Estimating the Conditional Gender Pay Gap: 
The Role of Omitted Controls

STEPHANIE BRIEL and MARINA TÖPFER

University of Hohenheim
University of Erlangen-Nuremberg

This draft: February 15, 2019
Preliminary, please do not cite or circulate.

Abstract: We estimate the gender pay gap using data for Germany in the years 2005, 2009 and 2013. Main interest of our analysis lies in accounting for potential omitted variable bias and its consequences for conventionally estimated gender pay gaps. In order to minimize the threat of omitted variable bias, we use a machine learning technique for model selection. This ensures that we control for the most important gender differences and wage predictors. We compare the estimated gender pay gaps and decomposition results to those of a baseline model specified in line with recent literature. Further, we evaluate the performance of the different models based on the method proposed by Oster (2017). The complete analysis is conducted at the mean as well as at specific quantiles of the wage distribution. Overall, we find evidence that more flexible model specifications as well as different model specifications along the wage distribution might be required if an (as) unbiased (as possible) estimation of the gender pay gap is of main interest.

JEL: J16, C21, C52

Keywords: Gender Pay Gap, Machine Learning, Omitted Variable Bias, Selection on Unobservables

Correspondence to: University of Hohenheim, Department of Economics (520B), 75093 Stuttgart, Germany.
1 Introduction

Gender differences in pay are a frequently analyzed topic. The Gender Pay Gap (GPG) persists worldwide despite political emphasis to close it (see for example Blau and Kahn, 2017; Godin, 2014). In the European Union men earn on average 16% more than women (Eurostat, 2016). In Germany, the wage gap lies with more than 20% above the EU-average (Eurostat, 2016). A large number of studies have shown that the GPG varies substantially across the wage distribution. Most of these studies find particularly pronounced differences at the top and bottom of the wage distribution (‘glass ceiling’ and ‘sticky floors’, e.g. Albrecht et al., 2003; Arulampalam et al., 2007). Generally, regression models used for estimation of GPGs are based on a Mincer-type wage equation. 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’ (unbiased) GPG, the conditional independence assumption or unconfoundness has to hold, i.e. we have to control for all factors that are characterized by important gender differences and that are highly predictive for wages (see e.g. Fortin et al., 2011).

Key characteristics for estimating a standard Mince-type wage equation include observable labor market and human capital variables such as job market tenure, labor market experience, schooling as well as industrial and occupational controls (Blau and Kahn, 2017; Goldin, 2015; Mandel and Semyonov, 2014). Individual characteristics such as family background (Bailey, 2012; Fortin, 2008) and union status (Heinze and Wolf, 2010) are also important controls. Firm-specific characteristics like firm-size or presence of a works councils are found to be important as well (Heinze and Wolf, 2010). A more recent part of the literature considers character skills as important controls when estimating the GPG. Character skills are key determinants of social and economic success. For instance, conscientiousness, which describes the tendency of an individual to be hard-working and well organized, is positively related to a wide range of economic outcomes (Almlund et al., 2011; Heckman et al., 2011; Heckman and Rubinstein, 2001). In general, there are significant gender differences in character skills. Recent literature analyzed their role in explaining gender differences in pay. Most of these studies focus on the Big Five Personality Traits as measures for character skills (Brenzel and Laible, 2016; Mueller and Plug, 2006), some additionally control for Locus of Control (Nyhus and Pons, 2012) as well as Reciprocity (Heineck and Anger, 2010) and Willingness to take a Risk (Collischon, 2017b). Nowadays, human capital variables - including character skills - are found to explain only minor to moderate parts of the GPG, while occupations and industries are found to be more important in explaining gender differences in pay (Blau and Kahn, 2017).

Given these findings, we are, first, interested in identifying the factors that are important if an (as) unbiased (as possible) estimation of the GPG is of main interest. Second, we analyze whether the same model should be used across the wage distribution or whether the determinants of wages might change along the wage distribution. In order to answer these questions, we rely on a machine learning technique. Starting point of our analysis

1
is a Mincer-type wage equation in line with the literature that we augment by individual character skills in order to estimate the GPG. In the following we will refer to this model specification as the baseline model. For identification of the pivotal control variables, in the next step of our analysis, we rely on the post-double-Least Absolute Shrinkage and Selection Operator (LASSO) estimator proposed by [Belloni et al. (2014a,b)]. The advantage of this type of model selection is that it results in a model specification with controls for the most important gender differences as well as the most important wage predictors. Thus, it immunizes against omitted variable bias. Besides estimation at the mean, we rely on unconditional quantile regression or RIF-OLS to analyze the GPG along the wage distribution. Precisely the post-double-LASSO estimator is used at the mean as well as at specific quantiles. Therefore, we can analyze whether the wage predictors change along the distribution. Next, we compare the performance of our data-driven models to a conventional model using a method recently proposed by [Oster (2017)]. With this method, for each model specification, we can calculate the degree of remaining selection on unobservables relative to selection on observables necessary to result in a zero GPG. Finally, we compare the decomposition results of the different model specifications. For separate identification of the wage structure and composition effects ignorability and common support are key assumptions [Firpo et al. (2018)]. Indeed, the pivotal assumption is that the distribution of unobserved explanatory factors in the wage determination is the same for men and women, once we condition on a vector of observed components. This is more likely to hold if model selection is based on the post-double-LASSO estimator.

We use data from the German Socio-Economic Panel Study (SOEP), providing a large nationally representative sample with extensive information on a broad range of validated behavioral characteristics such as character skills [Wagner et al. (2008)]. It provides detailed information on individual characteristics that might be important controls for estimation of the GPG. Moreover, it contains information on important job and employer characteristics like firm size, sectors, union status, presence of workers councils, working time, occupations and unemployment history. Overall, we are able to provide similar firm and job-specific information as in related studies (see e.g. [Brenzel and Laible (2016)]. The main advantage of the SOEP lies in its extensive individual level information that we exploit for the machine learning based model selection procedure.

Model selection with the post-double-LASSO estimator reveals that character skills are important control variables for the estimation of the GPG. Further, the selected models are more flexible than conventional models for estimation of GPGs, i.e. they contain more interactions and higher order polynomials. Besides this, we find that the wage predictors change along the wage distribution. In case of the baseline model we find evidence for glass ceiling, i.e. a particularly pronounced GPG at the top of the distribution. Compared to the baseline model, the models selected by the post-double-LASSO estimator deliver

\[1\] We use the post-double-LASSO because, as a following step, we want to use the method by [Oster (2017)] to compare different model specifications. Therefore, we need wage models can be estimated by OLS.
only slight evidence for glass ceiling and suggest a U-shape of gender differences in pay along the wage distribution. The decomposition results show that besides at the bottom of the wage distribution the models selected by the post-double-LASSO estimator explain (substantially) larger shares of the GPG. This is most pronounced at the top of the wage distribution. Overall, we find evidence that more flexible model specifications as well as different model specifications along the wage distribution might be required for an (as) unbiased (as possible) estimation of the GPG.

This paper is organized as follows. Section 2 outlines the estimation methods applied. Next, Section 3 describes the data set used and Section 4 presents the estimation results. Finally, Section 5 concludes.
2 Estimation Method

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

$$y_i = \beta d_i + x_i \gamma + u_i.$$  \hspace{1cm} (1)

where $y_i$ is the log of hourly earnings, $d_i$ 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. $u_i$ is the error term. The coefficient of $d_i$ gives the GPG. For unbiased estimation of the conditional GPG\(^2\) ($\beta$) the conditional independence assumption has to hold. This means that the set of control variables ($x_i$) has to contain all variables that are simultaneously correlated with $d_i$ and $y_i$, and $E(u_i|d_i, x_i) = 0$.

For our analysis we will consider different specifications for (1), which differ in the elements of the vector $x_i$. In case of the short regression model, $x_i$ only contains the constant. Thus, the GPG based on this model is the raw GPG. In the baseline model, the set of control variables $x_i$ contains, besides the constant, standard Mincer-type controls (conventional control variables) plus a set of character skill controls. By construction, this model specification is similar to model specifications used in the recent literature and is supposed to represent the conventional estimation approach, i.e. a Mincer-type wage model estimated by OLS.

Concerning model specification, a number of questions arise. Regarding the inclusion of the Big Five character skills in the model, for instance, in most settings (if used) the complete (available) set of character skill controls is included linearly. One question that arises is whether it is appropriate to include the full 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 might be more appropriate 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 her (employer learning). This 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 are also conceivable for the other control variables. Thus, it might be appropriate to consider more flexible model specifications than usual in case of estimation of GPGs. Consequently, we can think of a variety of potential interactions and higher-order polynomials, and therefore, a high-dimensional set of potential controls. In order to decide which controls to include for estimation of the GPG, we rely on a machine learning technique for model selection. Indeed, more and more research uses machine learning approaches when causal inference is the main goal (Chernozhukov et al. 2018). Therefore, in the restricted and unrestricted LASSO specification that we present

\(^2\)By conditional GPG we mean the GPG estimated with control variables $x_i$, i.e. conditioned on $x_i$. 4
in the following Section, the set of control variables $x_i$ contains regressors selected by a machine learning technique.

### 2.1 Model Selection: Post-Double-LASSO Estimator

We use the post-double-LASSO estimator proposed by Belloni et al. (2014a,b) for model selection. The idea is to use the data at hand to identify pivotal control variables for the estimation of the GPG. Relaxing the linearity assumption of equation (1), the partially linear wage model is given by:

$$y_i = \beta d_i + g(z_i) + \zeta_i.$$  \hspace{1cm} (2)

Further, it holds for $d_i$, that

$$d_i = m(z_i) + \nu_i$$  \hspace{1cm} (3)

where $z_i$ corresponds to the set of potential control variables. $\zeta_i$ and $\nu_i$ are the respective error terms. The functions $g(\cdot)$ and $m(\cdot)$ are not necessarily linear. Thus, linear combinations of the potential control variables – interactions and higher order polynomials with $x_i = P(z_i)$ – are used for their approximation, i.e.:

$$g(z_i) = x_i' \alpha_g + r_{gi}$$  \hspace{1cm} (4)

$$m(z_i) = x_i' \alpha_m + r_{mi}$$  \hspace{1cm} (5)

where $x_i' \alpha_g$ and $x_i' \alpha_m$ give the linear approximations and $r_{gi}$ and $r_{mi}$ are the corresponding approximation errors.

As stated before, in the case of the baseline model we rely on $x_i$ defined as a set of commonly used control variables. If, however, model selection is based on the post-double-LASSO estimator, the elements of $x_i$ are selected among the set of potential control variables that contains, additionally to the variables commonly used, additional factors as well as higher order polynomials and interactions of those variables (more details are given in Section 3). Theoretically, the set of potential controls can be huge. We denote the number of potential controls by $P$. As this method uses the LASSO, we have to assume approximate sparsity. This implies that we need only a relatively small number of controls for estimation of the model.

There are 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
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 \cite{Hastie2009}. In what follows, we rely on a slightly modified version of the LASSO that was first introduced by \cite{Belloni2012}. This procedure defines the LASSO as:

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \gamma_j$$

where $P$ potential controls are considered, $\gamma_j$ 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. Compared to the standard LASSO, the absolute value of each coefficient is additionally multiplied by its penalty loading $\gamma_j$. The inclusion of $\gamma_j$ 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 \cite{Belloni2012}. The penalty term in equation (6) has a kink at 0. 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 \cite{Belloni2014}. If prediction of the outcome is of main interest $\lambda$ is often chosen by cross-validation. However, as shown by \cite{Belloni2012}, there are superior methods for the choice of $\lambda$ if prediction is not the main goal.

The LASSO - at least in the current setting - is associated with two main challenges. First, the nonzero 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 about model parameters. We can address the first concern by employing the Post-LASSO estimator proposed by \cite{Belloni2012}. Post-LASSO estimation proceeds in two steps. First, the LASSO is run to determine important predictors for the outcome variable. Second, 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. The

\footnote{In comparison with the standard LASSO, the intercept is not part of the minimization problem. Note that the intercept is in neither equation included in the penalty term. Penalization of the intercept would make the procedure dependent on the origin chosen for $y_i$. The solution to the standard LASSO can be separated in two parts. First, one has to replace all variables of the set of potential controls with their centered values ($x_{ij} - \bar{x}_j$). Then, the intercept equals $\bar{y} = 1/N \sum_{i=1}^{N} y_i$. Finally, one can estimate the remaining coefficients with LASSO without an intercept using the centered covariates.}

\footnote{See \cite{Belloni2012} for more details on the calculation of $\lambda$ and $\gamma_j$. $\lambda$ was chosen dependent on the sample size and on the number of potential controls.}

\footnote{\cite{Belloni2013} and \cite{Belloni2012} have shown that the post-LASSO...}
number of variables in the set of selected controls is denoted by $S$.

The second concern is more challenging in our context. Estimation of the GPG in the standard context (in equation (1)) is based on a conditional-on-observables identification strategy - meaning that the main goal is to find among the set of potential controls those factors that are simultaneously correlated with $d_i$ (gender differences) and $y_i$ (wage predictors). Applying the LASSO to the wage equation given in equation (1), where $d_i$ is forced to remain in the model by excluding its coefficient from the penalty term, would result in a set of selected controls, where any variable that is highly correlated with gender is (likely to be) dropped since its addition does not add much explanatory power if the model already controls for gender. However, if we omit variables that are highly correlated with $d_i$, our estimated GPG suffers from the potential threat of an omitted variable bias. Hypothetically, we can rely on two different equations. A regression of $y_i$ on the set of potential control variables (see equation (4)) and a regression of $d_i$ on the set of potential controls see equation (5). Relying on one single equation is only feasible if there are no errors in variable selection. If we apply 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. In fact, the LASSO tends to miss variables with coefficients of only moderate size. If the neglected variables have large coefficients in the second regression ($d_i$ on $x_i$) this results in omitted variable bias when estimating the GPG and only include control variables selected by the LASSO of $y_i$ on $x_i$ (Belloni et al., 2014a).

Due to the potential drawbacks if model selection is based on a single equation, it is important to consider a prediction for both the natural logarithm of hourly earnings, $y_i$ and gender, $d_i$. 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 useful to predict $d_i$. We refer to this as the Gender-LASSO. The Gender-LASSO results in the set of selected gender controls $x_{ig}$. Next, the LASSO is used to select the set of controls that are useful to predict $y_i$. We call this the Wage-LASSO. The Wage-LASSO results in the set of selected wage controls $x_{iw}$. The GPG is estimated by OLS with the union of the two sets of selected controls, i.e. $x_i = x_{ig} + x_{iw}$, as control variables. 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 GPG by immunization against the types of model selection errors mentioned above (Belloni and Chernozhukov, 2013). While the use of the Gender-LASSO ensures that we control for all important gender differences, the Wage-LASSO further 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 baseline model. It helps to reduce the omitted variable bias and enables performance of uniform inference (for the GPG) after

is equal to or better than the LASSO in terms of rates of convergence and bias.
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. In case of sample splitting, the regularity condition becomes \( s \ll N \) (Belloni et al., 2012; Chernozhukov et al., 2018). For sample splitting one divides the sample into a main and an auxiliary sample. The main sample is used to select the control variables among the set of potential controls and the auxiliary sample for the post-LASSO estimation. Changing the ordering of the samples results in two different post-double-LASSO estimators. The average of these two different estimator gives then the post-double-LASSO estimator based on sample split (Chernozhukov et al., 2018).

If some variables are considered to be indispensable for estimation of the GPG, one can exclude the corresponding coefficients from the penalty term in equation (6). Such a set of variables is called amelioration set. Model selection is, thus, restricted to a smaller set of potential controls (\( P \) minus the amelioration set). That is, in the case of the restricted LASSO, we do not allow for shrinkage of the coefficients of the controls of the amelioration set. A further regularity condition for the use of the post-double-LASSO estimator is that the amelioration set is not allowed to be substantially larger than any of the two selected sets of controls. Thus, the number of variables in the amelioration set is allowed to be at maximum as high as the number of elements in the larger of the two sets of selected controls resulting from the Gender-LASSO or Wage-LASSO, respectively (Belloni et al., 2014b). Indeed, as one might 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 restricted LASSO.

2.2 Remaining Selection on Unobservables

One question of interest of this paper is whether models selected by the post-double-LASSO estimator are less prone to omitted variable bias than conventionally estimated gender pay gaps. To shed more light on this, we employ a method recently proposed by Oster (2017), that provides a refinement of the method proposed by Altonji et al. (2005).

Based on coefficient stability and changes in the degree of explained outcome variation when switching from the short to the intermediate regression model, we can calculate the degree of remaining selection on unobservables relative to selection on observables that is necessary to produce a zero GPG. As stated before, in the case of the short regression model no control variables are included, i.e. \( d_i \) is the only right-hand-side variable in equation (1). The intermediate regression model is the regression model with the complete set of control variables. In our setting, we consider three different intermediate models: the baseline model, the unrestricted and the restricted LASSO and compare their performance with respect to robustness against remaining selection on unobservables. We refer to the GPG effect in the short (intermediate) regression model by \( \beta_{\text{short}} \) (\( \beta_{\text{inter}} \)). \( R^2_{\text{short}} \) gives the \( R^2 \) corresponding to the short regression model. In case of the intermediate regression
Further, the hypothetical construct of the full regression model refers to the model that contains all factors (observables and unobservables) that are simultaneously correlated with gender and the log of hourly earnings. The corresponding unbiased GPG effect is given by $\beta_{\text{full}}$. Proportional selection, with proportionality parameter $\delta$, is the key assumption in this framework. It implies that the degree of remaining selection on unobservables necessary to produce a zero GPG can be expressed as the degree of selection on observables multiplied by the proportionality parameter $\delta$. Further, we assume that the relative contribution of each variable in $x_i$ to $d_i$ is the same as their contribution to $y_i$. Under these assumptions, we obtain the following approximation of the bias-adjusted GPG:

$$
\beta_{\text{rest}} = \beta_{\text{inter}} - \delta [\beta_{\text{short}} - \beta_{\text{inter}}] \frac{R^2_{\text{full}} - R^2_{\text{inter}}}{R^2_{\text{inter}} - R^2_{\text{short}}} \quad (7)
$$

The adjustment is based on a change in coefficient $\beta$ from the short to the intermediate model ($\beta_{\text{short}} - \beta_{\text{inter}}$), proportionality of selection ($\delta$) and the relation of the shift in explained outcome variation when switching from the intermediate to the full model ($R^2_{\text{full}} - R^2_{\text{inter}}$) compared to the switch from the short to the intermediate model ($R^2_{\text{inter}} - R^2_{\text{short}}$). Thus, the coefficient movement is scaled by the proportionality parameter and the movements in the explained outcome variation. Besides $R^2_{\text{full}}$ and $\delta$ all inputs are observable in standard regression outputs. Based on assumptions about $R^2_{\text{full}}$ and $\delta$, equation (7) can, thus, be easily calculated. Usually it is assumed that $\delta = 1$, implying that selection on unobservables is as important as selection on observables (following Altonji et al., 2005). The most conservative assumption for $R^2_{\text{full}}$ is $R^2_{\text{full}} = 1$. However, as this may be too restrictive, we use the threshold proposed by Oster (2017), i.e. $R^2_{\text{full}} = 1.3 \times R^2_{\text{inter}}$.

Besides restrictive assumptions, (7) gives an easy-to-calculate approximation for the bias-adjusted GPG and delivers a reasonable approximation of the more precise and less restrictive bias-adjusted GPG (see Oster, 2017). This less restrictive bias-adjusted GPG does not rely on the assumption of equal relative contributions.

Instead of calculating the bias-adjusted GPG, we use the procedure to determine the proportionality parameter ($\delta$). Thus, we calculate the degree of remaining selection on unobservables relative to selection on observables that is necessary to yield a zero GPG.

---

6 Based on the model specification, $x_i$ contains the set of controls given in our baseline specification or the set of controls selected by the corresponding LASSO procedure.

7 Later we will rely on RIF-OLS, where the dependent variable is binary, thus assuming $R^2_{\text{full}} = 1$ seems not reasonable.

8 The calculation of the less restrictive bias-adjusted GPG is further based on the assumption that the bias arising from remaining selection on unobservables does not result in a change of the sign of the correlation between the gender and the index of observables. To ensure the existence of a unique solution, it is necessary to assume equal selection ($\delta = 1$). For more details see Oster (2017).

9 Again, we use $R^2_{\text{full}} = 1.3 \times R^2_{\text{inter}}$. 
\[ \delta = \frac{\text{Degree of Remaining Selection on Unobservables necessary to Result in a Zero GPG}}{\text{Degree of Selection on Observables}} \]  

(8)

Large values of the ratio in (8) can occur for two reasons, either a high degree of remaining selection on unobservables is needed for a zero GPG (high values in the numerator) or as a result of small selection on observables (low values in the denominator). If selection on observables is small, i.e. the set of control variables does not contain all relevant observable factors, a high \( \delta \) could translate, nevertheless, into low (absolute) remaining selection on unobservables necessary to yield a zero GPG. Compared to the baseline model, we would expect lower values for \( \delta \) in case of model specifications selected by the post-double-LASSO estimator, since selection on observables should be higher (or at least not lower) in these cases.

### 2.3 Decomposition

After having obtained the set of potential controls in order to reduce omitted variable bias (Section 2.1) and investigated the robustness with respect to remaining selection on unobservables (Section 2.2), we estimate the wage equations separately for men and women and decompose it following Blinder (1973) and Oaxaca (1973). The GPG is therefore, as in the standard two-fold Oaxaca-Blinder decomposition, decomposed in an endowments and a coefficients component:

\[
\bar{Y}_G - \bar{Y}_G = (\bar{X}_M - \bar{X}_F)\hat{\beta}_M + \bar{X}_F(\hat{\beta}_M - \hat{\beta}_F)
\]

(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. The endowments effect is often referred to as ‘explained’ and the coefficients effect as ‘unexplained’ component of the gap.

Despite the choice of the appropriate set of control variables for estimation of the GPG, there might also be differences in the GPGs across the wage distribution. Indeed, the GPG varies across the wage distribution (Albrecht et al., 2003; Arulampalam et al., 2007). We rely on unconditional quantile regression for the estimation beyond the mean. Unconditional quantile regression allows to extend equation (9) beyond the mean and to estimate it with OLS. The resulting estimates of the endowments and coefficients effect are then quantile-specific.
2.4 Unconditional Quantile Regression: RIF-OLS

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

\[ Y = D\beta + X\gamma + u \] (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, \( X \), on the unconditional quantile, \( Q_\tau \), of the log hourly wage \( Y \) (Firpo et al., 2009).

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

\[ RIF(Y; Q_\tau) = Q_\tau + \frac{\tau - 1\{Y \leq Q_\tau\}}{f_Y(Q_\tau)} \] (11)

where \( f_Y(Q_\tau) \) is the density of the marginal distribution of \( Y \) computed at \( Q_\tau \). \( 1\{Y \leq Q_\tau\} \) is an indicator function that takes value one if the condition in \{\cdot\} is true and zero otherwise. We estimate the RIF separately for each wave \( t \) of the data in order to account for potential changes in the wage distribution over time (DiNardo et al., 1996). 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 a set of 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, \( X \), defined as unconditional quantile regression. The unconditional quantile regression model can then be estimated as a linear probability model and reads as:

\[ RIF(Y; Q_\tau) = X\beta_\tau + \epsilon_\tau \] (12)

where \( X \) is a regression matrix with \( k \) explanatory variables including the constant and \( \beta_\tau \) is the corresponding quantile-specific vector of coefficients. The error term is represented by \( \epsilon_\tau \). 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 conditional mean interpretation of the coefficient estimates.

To take the additional uncertainty of each step of the estimation procedure (calculation of the RIFs) into account, we compute standard errors through bootstrapping with 500 replications. 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 choice of the set of selected controls (post-double-LASSO estimator). Thus, each bootstrap replication relies on the same sets of selected controls.
3 Data

Our empirical analysis uses 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 different household members. It contains extensive information on a broad range of validated behavioral characteristics such as character skills. As the SOEP combines high quality individual level information with information on 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 workers.\footnote{We repeat the analysis over the full period covered in the SOEP as a robustness check in Appendix A.2.}

The dependent variable of our analysis is the natural logarithm of hourly gross earnings. We calculate this variable based on monthly gross earnings and weekly working hours (agreed in contract). To control for labor market characteristics of the individuals, we use information on labor market experience and job tenure. We further control for age, years of education, marital status, migration background, partnership, urban surrounding and the federal state. In order to capture job characteristics, we control for type of contract (limitation yes or no), firm size, union status as well as presence of a workers’ council. Additionally, we include occupational and industrial or sectoral dummies. For classification, we use the ISCO88 (1-Digit) in case of occupations and NACE (2-Digits) for industries and sectors. Further, we include measures for the Big Five Personality Traits, Locus of Control, Reciprocity and Willingness to take a Risk.\footnote{For further details on the measures of character skills applied, see Appendix A.3.}

In order to account for the potential endogeneity between character skills and earnings, in a first step, we restrict our sample to individuals aged 30 to 60 years. Recent research has shown that character skills are only stable for adults in working-age. Changes are found to be concentrated among young or elder individuals [Cobb-Clark and Schurer 2012, 2013]. Further, in order to account for the fact that 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 measure free of age effects [Heineck and Anger 2010].

The main contribution of our paper is the careful identification of the set of control variables. In the first step of our analysis we estimate equation (1) in line with the literature. We refer to this as the baseline model. In this setting, the set of control variables contains actual labor market experience, experience squared, tenure with the current employer, age, age squared, years of education as well as dummies for the survey year, federal state, type of contract, firm size, union status, presence of a workers’ council, occupation and industry, marital status, migration background and urban surrounding.
In order to control for character skills, the model includes measures for the Big Five Personality Traits, Locus of Control, Reciprocity and Willingness to take a Risk. This specific set of controls for character skill is motivated by their common use in the literature and data availability (Collischon, 2017b; Heineck and Anger, 2010; Mueller and Plug, 2006; Nyhus and Pons, 2012). This model specification is supposed to yield an estimate of the conditional GPG in line with recent literature.

Besides the above mentioned variables used in the baseline specification, the data set contains information about the religion of the individual, i.e. dummies for being catholic, protestant, muslim or belonging to an other religion and in case of migrants 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 Maths grades in last record, parental educational attainement, asset income, labor-market experience as self-employed or part-time employee and time of unemployment. Further, we can create controls for life satisfaction, satisfaction with working conditions, 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 a control for interview conditions using average sun hours on the interview day in the federal state (based on data from the German weather service).

Our final analysis sample consists of 7,919 positive wage observations of full-time employed individuals aged between 30 and 60 years (2,604 women and 5,315 men, see Table 1). Table 1 shows descriptive statistics for selected variables separately for men and women. The final column shows the p-value of a t-test for equality of means. In our sample women have significantly lower hourly earnings: women earn on average 15.3% (log approximation) less than men. Regarding human capital measures women outperform men. Women are more educated, and have better (lower) German grades in their last record. If we consider the socio-economic background, the most appealing gender differences occur with respect to marital status. While not even half of all full-time working females are married, the share for males amounts to more than 60%. In case of working conditions the most pronounced differences are, that permanent contracts and union membership are significantly rarer among females. In line with the literature we find particularly pronounced differences with respect to individual character skills (Weisberg et al., 2011). In particular, women are more open, more conscientious, more extraverted, more agreeable and more neurotic. On average female scores for conscientiousness are 0.119 standard deviations higher. For risk-taking and negative reciprocity women score lower on average.

---

13We identify the individual interview condition using indicator variables of the interviewee, its gender and region and match it with the sun hours in that region at the date of the interview. 
Table 1: Descriptive Statistics by Gender

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Women</td>
<td>Std.Dev.</td>
<td>Men</td>
<td>Std.Dev.</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>Dependent Variable</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Hourly Wage</td>
<td>2.786</td>
<td>0.433</td>
<td>2.939</td>
<td>0.406</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td><strong>Baseline Control Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>44.97</td>
<td>8.385</td>
<td>45.07</td>
<td>8.012</td>
<td>0.787</td>
</tr>
<tr>
<td>Education (in years)</td>
<td>12.78</td>
<td>2.642</td>
<td>12.33</td>
<td>2.553</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Married (Dummy)</td>
<td>0.471</td>
<td>0.499</td>
<td>0.641</td>
<td>0.480</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Migration (Dummy)</td>
<td>0.169</td>
<td>0.375</td>
<td>0.161</td>
<td>0.367</td>
<td>0.632</td>
</tr>
<tr>
<td>Urban Area (Dummy)</td>
<td>0.698</td>
<td>0.459</td>
<td>0.700</td>
<td>0.458</td>
<td>0.926</td>
</tr>
<tr>
<td>Permanent Contract (Dummy)</td>
<td>0.915</td>
<td>0.279</td>
<td>0.936</td>
<td>0.245</td>
<td>0.066</td>
</tr>
<tr>
<td>Medium Firm ∈ [20; 200] (Dummy)</td>
<td>0.275</td>
<td>0.446</td>
<td>0.297</td>
<td>0.457</td>
<td>0.267</td>
</tr>
<tr>
<td>Large Firm ≥ 200 (Dummy)</td>
<td>0.537</td>
<td>0.499</td>
<td>0.550</td>
<td>0.498</td>
<td>0.566</td>
</tr>
<tr>
<td>Workers Council (Dummy)</td>
<td>0.593</td>
<td>0.491</td>
<td>0.608</td>
<td>0.488</td>
<td>0.500</td>
</tr>
<tr>
<td>Union Member (Dummy)</td>
<td>0.179</td>
<td>0.383</td>
<td>0.278</td>
<td>0.448</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td><strong>Character Skill Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Openness</td>
<td>0.0379</td>
<td>1.005</td>
<td>-0.157</td>
<td>0.933</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.243</td>
<td>0.880</td>
<td>0.124</td>
<td>0.901</td>
<td><strong>0.002</strong></td>
</tr>
<tr>
<td>Extraversion</td>
<td>0.0523</td>
<td>0.976</td>
<td>-0.188</td>
<td>0.985</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0.0828</td>
<td>0.940</td>
<td>-0.145</td>
<td>0.969</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>0.0351</td>
<td>0.968</td>
<td>-0.207</td>
<td>0.933</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>External Locus of Control</td>
<td>-0.0674</td>
<td>0.964</td>
<td>-0.0048</td>
<td>0.958</td>
<td>0.154</td>
</tr>
<tr>
<td>Risk-taking</td>
<td>-0.239</td>
<td>0.938</td>
<td>0.0674</td>
<td>0.951</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Positive Reciprocity</td>
<td>0.0780</td>
<td>0.996</td>
<td>0.0599</td>
<td>0.923</td>
<td>0.685</td>
</tr>
<tr>
<td>Negative Reciprocity</td>
<td>-0.0986</td>
<td>1.014</td>
<td>0.154</td>
<td>0.999</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td><strong>Grades in Last Record</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German Grade in Last Record</td>
<td>2.454</td>
<td>0.577</td>
<td>2.682</td>
<td>0.586</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>Maths Grade in Last Record</td>
<td>2.607</td>
<td>0.712</td>
<td>2.616</td>
<td>0.701</td>
<td>0.820</td>
</tr>
</tbody>
</table>

Observations 2,604 5,315

Source: SOEP v33. Notes: Calculations use SOEP sample weights. Measures for character skills are standardized. Statistically significant gender differences in bold tested with a t-test for equality of means in bold.
4 Results

In the following, we discuss the sets of selected controls by the post-double-LASSO estimator. Next, we present the estimation results of the (conditional) GPG obtained from the baseline model and both the restricted and unrestricted LASSO models. Further, we analyze the estimated GPGs with respect to robustness to remaining selection on unobservables, and thereby evaluate the capability of the corresponding regression models to control for observables in our data set. Finally, we compare the decomposition results of the three different model specifications (baseline, restricted and unrestricted LASSO).

4.1 Model Selection using the Post-Double-LASSO Estimator

The set of potential controls, among which the LASSO selects includes all control variables that are part of the baseline model. We further take additional controls like religion, country of origin (for migrants), school leaving degree, parental education, last grades in Maths and German, MINT subject, experience as self-employed and part-time worker as well as periods of unemployment, of parental leave, controls for life and work satisfaction working, asset income, number of children, a dummy for children younger than six and sun hours at the interview day in the federal state. After collecting these variables commonly used in and/or suggested for estimation of the GPG, we end up with a set of 111 raw potential controls. For all non-binary potential controls (23 variables), we construct second order polynomials. In a next step, we interact all 134 potential controls with each other. This 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 (3,182 or 35% of the) variables with extremely small standard deviation.\(^{14}\) 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. This leads to the exclusion of further 109 variables and a set of 5,754 potential controls (see Table 2).

In case of the unrestricted LASSO specification, we force no variables to be part of the wage regression. Thus, the LASSO selects among the complete set of 5,754 potential controls.

\(^{14}\)For all potential control variables, we calculate the ratio of the standard deviation to the range between the 99th and 1th percentile, whenever this ratio is smaller than 0.1 (the 1th percentile of the ratio for all potential controls) the corresponding variable is excluded from the set of potential controls. Moreover variables where there is no difference between the 99th and 1th percentile are dropped.
Table 2: Set of Potential Controls

<table>
<thead>
<tr>
<th>Set of controls</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set of raw controls</td>
<td>111</td>
</tr>
<tr>
<td>Set of higher order polynomials</td>
<td>23</td>
</tr>
<tr>
<td>Set of interaction terms</td>
<td>8,911</td>
</tr>
<tr>
<td>Theoretical set of control variables</td>
<td>9,045</td>
</tr>
<tr>
<td>Excluded: small standard deviations</td>
<td>3,182</td>
</tr>
<tr>
<td>Excluded: highly correlated</td>
<td>109</td>
</tr>
<tr>
<td>Final set of potential controls</td>
<td>5,754</td>
</tr>
</tbody>
</table>

The restricted LASSO model specification, on the other hand, is selected while forcing variables that are considered to be indispensable (amelioration set) to be part of the model. The amelioration set consists of all control variables of the baseline model besides the character skill measures and the dummies for occupations and industries. The variables of the amelioration set are forced to be part of the final set of controls by excluding their coefficients from the penalty term in (6).

Tables A.6 to A.9 in Appendix A show the list of selected controls for the two different LASSO specification at the mean, as well as along the wage distribution. Table A.6 (A.8) gives all variables selected by the unrestricted (restricted) Gender-LASSO, while Table A.7 (Table A.9) shows the set of variables chosen by the unrestricted (restricted) Wage-LASSO at the mean as well as along the wage distribution. According to the unrestricted Gender-LASSO, important gender differences are found for part-time experience, German grade in last record, educational attainment, occupations and industries as well as with respect to character skills. Among the nine character skill measures eight are selected (as element of interactions). To be precise, positive reciprocity is the only character skill not selected. The most noticeable difference compared to the baseline model is that the Gender-LASSO selects a huge variety of interaction and second order polynomials, which is in favor of a more flexible functional form. As already stated we use the post-double-LASSO at different points of the wage distribution. Thus, we use the LASSO given in (6) to select important predictors of wages at the mean as well as along selected points of the wage distribution (10th, 50th and 90th quantile). Comparing the set of selected controls by the different wage-LASSOs reveals that different models should be considered at different points of the wage distributions, i.e. wage predictors change along the distribution. The models differ for example with respect to the selected occupations and industries as well as with respect to the selected character skill measures. Again, numerous interactions and higher order polynomials are selected, indicating that we need more flexible model specifications than in case of the baseline model.

The results in case of the restricted
LASSO are line with the findings of the unrestricted LASSO. Over all, model selection based on the post-double-LASSO estimator underlines that for estimation of the GPG, character skills are important control variables, more flexible models should be considered and that the wage predictors change along the wage distribution. This implies that we should use different model specifications at different points of the wage distribution.

According to the regularity condition, $s^2 \ll N$, the post-double-LASSO estimator should not select more than 89 controls. However, in our setting the restricted (unrestricted) LASSO chooses between 117 (117) and 161 (116) control variables. This implies that the regularity condition is not fully met. Therefore, we conduct a sample split as a robustness check in Section 4.5. With sample splitting, the regularity condition becomes $s \ll N$.

### 4.2 Regression Outcome

Table 3 compares the estimated GPGs for different model specifications at the mean as well as along the wage distribution. Panel (A) shows the results for the short regression model, i.e. the raw gaps obtained from regressing the log hourly wages ($y_i$) on a female dummy ($d_i$) only. The (raw) GPG is found to be highest at the bottom of the distribution in Germany for the years 2005, 2009 and 2013. In particular, we find a U-shaped pattern of gender differences in pay with most pronounced differences at the bottom and top compared to the median or mean. In line with the U-shape, the results in Panel (A) suggest both sticky floors and glass ceiling. That is a more than three percentage points higher GPG at the bottom or top compared to the median or 90th or 10th percentile pay gap, respectively ([Arulampalam et al., 2007](#)).

The estimation outcome of the baseline model is given in Panel (B) of Table 3. Here, we find also substantial gender differences in pay across the wage distribution. This is in line with results from the literature ([Arulampalam et al., 2007](#); Collischon, 2017a). Compared to Panel (A) the differential is corrected downwards at all points of the wage distribution. However, the results suggest most pronounced differences at the top of the wage distribution. This implies that, conditional on the set of standard controls (plus character skills measures), the remaining GPG is highest for top-income earners. Further, we find no longer evidence for sticky floors. In contrast, the results suggest again glass ceiling in Germany for the years 2005, 2009 and 2013.

The results of the unrestricted and restricted LASSO specifications are shown in Panels (C) and (D) of Table 3 respectively. The gap at the median and top is corrected downwards compared to the baseline model, while it is slightly corrected upwards at the mean and bottom. Contrary to the results of the baseline model, the more flexible model specifications suggest again evidence for sticky floor but only slight evidence for glass ceiling. Moreover, we find again a U-shaped pattern of the (conditional) GPG across the wage distribution.

---

<table>
<thead>
<tr>
<th>Year</th>
<th>Panel (A)</th>
<th>Panel (B)</th>
<th>Panel (C)</th>
<th>Panel (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Sticky</td>
<td>Glass</td>
<td>Sticky</td>
<td>Glass</td>
</tr>
<tr>
<td>2009</td>
<td>Sticky</td>
<td>Glass</td>
<td>Sticky</td>
<td>Glass</td>
</tr>
<tr>
<td>2013</td>
<td>Sticky</td>
<td>Glass</td>
<td>Sticky</td>
<td>Glass</td>
</tr>
</tbody>
</table>

---

*17*
distribution. Thus, the LASSO specifications, though corrected in size, provide results in line with the raw GPG.

Table 3: Gender Pay Gaps at the Mean & Selected Percentiles, 2005-2013

<table>
<thead>
<tr>
<th>Percentile</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>10th</td>
<td>50th</td>
<td>90th</td>
</tr>
<tr>
<td>Panel A: Raw Gaps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.153***</td>
<td>-0.235***</td>
<td>-0.155***</td>
<td>-0.168***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.039)</td>
<td>(0.020)</td>
<td>(0.030)</td>
</tr>
<tr>
<td></td>
<td>0.071</td>
<td>0.019</td>
<td>0.029</td>
<td>0.012</td>
</tr>
<tr>
<td>Panel B: Baseline Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.102***</td>
<td>-0.108***</td>
<td>-0.106***</td>
<td>-0.149***</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.038)</td>
<td>(0.019)</td>
<td>(0.031)</td>
</tr>
<tr>
<td></td>
<td>0.616</td>
<td>0.282</td>
<td>0.391</td>
<td>0.263</td>
</tr>
<tr>
<td>Panel C: Unrestricted Post-Double-LASSO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.104***</td>
<td>-0.141***</td>
<td>-0.102***</td>
<td>-0.123***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.039)</td>
<td>(0.020)</td>
<td>(0.035)</td>
</tr>
<tr>
<td></td>
<td>0.621</td>
<td>0.338</td>
<td>0.407</td>
<td>0.314</td>
</tr>
<tr>
<td>Panel D: Restricted Post-Double-LASSO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.104***</td>
<td>-0.146***</td>
<td>-0.100***</td>
<td>-0.128***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.038)</td>
<td>(0.020)</td>
<td>(0.034)</td>
</tr>
<tr>
<td></td>
<td>0.632</td>
<td>0.348</td>
<td>0.409</td>
<td>0.327</td>
</tr>
<tr>
<td>Observations</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
</tr>
</tbody>
</table>

Source: SOEP v33. Notes: Calculations use SOEP sample weights. Outcome variable is the natural logarithm of gross hourly earnings. Bootstrapped standard errors (500 replications) in parentheses. *, ** and *** denote significance at the 10%- , 5%- and 1%-level, respectively.
4.3 Robustness to Remaining Selection on Unobservables

In the next step of our analysis, we use the method of Oster (2017) in order to make statements about robustness to remaining selection on unobservables of the estimated (conditional) GPGs. Based on the assumption that selection on observables is informative about selection on unobservables, this method allows to determine the degree of selection on unobservables (relative to selection on observables) that would be necessary to result in a zero GPG ($\delta$). In line with Oster (2017), we consider the adjusted GPG to be robust to remaining selection on unobservables whenever $\delta \geq 1$.

We report the proportionality parameters $\delta$ for the corresponding models in Table 4. All reported values of $\delta$ exceed the threshold of 1. Precisely, $\delta$ ranges between 1.5 and 4.9. For the baseline model, shown in column (1), the proportionality parameter ranges between 1.5 and 4.8. The latter number corresponds to the 90th percentile and would suggest that the estimated GPG at the 90th percentile would turn to zero if remaining selection on unobservables is 4.9 times as high as the selection on observables. This either means that the 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 (2) and (3) of Table 4, respectively. Besides the bottom of the wage distribution, the estimated proportionality parameters are (substantially) 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) one could conclude that the LASSO models control more appropriately for selection on observables.

4.4 Decomposition Results

The decomposition analysis allows 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.

Table 5 shows that, compared to the baseline model, restricted and unrestricted LASSO models explain a substantially larger fraction of the GPG in the middle and at the top of the wage distribution. While at the 90th percentile the explained part in case of the baseline model amounts to 33%, the respective share in case of the unrestricted LASSO equals 49%. The unrestricted and restricted LASSO specification perform similar in explaining the GPG at the median, while the unrestricted LASSO outperforms the restricted LASSO at the mean, bottom, and top of the distribution. At the bottom, the fully pre-defined baseline model performs surprisingly well in explaining gender differences in pay. At this part of the wage distribution, the unrestricted LASSO explains least. This might suggest that the baseline wage model is appropriate for explaining the GPG at the bottom, while

---

16The estimates of the raw GPG are represented in Panel A of Table 3.
in the middle and in particular at the top as well as at the mean more flexible specifications are required.

Table 4: Remaining Selection on Unobservables at the Mean & Selected Percentiles, 2005-2013

<table>
<thead>
<tr>
<th></th>
<th>Panel A: At the Mean</th>
<th>Panel B: 10th Percentile</th>
<th>Panel C: 50th Percentile</th>
<th>Panel D: 90th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Baseline Model</td>
<td>(2) Unrestricted LASSO</td>
<td>(3) Restricted LASSO</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.102***</td>
<td>-0.104***</td>
<td>-0.104***</td>
<td></td>
</tr>
<tr>
<td>Proportionality Parameter (δ)</td>
<td>2.661</td>
<td>2.203</td>
<td>2.187</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.108***</td>
<td>-0.141***</td>
<td>-0.146***</td>
<td></td>
</tr>
<tr>
<td>Proportionality Parameter (δ)</td>
<td>1.495</td>
<td>1.810</td>
<td>1.891</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.106***</td>
<td>-0.102***</td>
<td>-0.100***</td>
<td></td>
</tr>
<tr>
<td>Proportionality Parameter (δ)</td>
<td>2.748</td>
<td>2.047</td>
<td>1.965</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.149***</td>
<td>-0.123***</td>
<td>-0.128***</td>
<td></td>
</tr>
<tr>
<td>Proportionality Parameter (δ)</td>
<td>4.883</td>
<td>2.547</td>
<td>2.705</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
<td></td>
</tr>
</tbody>
</table>

Source: SOEP v33. Notes: Calculations use SOEP sample weights. Outcome variable is the natural logarithm of gross hourly earnings. δ based on the assumption, that the $R^2$ of the full model equals $1.3 \times R^2$ of the Baseline Model. *, ** and *** denote significance at the 10%- , 5%- and 1%-level, respectively.
Table 5: Decomposition of the Raw GPG at the Mean & Selected Percentiles, 2005-2013

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Baseline Model At the Mean</th>
<th>Unrestricted LASSO At the Mean</th>
<th>Restricted LASSO At the Mean</th>
<th>Baseline Model 10th</th>
<th>Unrestricted LASSO 10th</th>
<th>Restricted LASSO 10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw GPG</td>
<td>0.153*** (0.018)</td>
<td>0.153*** (0.018)</td>
<td>0.153*** (0.015)</td>
<td>0.235*** (0.038)</td>
<td>0.235*** (0.032)</td>
<td>0.235*** (0.032)</td>
</tr>
<tr>
<td>Explained</td>
<td>0.062*** (0.017)</td>
<td>0.074*** (0.021)</td>
<td>0.063*** (0.019)</td>
<td>0.129*** (0.032)</td>
<td>0.052 (0.041)</td>
<td>0.030 (0.040)</td>
</tr>
<tr>
<td>Unexplained</td>
<td>0.092*** (0.015)</td>
<td>0.079*** (0.019)</td>
<td>0.091*** (0.017)</td>
<td>0.107** (0.046)</td>
<td>0.183*** (0.041)</td>
<td>0.205*** (0.045)</td>
</tr>
<tr>
<td>% Explained</td>
<td>40.26</td>
<td>48.37</td>
<td>40.91</td>
<td>54.66</td>
<td>22.13</td>
<td>12.77</td>
</tr>
<tr>
<td>% Unexplained</td>
<td>59.74</td>
<td>51.63</td>
<td>59.09</td>
<td>45.34</td>
<td>77.87</td>
<td>87.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Baseline Model 50th</th>
<th>Unrestricted LASSO 50th</th>
<th>Restricted LASSO 50th</th>
<th>Baseline Model 90th</th>
<th>Unrestricted LASSO 90th</th>
<th>Restricted LASSO 90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw GPG</td>
<td>0.155*** (0.019)</td>
<td>0.155*** (0.016)</td>
<td>0.155*** (0.016)</td>
<td>0.168*** (0.029)</td>
<td>0.168*** (0.024)</td>
<td>0.168*** (0.024)</td>
</tr>
<tr>
<td>Explained</td>
<td>0.059*** (0.018)</td>
<td>0.067*** (0.023)</td>
<td>0.068*** (0.025)</td>
<td>0.055* (0.032)</td>
<td>0.082* (0.045)</td>
<td>0.061 (0.051)</td>
</tr>
<tr>
<td>Unexplained</td>
<td>0.095*** (0.019)</td>
<td>0.088*** (0.024)</td>
<td>0.087*** (0.026)</td>
<td>0.113** (0.034)</td>
<td>0.086* (0.047)</td>
<td>0.107*** (0.052)</td>
</tr>
<tr>
<td>% Explained</td>
<td>38.31</td>
<td>43.23</td>
<td>43.87</td>
<td>32.74</td>
<td>48.81</td>
<td>36.31</td>
</tr>
<tr>
<td>% Unexplained</td>
<td>61.69</td>
<td>56.77</td>
<td>56.13</td>
<td>67.26</td>
<td>51.19</td>
<td>63.69</td>
</tr>
<tr>
<td>Observations</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
<td>7,919</td>
</tr>
</tbody>
</table>

Standard error clustered at the individual level in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1.
5 Conclusion

This paper analyzes gender differences in pay in Germany along the wage distribution. The identification of the relevant set of controls for an (as) unbiased (as possible) estimation of the GPG and thus on reducing omitted variable bias is of main interest. We perform model selection with the post-double-LASSO estimator, a machine-learning approach. We attempt to learn more about robustness to remaining selection on unobservable by calculating the degree of remaining selection on unobservables relative to selection on observables necessary to yield a zero GPG and decompose the GPG into an explained and an unexplained part. Both LASSO specifications provide smaller and more persistent estimates of the GPG. Further, we find substantial differences in the estimated GPG, the selected set of controls as well as the explained component of the GPG at different points of the wage distribution.

Model selection with the post-double-LASSO estimator reveals that character skills are important control variables for the estimation of the GPG. Further, the selected models are more flexible than conventional models for estimation of GPGs, i.e. they contain more interactions and higher order polynomials. Besides this, we find that the wage predictors change along the wage distribution. In case of the baseline model we find evidence for glass ceiling, i.e. a particularly pronounced GPG at the top of the distribution. Compared to the baseline model, the models selected by the post-double-LASSO estimator deliver only slight evidence for glass ceiling and suggest a U-shape of gender differences in pay along the wage distribution. The decomposition results show that besides at the bottom of the wage distribution the models selected by the post-double-LASSO estimator explain (substantially) larger shares of the GPG. This is most pronounced at the top of the wage distribution. Overall, we find evidence that more flexible model specifications as well as different model specifications along the wage distribution should be used for estimation of the GPG.

References


— (2017b): “The Returns to Personality Traits across the Wage Distribution,” SOEP-
papers 912.


26.

Firpo, S., N. M. Fortin, and T. Lemieux (2009): “Unconditional Quantile Regress-


## Appendix

### A.1 Set of Selected Controls by LASSO Specifications

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

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

Source: SOEP Questionnaire.

*Notes: (+) indicates a positive relation. (−) indicates a negative relation.*
Table A.2: SOEP Survey Questions Corresponding to Locus of Control

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

Source: SOEP Questionnaire.

Table A.3: SOEP Survey Questions Corresponding to Reciprocity

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

Source: SOEP Questionnaire.
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.

Table A.5: Set of Selected Controls

<table>
<thead>
<tr>
<th>Unrestricted LASSO</th>
<th>Combined Set (Gender + Wage LASSO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender LASSO</td>
<td>71</td>
</tr>
<tr>
<td>Wage LASSO</td>
<td></td>
</tr>
<tr>
<td>- At the mean</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>161</td>
</tr>
<tr>
<td>- 10th Percentile</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>146</td>
</tr>
<tr>
<td>- 50th Percentile</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>137</td>
</tr>
<tr>
<td>- 90th Percentile</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>117</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restricted LASSO</th>
<th>Combined Set (Gender + Wage LASSO+ Amelioration Set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amelioration Set</td>
<td>32</td>
</tr>
<tr>
<td>Gender LASSO</td>
<td>58</td>
</tr>
<tr>
<td>Wage LASSO</td>
<td></td>
</tr>
<tr>
<td>- At the mean</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>125</td>
</tr>
<tr>
<td>- 10th Percentile</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>121</td>
</tr>
<tr>
<td>- 50th Percentile</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>120</td>
</tr>
<tr>
<td>- 90th Percentile</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>117</td>
</tr>
</tbody>
</table>

\[n^{1/2}\] 89
Table A.6: Set of Selected Controls: Unrestricted LASSO, Gender-LASSO

Gender-LASSO:

Part-time experience
Part-time experience interacted with labor market experience, middle secondary degree, Maths grade in last Record, Saxony and Saxony-Anhalt
Last grade in German
Last grade in German interacted with labor market experience and mint subject
Marital status interacted with last grade in German squared

**Neuroticism**
*Neuroticism* interacted with years of education and grades in last school record

Sectors/Industries: Education, Health and Social Work, Basic Metals and Fabricated Metal Products
Education interacted with *agreeableness* and *agreeableness squared*
Health and social work interacted with experience, tenure, sun hours at the interview day, middle secondary degree, permanent contract, small firm, migration, lower secondary degree of the father and *negative reciprocity* squared
Basic metals and fabricated metal products interacted with grades in last record squared
Occupation: Craftsmen
Also interacted with small firm, construction sector, Lower Saxony, grades in last record and *external locus of control* squared

**Further Interactions:**
*Openness* interacted with tenure
*Conscientiousness* interacted with years of education
*Extraversion* interacted with years of education
*Risk-taking* interacted with years of education, relationship status
*Negative Reciprocity* interacted with years of education, presence of a work council
Number of children interacted with qualification for a university of applied sciences, relationship status, marital status, urban surrounding
Mint degree interacted with qualification for a university of applied science and permanent contract
Union status interacted with large firm and relationship status
Marital status interacted with being catholic

Sectors/Industries:
Machinery and Equipment n.e.c interacted with grades in German in last record squared
Construction interacted with lower secondary degree, NRW and grades in last record
Transport Storage and Communication interacted with lower secondary degree
Real estate, renting and business activities interacted with BW
Occupations:
Technicians interacted with experience unemployment, education and health and social work
Office workers interacted with years of education, and lower secondary degree
Service Occupations interacted with health and social work
Assistants interacted with last Maths and German grade in last record squared, years of education squared and experience
Interviewer of the same gender interacted with last math grade, large firm and union status

Table A.7: Set of Selected Controls: Unrestricted LASSO, Wage-LASSO at the Mean & Selected Percentiles

Wage-LASSO at the mean:

Years of education
Years of education interacted with experience unemployment, tenure, agreeableness, large firm, presence of a workers council, coke and chemicals industry, assistants, Bavaria, income out of assets and survey year
Years of education squared interacted with number of children, permanent contract
Experience unemployment
Experience unemployed interacted with Maths grade in last record, middle secondary degree, relationship status, office worker and life satisfaction
External locus of control
External locus of control interacted permanent contract
Risk-taking
Risk-taking interacted with German grade in last record, permanent contract, and sun hours on the interview day
Agreeableness interacted with urban surrounding
Neuroticism interacted with relationship status
Openness squared interacted with machine operator
Positive reciprocity interacted with part-time experience
Small firm
Small firm interacted with lower secondary degree, Saxony-Anhalt, Service Occupation
Year dummies
Year dummies interacted with being catholic, presence of a workers council, Bavaria, German grade in last record, experience
Sector/Industry: Wholesale and retail
Occupations: Scientists, office workers, service occupations
Scientist and service occupations interacted with age
Service occupations interacted with life satisfaction
Further interactions
Maths grade in last record interacted with part-time experience
German grade in last record interacted with tenure and mint subject
Age interacted with life satisfaction
Number of children interacted with experience and academic qualification
Being catholic interacted with tenure
Qualification for a university of applied sciences interacted with relationship status
Large firm interacted with tenure
Presence of a workers’ council interacted with experience and permanent contract
Marital status interacted with being catholic, qualification for a university of applied science and large firm
Hessen interacted with urban surrounding
BW interacted with secondary degree of mother (lower) and permanent contract
Bavaria interacted with number of children
Brandenburg interacted with grades in last record
Saxony interacted with age, middle secondary degree, grades in last record
Saxony-Anhalt interacted with grades in last record
Income out of assets interacted with permanent contract and presence of a workers council
Sectors/Industries
Coke and chemicals interacted with tenure and large firm
Basic metals and fabricated metal products interacted with large firm and experience
Electrical and optical equipment interacted with tenure and permanent contract
Transport equipment interacted with large firm and presence of a workers’ council
Financial Intermediation interacted with permanent contract and urban surrounding
Public administration and defence interacted with highest secondary degree and large firm
Occupations:
Technicians interacted with transport equipment
Craftsmen interacted with middle secondary degree
Machine Operator interacted with part-time experience
Assistants interacted with age and permanent contract
Wage-LASSO at the 10th Percentile:

```
exper_part note_d experXnote_d mintXnote_d z_neuro_adj sec_fil23 sec_fil24 isco88_7 note_mXexper_part experXexper_part
z_open_adjXtenure_fil z_cons_adjXyears_educ z_extra_adjXyears_educ
z_neuro_adjXyears_educ z_RISK_adjXyears_educ z_REC_neg_adjXyears_educ
relation2Xz_RISK_adj
relation2Xnum_kids
degree_rsXexper_part degree_rsXdegree_aca degree_fhXnum_kids degree_fhXmint
contract_permXmint
unionXrelation2 unionXfirm_large
wor counXz_REC_neg_adj
sec_fil10Xnote_d
sec_fil16Xdegree_hs
sec_fil19Xdegree_hs
sec_fil23Xz_agree_adj
ec_fil24Xexper sec_fil24Xtenure_fil sec_fil24Xsun_hours sec_fil24Xdegree_rs
sec_fil24Xcontract_perm sec_fil24Xfirm_small
isco88_3Xexper_unem isco88_3Xsec_fil23 isco88_4Xsec_fil24
isco88_4Xyears_educ isco88_4Xdegree_rs
isco88_5Xsec_fil24
isco88_7Xfirm_small isco88_7Xsec_fil16
isco88_8Xnote_d isco88_8Xnote_m isco88_8Xexper
marriedXnum_kids marriedXcath
migrationXsec_fil24
urbanXnum_kids bl3Xisco88_7
bl5Xsec_fil16 bl8Xsec_fil21 bl14Xexper_part bl15Xexper_part HSdegree_fathXsec_fil24
note_missXz_neuro_adj note_missXsec_fil16 note_missXisco88_7 samesex_intXnote_d
samesex_intXfirm_large samesex_intXunion note_d2Xsec_fil11 note_d2Xmarried
note_m2Xsec_fil10 note_m2Xisco88_8 years_educ2Xisco88_8 z_agree_adj2Xsec_fil23
z_LOC_ext_adj2Xisco88_7 z_REC_neg_adj2Xsec_fil24
```
Wage-LASSO at the 50th Percentile:

\[
\begin{align*}
z_{\text{sat19}} & \text{adj mint firm\_small sec\_fil17 isco88\_5 note\_mXexper\_unem} \\
\text{years\_educXexper\_unem years\_educXtenure\_fil} \\
z_{\text{LOC\_ext\_adjXyears\_educ}} & \text{num\_kidsXexper num\_kidsXtenure\_fil} \\
mintXnote\_d degree acax\_catheX\text{degree\_fhXexper} \\
\text{contract\_permXz\_RISK\_adj} & \text{firm\_smallXdegree\_rs firm\_largeXxeper} \\
\text{firm\_largeXtenure\_fil firm\_largeXyears\_educ wor\_comX\text{Xtenure\_fil}} \\
\text{sec\_fil7Xtenure\_fil sec\_fil7Xfirm\_large} \\
\text{sec\_fil11Xexper sec\_fil11X\text{tenure\_fil}} \\
\text{sec\_fil13Xfirm\_large} \\
\text{sec\_fil15Xsun\_hours} \\
\text{sec\_fil17Xexper\_part} \\
\text{sec\_fil20Xnote\_d sec\_fil20Xsun\_hours sec\_fil20Xcontract\_perm} \\
\text{sec\_fil\_other\_Xnote\_d} \\
isco88\_2X\text{age} \\
isco88\_3Xsec\_fil13 \\
isco88\_4Xexper\_unem \\
isco88\_5Xfirm\_small isco88\_5Xsec\_fil24 \\
isco88\_7X\text{degree\_rs} \\
isco88\_8X\text{exper\_part} \\
isco88\_9X\text{age isco88\_9Xyears\_educ marriedX\text{degree\_fh marriedXfirm\_large migraXtionXfirm\_small urbanXsec\_fil20 bl13Xdegree\_rs bl14Xdegree\_rs bl14Xisco88\_7 bl14Xmarried bl15Xnote\_d H\text{degree\_mothXbl8 assetsXyears\_educ assetsXsun\_hours assetsXcontract\_perm z\_sat19\_adj2Xfirm\_small note\_d2Xbl14 note\_d2Xassets note\_m2Xbl12 years\_educXrelaXtion2 years\_educXcontract\_perm years\_educXwor\_com years\_educXisco88\_8 years\_educXurban years\_educX\text{wave9 z\_open\_adj2Xisco88\_8 z\_cons\_adj2Xbl15 z\_agree\_adj2Xbl14 z\_neuro\_adj2Xmint z\_neuro\_adj2Xisco88\_5 z\_neuro\_adj2Xisco88\_9}}
\end{align*}
\]
Wage-LASSO at the 90th Percentile:

\[
\begin{align*}
\text{sun\_hours}_z \cdot \text{RISK\_adj} & \\
mintXnote\_d \cdot \text{mintXexper} & \cdot \text{mintXnum\_kids} \\
\text{relation2Xz\_LOC\_ext\_adj} & \cdot \text{relation2Xz\_RISK\_adj} \cdot \text{firm\_largeXz\_neuro\_adj} \\
\text{firm\_largeXz\_RISK\_adj} & \cdot \text{z\_open\_adj2Xisco88\_8} \cdot \text{z\_agree\_adj2Xbl15} \\
\text{z\_neuro\_adj2Xisco88\_8} & \\
degree\_acaXnum\_kids & \cdot degree\_acaXcath \\
\text{contract\_permXexper\_unem} & \cdot \text{contract\_permXz\_RISK\_adj} \\
\text{firm\_smallXage} & \cdot \text{firm\_smallXrelation2} \cdot \text{firm\_largeXcivil\_ser} \cdot \text{firm\_largeXdegree\_aca} \\
\text{sec\_fil22Xnote\_d} & \cdot \text{s\_ec\_fil22Xrelation2} \cdot \text{sec\_fil22Xdegree\_aca} \cdot \text{sec\_fil22Xfirm\_large} \\
isco88\_2Xexper & \cdot \text{isco88\_2Xnum\_kids} \cdot isco88\_2Xcontract\_perm \cdot isco88\_2Xfirm\_large \\
isco88\_3Xsec\_fil23 & \cdot isco88\_3Xsec\_fil24 \\
isco88\_5Xdegree\_rs & \cdot isco88\_5Xsec\_fil24 \\
isco88\_7Xdegree\_rs & \\
isco88\_9Xexper & \cdot isco88\_9Xfirm\_large \cdot isco88\_9Xwor\_coun \\
migrationXisco88\_8 & \cdot \text{urbanXdegree\_aca} \cdot \text{bl13Xdegree\_rs} \cdot \text{bl14Xnote\_d} \cdot \text{bl14Xexper} \\
bl14Xtenure\_fil & \cdot \text{bl14Xnum\_kids} \cdot \text{assetsXcivil\_ser} \cdot \text{assetsXurban} \cdot \text{note\_missXsec\_fil22} \\
\text{note\_missXisco88\_7} & \cdot \text{note\_missXbl13} \cdot \text{note\_m2Xbl15} \cdot \text{years\_educ2Xcontract\_perm} \\
\text{years\_educ2Xfirm\_large} & \cdot \text{years\_educ2Xmarried} \cdot \text{years\_educ2Xassets} 
\end{align*}
\]
Table A.8: Set of Selected Controls: Restricted LASSO, Gender-LASSO

Gender-LASSO:

<table>
<thead>
<tr>
<th>Part-time experience</th>
<th>Last Grade in German</th>
<th>sec_fil23</th>
<th>isco88_3</th>
<th>isco88_7</th>
</tr>
</thead>
</table>

Note: 
- mXexper_part
- experXpart
- experXnote_d
- ageXexper
- z_open
- adjXtenure
- fil
- z_cons
- adjXyears
- educ
- z_extra
- adjXyears
- educ
- z_neuro
- adjXyears
- educ
- z_RISK
- adjXyears
- educ
- z_REC
- neg
- adjXyears
- educ
- z_sat
- adj2Xsec
- fil16
- note_d2Xsec
- fil11
- years
- educ
- 2Xisco88
- 8
- z_agree
- adj2Xsec
- fil23
- z_LOC
- ext
- adj2Xisco88
- 7
- z_REC_neg
- adj2Xsec
- fil24
num_kidsXexper
mXexper_d
relation2Xz
- cons
- adj
relation2Xz
- RISK
- adj
relation2Xnum
- kids
degreeXrsXexperXpart
degreeXfhXmintXcontract
permXnote_d
contract
permXmint
sec_fil19Xdegree
hs
sec_fil23Xz
agree
adj
sec_fil24Xexper
sec_fil24Xtenure
fil
sec_fil24Xdegree
rs
secFil24Xfirm
small
isco88
3Xsec
fil23
isco88
4Xexper
isco88
4Xyears
educ
isco88
4Xdegree
rs
isco88
5Xsec
fil24
isco88
7Xdegree
rs
isco88
7Xfirm
small
isco88
7Xsec
fil16
isco88
8Xnote
isco88
8Xnote
m
isco88
8Xexper
migrationXsec
fil24
bl8Xsec
fil21
bl14Xz
neuro
adj
HSdegreeXfathXsec
fil24
note
miss
Xz
neuro
adj	note
miss
Xsec
fil16
note
miss
isco88
7
cityXsec
fil19
samesex
int
Xnote_d
samesex
int
Xisco88
8
samesex
int
Xbl15
Table A.9: Set of Selected Controls: Restricted LASSO, Wage-LASSO at the Mean & Selected Percentiles

Wage-LASSO at the mean:

<table>
<thead>
<tr>
<th>z_RISK_adj</th>
<th>z_LOC_ext_adjXage</th>
<th>sun_hoursXz_RISK_adj</th>
<th>contract_permXz_LOC_ext_adj</th>
<th>contract_permXz_RISK_adj</th>
<th>urbanXz_agree_adj</th>
<th>z_open_adj2Xisco88_8</th>
<th>z_neuro_adj2Xisco88_9</th>
<th>z_RISK_adj2Xisco88_5</th>
<th>contract_permXz_sat11_adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>assets num_kidsXyears_educ</td>
<td>sec_fil17Xyears_educ</td>
<td>sec_fil11Xexper</td>
<td>sec_fil13Xfirm_large</td>
<td>sec_fil13Xwor_coun</td>
<td>sec_fil17Xyears_educ</td>
<td>sec_fil22Xdegree_abi</td>
<td>sec_fil22Xfirm_large</td>
<td>isco88_2Xage</td>
<td>isco88_3Xsec_fil13</td>
</tr>
</tbody>
</table>

Wage-LASSO at the 10th Percentile:

exper_unem isco88_5 note_mXexper_unem experXexper_unem relation2Xexper_unem civil_serXexper_part degree_acaXcath degree_rsXexper_unem firm_mediumXmint | sec_fil22Xdegree_rs | sec_fil23Xdegree_rs | isco88_2Xsun_hours | isco88_2Xfirm_medium | isco88_3Xage | isco88_3Xcivil_ser | isco88_3Xdegree_rs | isco88_3Xfirm_medium | isco88_5Xz_RISK_adj | isco88_5Xfirm_small | isco88_9Xexper_unem | urbanXfirm_large | bl14Xdegree_aca | bl14Xfirm_large | bl15Xisco88_2 | assetsXisco88_3 | note_missingXsec_fil23 | note_missingXisco88_5 | note_d2Xsec_fil16 | years_educ2Xfirm_large | years_educ2Xisco88_9 | z_RISK_adj2Xisco88_5 |

Wage-LASSO at the 50th Percentile:

Number of Kids

| sec_fil17 | sec_fil20 | contract_permXz_RISK_adj | sec_fil8Xfirm_large | sec_fil11Xexper | sec_fil11Xtenure_fil | sec_fil13Xfirm_large | sec_fil15Xsun_hours | sec_fil17Xnote_d | sec_fil20Xnote_d | sec_fil20Xsun_hours | sec_fil_otherXtenure_fil | isco88_2Xage | isco88_2Xdegree_rs | isco88_3Xsec_fil13 | isco88_5Xcath | isco88_5Xcontract_perm | isco88_5Xsec_fil24 | isco88_8Xexper_part | isco88_9Xyears_educ | isco88_9Xwor_coun | marriedXexper_unem | assetsXsun_hours | assetsXcontract_perm | note_missingXz_sat11_adj | samesex_intXz_LOC_ext_adj | years_educ2Xisco88_5 | years_educ2Xisco88_8 | years_educ2Xisco88_9 | z_neuro_adj2Xmint | z_neuro_adj2Xisco88_9 |

36
Wage-LASSO at the 90th Percentile:

| zLOC_ext_adjXexper | sun_hoursXz_RISK_adj | mintXnum_kids | relation2Xz_RISK_adj | civil_serXtenure_fil | civil_serXyears_educ | degree_aXdegree_aXnum_kids | degree_aXcivil_serXdegree_abiXexper_unem | firm_largeXz_RISK_adj | firm_largeXcivil_serXdegree_aXnum_kids | firm_largeXdegree_aXwor_counXcivil_ser | sec_fil22Xnote_d | sec_fil22Xrelation2 | sec_fil22Xdegree_aXsec_fil22Xfirm_large | isco88_2Xexper | isco88_2Xnum_kids | isco88_2Xfirm_large | isco88_3Xsec_fil23 | isco88_9Xfirm_large | urbanXcivil_serXdegree_aXterm | bl14Xdegree_aXassets | Xurban | note_missXsec_fil22 | z_sat11_adj2Xisco88_8 | years_educX2Xsec_fil22 | years_educX2Xassets |

A.2 Robustness Analysis: Estimation over the Period 1984-2016

TO BE ADDED

A.3 Measures for Character Skills

The Big Five are a widely used approach to measure personality, which 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 of 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 beliefs 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 8 statements the same seven-point Likert scale as in case of the Big Five. Table A.2 shows the eight different statement 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 were 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. 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 2017b; Heineck and Anger 2010).

\footnote{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 to A.4. In case of all measures the choice of the corresponding items is confirmed.}