, M., Oberfichtner, M., Schnabel, C. | Employment adjustments following rises and reductions in minimum wages: New insights from a survey experiment | 08/2018 |
| 105 | Töpfer, M. | The Age Pay Gap and Labor Market Heterogeneity: A New Empirical Approach Using Data for Italy | 07/2018 |
| 104 | Fackler, D., Fuchs, M., Hölscher, L., Schnabel, C. | Do startups provide employment opportunities for disadvantaged workers? | 05/2018 |
| 103 | Castagnetti, C., Rosti, L., Töpfer, M. | Discriminate Me – if You Can! The Disappearance of the Gender Pay Gap among Public-Contest Selected Employees | 02/2018 |
| 102 | Oberfichtner, M., Schnabel, C. | The German Model of Industrial Relations: (Where) Does it still Exist? | 10/2017 |
| 101 | Lechmann, D. | Estimating labor supply in self-employment: pitfalls and resolutions | 09/2017 |

Eine aktualisierte Liste der Diskussionspapiere findet sich auf der Homepage:
http://www.arbeitsmarkt.rw.fau.de/

An updated list of discussion papers can be found at the homepage:
http://www.arbeitsmarkt.rw.fau.de