On the sensitivity of wage gap decompositions

Martin Huber and Anna Solovyeva

Department of Economics, University of Fribourg, Switzerland

Abstract: This paper investigates the sensitivity of average wage gap decompositions to methods resting on different assumptions regarding endogeneity of observed characteristics, sample selection into employment, and estimators' functional form. Applying five distinct decomposition techniques to estimate the gender wage gap in the U.S. using data from the National Longitudinal Survey of Youth 1979, we find that the magnitudes of the wage gap components are generally not stable across methods. Furthermore, the definition of the observed characteristics matters: merely including their levels (as frequently seen in wage decompositions) entails smaller explained and larger unexplained components than when both their levels and histories are included in the analysis. Given the sensitivity of our results, we advise caution when using wage decompositions for policy recommendations.

Keywords: wage decomposition, gender wage gap, causal mechanisms, mediation.

JEL classification: C14, C21, J31, J71.

We have benefited from comments by conference participants in Landeck-Zams (Ski and Labor Seminar 2017) and Bern (Conference on Discrimination in the Labor Market 2017), as well as by seminar participants in Zürich (ETH), Oslo (Frisch Centre), and Ispra (European Commission Competence Centre on Microeconomic Evaluation). Addresses for correspondence: Martin Huber, Anna Solovyeva, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch, anna.solovyeva@unifr.ch.

1 Introduction

A vast empirical literature is concerned with the analysis and decomposition of gender wage gaps. Blinder (1973) and Oaxaca (1973) (see also Duncan (1967)) suggested a linear method allowing disentangling the total gap into an explained part that is linked to differences in observed characteristics, for instance education, and an unexplained part that is linked to unobserved factors, for instance discrimination. Several studies proposed non-parametric decomposition methods dropping the linearity assumptions, see for instance DiNardo, Fortin, and Lemieux (1996), Barsky, Bound, Charles, and Lupton (2002), Frölich (2007), Mora (2008), and Ñopo (2008). Finally, another branch of the literature suggested decomposition methods at quantiles (rather than means) of the wage distribution, see for instance Juhn, Murphy, and Pierce (1993), DiNardo, Fortin, and Lemieux (1996), Machado and Mata (2005), Melly (2005), Firpo, Fortin, and Lemieux (2007), Chernozhukov, Fernandez-Val, and Melly (2009), and Firpo, Fortin, and Lemieux (2009).

The aforementioned methods ignore the potential endogeneity of the observed characteristics, which are typically 'bad controls' in the sense of Angrist and Pischke (2009) as they are determined later in life, i.e. after gender. This implies that the explained and unexplained parts do not correspond to the true causal mechanisms related to observed and unobserved factors, respectively, through which gender influences wage. For this reason, policy conclusions – for instance about the magnitude of discrimination – are difficult to derive from such conventional decompositions, see Kunze (2008), Huber (2015), and Yamaguchi (2014) for related criticisms. Using an approach that comes from the literature on nonparametric causal mediation analysis (see for instance Robins and Greenland (1992) and Pearl (2001)), Huber (2015) controls for observed confounders at birth as one possible approach to improve upon the endogeneity issue. However, a further threat to identification is sample selection (see Heckman (1976) and Heckman (1979)) due to the fact that wages are only observed for those who work. For this reason, Neuman and Oaxaca (2003) and Neuman and Oaxaca (2004) combine classic decompositions with Heckman-type sample selection correction.[1] Alternatively, Maasoumi and Wang (2016) apply the copula approach of Arellano and Bonhomme (2010) to model the joint distribution of the quantile of the wage distribution and selection. In the presence of panel data, Blau and Kahn (2006) and Olivetti and Petrongolo (2008)[2] consider proxying

[1] See also the method of Machado (2017), which permits arbitrary unobserved heterogeneity in the selection process.

[2] Olivetti and Petrongolo (2008) also estimate the Manski bounds (Manski (1989)) on the distribution of wages, using the actual and the imputed wage distributions. Bičakova (2014) derives bounds on gender unemployment gaps.

non-observed wages by the observed wage in the closest period.[3] Finally, few studies aim at controlling for both endogeneity and sample selection. García, Hernández, and López-Nicolás (2001) combine instrumental variable regression to control for the endogeneity of one of the observed characteristics (education) with Heckman-type sample selection correction in a parametric framework. The more flexible causal mediation method by Huber and Solovyeva (2018) aims at tackling endogeneity by conditioning on observed potential confounders and sample selection by controlling for the selection probability based on observables and/or instruments.

In this paper, we investigate the sensitivity of average wage gap decompositions to various methods ignoring and considering endogeneity and sample selection, to provide insights on the robustness of decompositions across identifying assumptions. To this end, we consider U.S. wage data collected in the year 2000 coming from the National Longitudinal Survey of Youth 1979 (NLSY). The latter is a panel study of young individuals in the U.S. aged 14 to 22 years in 1979. The analysed estimators include the Oaxaca-Blinder decomposition; semiparametric inverse probability weighting (IPW, see Hirano, Imbens, and Ridder (2003)), which eases linearity but ignores endogeneity and sample selection just as the Oaxaca-Blinder decomposition; IPW controlling for potential confounders at birth to mitigate endogeneity as in Huber (2015) but ignoring sample selection; and the approaches proposed in Huber and Solovyeva (2018) to tackle both endogeneity and sample selection.

We find that the explained and unexplained wage gap components are generally not stable across methods. Even the total gap estimates differ non-negligibly between methods ignoring and controlling for sample selection. Although we do not claim that any of the estimators is capable of fully tackling identification concerns, our results cast doubts about the usefulness of standard decompositions used in the vast majority of empirical studies, which ignore endogeneity and sample selection altogether. We also investigate the robustness of our findings w.r.t. the definition of the observed characteristics. In our main specification, we include both levels as well as histories of such characteristics (e.g., current occupation as well as years in current occupation). In a robustness check, we only keep the levels and omit histories (as it appears to be the convention in many decompositions) and find this to reduce the explained and increase the unexplained component across our estimators. In light of the sensitivity of some of our results w.r.t. methods and variable definitions, we advise caution when basing policy recommendations (which typically require a proper identification of the causal mechanisms underlying the wage gap) on the outcomes of wage

[3] As an alternative use of panel data, Lemieux (1998) combines fixed effect estimation with decomposition methods and allows for heterogeneity of the return to fixed effects across groups. However, this strategy depends on individuals switching groups, which is rarely the case for gender.

decompositions. This seems important given that the empirical literature on wage decompositions appears to have paid comparably little attention to identification issues that may jeopardize the interpretability of the parameters of interest.

Goraus, Tyrowicz, and van der Velde (2015) provide a further study systematically investigating the robustness of wage gap decompositions across specifications, considering the Polish Labor Force Survey. The authors compare estimates of the unexplained component across parametric and nonparametric methods for both means and quantiles. They also analyze issues of common support (or overlap) in observed characteristics across females and males and selection into employment based on Heckman-type sample selection corrections. Their results suggest that enforcing versus not enforcing common support in the characteristics has a non-negligible impact on the estimates. Also our IPW procedures enforce common support by specific trimming rules to ensure the comparability of observations across gender and employment states in terms of observables. The sample selection corrections, on the other hand, barely affect estimates of the unexplained component in Goraus, Tyrowicz, and van der Velde (2015). We also find that our weighting-based sample selection corrections change the unexplained component moderately when compared to IPW controlling for potential confounders alone, while more variation is observed for the total wage gap and the explained component. We point out that one major distinction of our study and Goraus, Tyrowicz, and van der Velde (2015) is that they do not consider methods that control for confounders at birth to tackle the endogeneity of the observed characteristics.

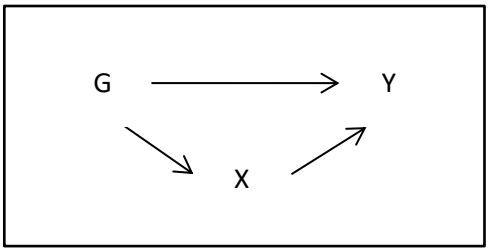
The remainder of this paper is organized as follows. Section 2 formally discusses the econometric parameters of interest and the identifying assumptions required for the various methods considered to consistently decompose wage gaps into observed and unobserved causal mechanisms. Section 3 discusses the NLSY data, sample definition, and descriptive statistics. Section 4 presents and interprets the estimation results. Section 5 concludes.

2 Identification

Fortin, Lemieux, and Firpo (2011) pointed out that while it is standard in econometrics to first discuss identification and then introduce appropriate estimators, most studies in the field of wage gap decompositions go directly to estimation without clarifying identification first. Here, we first define what in our opinion should be the parameters of interest to be able to derive useful policy recommendations. To this

end, let G denote a binary group dummy for gender, Y the outcome of interest (e.g., log wage) and X the vector of observed characteristics (e.g., education, work experience, occupation, industry, and others). We assume that G causally precedes X, which appears intuitive as gender is determined even prior to birth, while X is determined by decisions later in life. G might influence Y 'indirectly' via its effect on X, i.e. by a causal mechanism related to observed characteristics. For instance, gender may have an effect on wage because females and males select themselves into different occupations. G might affect Y also 'directly', i.e. through factors not observed by the researcher such that they do not appear in X. For instance, gender could have an impact on the perception of individual traits by decision makers in the labor market (see Greiner and Rubin (2011)), which in turn may entail discriminatory behavior. A graphical representation of this causal framework is given in Figure 1, where arrows represent causal effects: G influences Y either through X or 'directly'.

Figure 1: A graphical representation of the decomposition under Assumption 1



For a formal definition of the causal mechanisms running through observed characteristics X and unobserved factors as parameters of interest, we denote by $Y(g)$ and $X(g)$ the potential outcomes and characteristics when exogenously setting gender G to a specific g, with $g \in \{1, 0\}$.[4] $E(X(1)) - E(X(0))$ gives the average causal effect of G on X (represented by the arrow of G to X in Figure 1), so to speak the 'first stage' of the indirect effect. $E(Y(1)) - E(Y(0))$, on the other hand, gives the total average causal effect of G on Y, represented by the sum of direct and indirect (i.e. operating through X) effects. Following the causal mediation literature, see Robins and Greenland (1992) and Pearl (2001), we further refine the potential outcome notation to be able to distinguish between the causal mechanisms in Figure 1: Let $Y(g) = Y(g, X(g))$, to make explicit that the potential outcome is affected by the group variable both directly and indirectly via $X(g)$. This permits rewriting the total effect of G on Y as $E(Y(1)) - E(Y(0)) = E[Y(1, X(1))] - E[Y(0, X(0))]$ and more importantly, it allows disentangling the latter into the causal mechanisms of interest. That is, the difference in potential outcomes due to a switch

[4]See for instance Rubin (1974) for an introduction to the potential outcome framework.

from $X(1)$ to $X(0)$ while keeping gender fixed at $G = 1$ yields the indirect effect (denoted by ψ), while varying gender and fixing characteristics at $X(0)$ gives the direct effect (η). Both together add up to the total causal effect:

$$E[Y(1,X(1))] - E[Y(0,X(0))] = \underbrace{E[Y(1,X(1))] - E[Y(1,X(0))]}_{\psi} + \underbrace{E[Y(1,X(0))] - E[Y(0,X(0))]}_{\eta}. \quad (1)$$

We now introduce the first identifying assumption considered in our empirical analysis, which rules out endogeneities of G, X and sample selection issues.

Assumption 1 (sequential independence):

(a) $\{Y(g',x), X(g)\} \perp G$ for all $g', g \in \{0,1\}$ and x in the support of X,

(b) $Y(g',x) \perp X|G = g$ for all $g', g \in \{0,1\}$ and x in the support of X,

(c) $Y(g,X)$ is linear X for $g \in \{0,1\}$,

(d) $\Pr(G = 1|X = x) > 0$ for all x in the support of X,

where '$\perp$' denotes statistical independence. Under Assumption 1(a), G is as good as randomly assigned, i.e. there are no factors confounding G on the one hand and Y and/or X on the other hand. Under Assumption 1(b), observed characteristics like education are as good as randomly assigned within gender, i.e. given G, so that there are no factors confounding X and Y. Assumption 1(c) imposes potential outcomes to be linear in X. Finally, Assumption 1(d) is a common support restriction. It implies that the conditional probability (the so-called propensity score) to belong to the reference group ($G = 1$), e.g., males, is larger than zero for any value in the support of X, such that for each female observation ($G = 0$), there exists a male who is comparable w.r.t. X.

The Oaxaca-Blinder decomposition consistently estimates ψ and η under Assumptions 1(a)-1(c). To see this, note that under Assumption 1(a), $E(X(g)) = E(X|G = g)$. Under Assumptions 1(a), 1(b), and 1(c), $E[Y(g,x)] = E(Y|G = g, X = x) = c_g + x\beta_g$, where c_g denotes a gender-specific constant and β_g denotes a vector of gender-specific coefficients on X in the respective female or male population. Finally, by iterated expectations, $E[Y(g,X(g'))] = c_g + E(X|G = g')\beta_g$ for $g, g' \in \{0,1\}$. Therefore,

$$\psi = E[Y(1,X(1))] - E[Y(1,X(0))] = [E(X|G = 1) - E(X|G = 0)]\beta_1, \quad (2)$$

$$\eta = E[Y(1,X(0))] - E[Y(0,X(0))] = c_1 - c_0 + E(X|G = 0)(\beta_1 - \beta_0). \quad (3)$$

The left hand expressions in (2) and (3) correspond to the probability limits of the explained and unexplained components, respectively, in the Oaxaca-Blinder decompositions. For (2) and (3) to hold, Assumptions 1(a) and 1(b) could be relaxed to mean independence, while full independence needs to be maintained for decompositions of quantiles.[5]

Nonparametric approaches do not rely on the linearity assumption 1(c), but instead require common support as postulated in Assumption 1(d). This becomes obvious from considering the denominators of the following expressions based on inverse probability weighting (IPW) by the propensity score, which identify the parameters of interest as discussed in Huber (2015):

$$\psi = E\left[\frac{Y \cdot G}{\Pr(G=1)}\right] - E\left[\frac{Y \cdot G}{\Pr(G=1|X)} \cdot \frac{1-\Pr(G=1|X)}{1-\Pr(G=1)}\right], \tag{4}$$

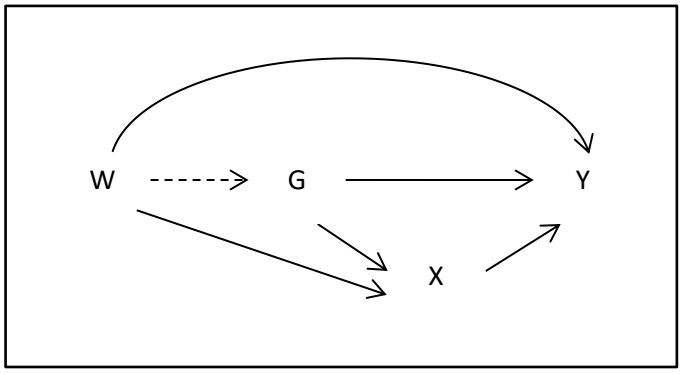
$$\eta = E\left[\frac{Y \cdot G}{\Pr(G=1|X)} \cdot \frac{1-\Pr(G=1|X)}{1-\Pr(G=1)}\right] - E\left[\frac{Y \cdot (1-G)}{1-\Pr(G=1)}\right]. \tag{5}$$

(5) is identical to the identification result for the average treatment effect on the non-treated (see Hirano, Imbens, and Ridder (2003) for IPW-based treatment evaluation in subgroups based on reweighing), even though the causal framework differs. In classic treatment evaluation, one typically controls for pre-treatment (or pre-group) variables to tackle the endogeneity of the treatment (or group). Here, X are post-group variables such that conditioning allows separating the indirect causal mechanism via X from the direct one related to unobservables. Obviously, this is only feasible if neither G nor X given G are endogenous as postulated in Assumption 1. In the empirical application presented in Section 4, we consider both the Oaxaca-Blinder decomposition and estimation based on the sample analogues of (4) and (5).

In a next step, we ease Assumption 1 by assuming that the identifying restrictions need not hold unconditionally, but conditional on a set of observed covariates measured at birth and denoted by W. This allows for endogeneity of X, as long as it can be tackled by W. The dashed arrow going from W to G in Figure 2 even points to the possibility of an endogenous G. This may appear unnecessary when assuming gender to be randomly assigned by nature. However, specific interventions like selective abortions could in principle jeopardize randomization, which is permitted in Assumption 2 below as long as W captures all confounding.

[5]However, analogous results to (2) and (3) cannot be applied to quantile decompositions, because the law of iterated expectations does not apply, see Fortin, Lemieux, and Firpo (2011).

Figure 2: A graphical representation of the decomposition under Assumption 2



Assumption 2 (sequential conditional independence):

(a) $\{Y(g',x), X(g)\} \perp G|W$ for all $g', g \in \{0,1\}$ and x in the support of X,

(b) $Y(g',x) \perp X|G=g, W=w$ for all $g', g \in \{0,1\}$ and x, w in the support of X, W,

(c) $\Pr(G=1|X=x, W=w) > 0$ and $0 < \Pr(G=1|W=w) < 1$ for all x, w in the support of X, W.

Identical or similar conditions as Assumption 2 have been frequently applied in the literature on causal mediation analysis, see for instance Pearl (2001), and Imai, Keele, and Yamamoto (2010). Assumptions 2(a) and (b) imply that after controlling for W, no unobserved variables confound either G and Y, G and X, or X and Y given G. Assumption 2(c) is a refined common support restriction, requiring that the conditional probability of belonging to the reference group given X, W is larger than zero, while the conditional probability given W must neither be zero nor one. The latter implies that for each female in the population, the exists a comparable observation in terms of W among males and vice versa. Under Assumption 2, it follows from the results on IPW-based identification of direct and indirect effects in Huber (2014) that

$$\psi = E\left[\frac{Y \cdot G}{\Pr(G=1|W)}\right] - E\left[\frac{Y \cdot G}{\Pr(G=1|X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right], \quad (6)$$

$$\eta = E\left[\frac{Y \cdot G}{\Pr(G=1|X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right] - E\left[\frac{Y \cdot (1-G)}{1-\Pr(G=1|W)}\right]. \quad (7)$$

Estimation of (ethnic) wage gaps based on (6) and (7) has been considered in Huber (2015) and is also among the methods investigated in our empirical application presented further below.

The approaches discussed so far abstract from sample selection stemming from the issue that wages are only observed for individuals in employment and that the decision to work is unlikely to be random.

However, the previous sets of assumptions, even if satisfied in the total population, do not hold in the working subpopulation if selection into employment is related to factors that also affect the outcome, for instance ability. To improve upon this problem both notationally and methodologically, we introduce a binary selection indicator S which is equal to one if an individual is employed such that the wage outcome Y is observed in the data and zero otherwise. We maintain that G, X, W are observed for all individuals and note that each of these variables might affect S which can be considered as yet another outcome variable.

Using the results of Huber and Solovyeva (2018), one may combine Assumption 2 with specific restrictions on the nature of selection into employment. The first approach of Huber and Solovyeva (2018) assumes selection to be related to the observed variables G, X, W only.

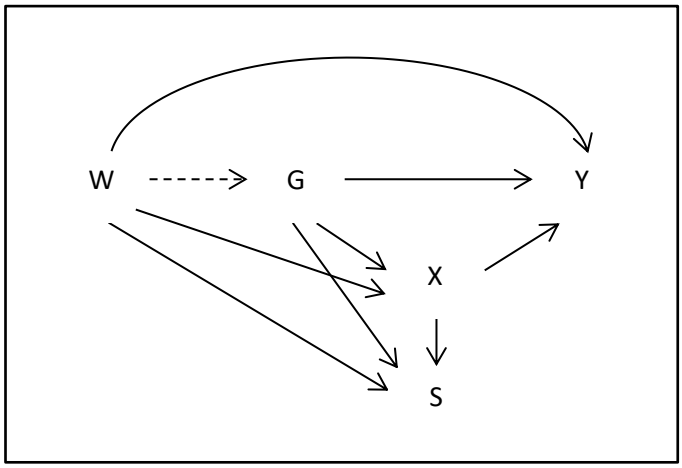
Assumption 3 (Selection on observables):

(a) $Y \perp S | G = g, X = x, W = w$ for all $g \in \{0, 1\}$ and x, w in the support of X, W,

(b) $\Pr(S = 1 | G = g, X = x, W = w) > 0$ for all $g \in \{0, 1\}$ and x, w in the support of X, W.

By Assumption 3(a), there are no unobservables confounding S and Y conditional on G, X, W, so that outcomes are missing at random (MAR) in the denomination of Rubin (1976). The common support restriction implies that conditional on the values of G, X, W in their joint support, the probability to be observed is larger than zero, otherwise no outcome is observed for some specific combinations of these variables and identification fails. Figure 3 presents a graphical illustration of the decomposition with selection on observables.

Figure 3: A graphical representation of the decomposition under Assumption 3

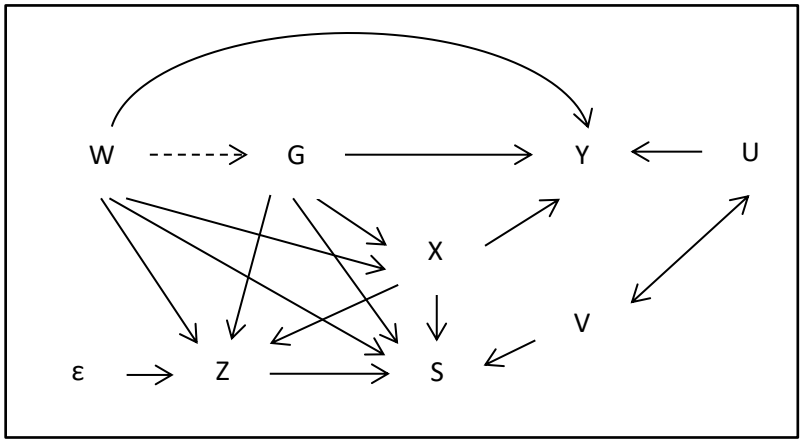


Under Assumptions 2 and 3, the parameters of interest are identified by the following IPW expression, which fits the general framework of IPW-based M-estimation of missing data models in Wooldridge (2002):

$$\begin{aligned}
\psi &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|W) \cdot \Pr(S=1|G,X,W)}\right] \\
&- E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W) \cdot \Pr(S=1|G,X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right], \qquad (8) \\
\eta &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W) \cdot \Pr(S=1|G,X,W)} \cdot \frac{1-\Pr(G=1|X,W)}{1-\Pr(G=1|W)}\right] \\
&- E\left[\frac{Y \cdot (1-G) \cdot S}{(1-\Pr(G=1|W)) \cdot \Pr(S=1|G,X,W)}\right]. \qquad (9)
\end{aligned}$$

Alternatively to Assumption 3, Huber and Solovyeva (2018) present a control function approach for the case that selection is related to unobservables affecting the outcome. This requires an instrument for selection, denoted by Z, which affects selection but is not directly associated with the outcome. Figure 4 provides a graphical representation of mediation with selection on unobservables and an instrument for selection. $\mathcal{E}$, V, and U denote unobserved variables that affect the instrument for selection Z, the selection indicator S, and the outcome Y, respectively.

Figure 4: A graphical representation of the decomposition under Assumption 4



Assumption 4 (Instrument for selection):

a) There exists an instrument Z that may be a function of G, X, i.e. $Z = Z(G,X)$, is conditionally correlated with S, i.e. $E[Z \cdot S|G,X,W] \neq 0$, and satisfies (i) $Y(g,x,z) = Y(g,x)$ and (ii) $\{Y(g,x), X(g')\} \perp Z(g'',x')|W=w$ for all $g,g',g'' \in \{0,1\}$ and z,x,x',w in the support of Z,X,W,

(b) $S = I\{V \leq \Pi(G,X,W,Z)\}$, where Π is a general function and V is a scalar (index of) unobservable(s)

with a strictly monotonic cumulative distribution function conditional on W,

(c) $V \perp (G, X, Z)|W$,

(d) $E[Y(1,x) - Y(0,x)|W = w, V = v, S = 1] = E[Y(1,x) - Y(0,x)|W = w, V = v]$ and $E[Y(g, X(1)) - Y(g, X(0))|W = w, V = v, S = 1] = E[Y(g, M(1)) - Y(g, M(0))|W = w, V = v]$, for all $g \in \{0,1\}$ and x, w, v in the support of X, W, V,

(e) $\Pr(G = 1|X = x, W = w, p(Q) = p(q)) > 0$, $0 < \Pr(G = 1|W = w, p(Q) = p(q)) < 1$, and $p(q) > 0$ for all $g \in \{0,1\}$ and x, w, z in the support of X, W, Z.

In contrast to Assumption 3(a), the unobservable V in the selection equation is now allowed to be associated with unobservables U affecting the outcome. Therefore, the distribution of V generally differs across values of G, X conditional on W, which entails confounding. Identification hinges on exogenous shifts in the conditional selection probability $p(Q) = \Pr(S = 1|G, M, X, Z)$ based on instrument Z, with $Q = (G, X, W, Z)$ for the sake of brevity. By using $p(Q)$ as additional control variable in the decompositions, one controls for the distribution of V and thus, for the confounding associations of V with (i) D and $\{Y(d,m), M(d')\}$ and (ii) M and $Y(d,m)$ that occur conditional on $S = 1$.

Z and S have to satisfy particular conditions. Z must not affect Y or be associated with unobservables affecting X or Y conditional on W, as invoked in Assumption 4(a). By the threshold crossing model in Assumption 4(b), $p(Q)$ identifies the distribution function of V given W. Assumption 4(c) implies the (nonparametric) identification of the distribution of V, as the latter is independent of (G, X, Z) given W. Assumption 4(d) imposes homogeneity of the observed and unobserved causal mechanisms across employed and non-employed populations conditional on W, V. Without this restriction, wage decompositions can merely be conducted for the employed but not the total population, as effects might be heterogeneous in unobservables, see also the discussion in Newey (2007). A sufficient condition for effect homogeneity in unobservables is separability of observed and unobserved components in the outcome variable, i.e. $Y = \eta(G, M, X) + \nu(U)$, where η, ν are general functions and U is a scalar or vector of unobservables. Finally, the first part of Assumption 4(e) strengthens the previous common support assumption 2(c) to also hold when including $p(Q)$ as additional control variable. The second part requires the selection probability $p(Q)$ to be larger than zero for any combination of values in the support of G, X, W, Z to ensure that outcomes are observed for all values occurring in the population. Under Assumptions 2 and 4, the causal

mechanisms are identified by the following expressions:

$$\begin{aligned}
\psi &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|W,p(Q)) \cdot p(Q)}\right] \\
&- E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W,p(Q)) \cdot p(Q)} \cdot \frac{1-\Pr(G=1|X,W,p(Q))}{1-\Pr(G=1|W,p(Q))}\right], \qquad (10) \\
\eta &= E\left[\frac{Y \cdot G \cdot S}{\Pr(G=1|X,W,p(Q)) \cdot p(Q)} \cdot \frac{1-\Pr(G=1|X,W,p(Q))}{1-\Pr(G=1|W,p(Q))}\right] \\
&- E\left[\frac{Y \cdot (1-G) \cdot S}{(1-\Pr(G=1|W,p(Q))) \cdot p(Q)}\right]. \qquad (11)
\end{aligned}$$

3 Data

Our data come from the National Longitudinal Survey of Youth 1979 (NLSY79), a panel survey of young individuals who were aged 14 to 22 years at the first wave in 1979.[6] Conducted annually until 1994, it then became biannual. The data contain a wealth of individual characteristics, including rich information relevant for labor market decisions, such as education, occupation, work experience and more. We estimate decompositions for wages reported in the year 2000 when respondents were 35 – 43 years old. After excluding 1,351 observations from the total NLSY79 sample in 2000 due to various data issues,[7] our evaluation sample consists of 6,658 individuals (3,162 men and 3,496 women). Table 3 in Appendix A provides descriptive statistics (mean values, mean differences, and respective p-values based on two-sample t-tests) for the key variables in our analysis. The group variable G is equal to zero for female and one for male respondents, such that male wages are regarded as reference wages, as it is frequently the case in the decomposition literature.[8] The outcome variable of interest (Y) is the log average hourly wage in the past calendar year reported in 2000. The selection indicator S is equal to one for individuals who indicated to have worked at least 1,000 hours in the past calendar year. This is the case for 87% of males and 70% of females.

[6]The NLSY79 data consist of three independent probability samples: a cross-sectional sample (6,111 subjects, or 48%) representing the non-institutionalized civilian youth; a supplemental sample (42%) oversampling civilian Hispanic, black, and economically disadvantaged nonblack/non-Hispanic young people; and a military sample (10%) comprised of youth serving in the military as of September 30, 1978 (Bureau of Labor Statistics, U.S. Department of Labor (2001)).

[7]Specifically, we excluded 502 persons who reported to have worked 1,000 hours or more in the past calendar year, but whose average hourly wages in the past calendar year were either missing or equal to zero. We also dropped 54 working individuals with average hourly wages of less than $1 in the past calendar year. Furthermore, 608 observations with missing values in mediators (see Table 3 for the full list of mediators) and 186 observations with missing values in the instruments for selection – the number of young children and the employment status of the respondent's mother back when the respondent was 14 years old – were excluded.

[8]We refer to Sloczynski (2013) for a discussion of reference group choice in the potential outcome framework.

The set of post-group characteristics X, which potentially mediate the effect of gender on wages, consists of individual variables reported in or constructed with reference to 1998: marital status, years in marriage, the region of residence and how many years an individual has been residing in that region, an indicator for living in an urban area (SMSA) and the number of years living in an urban area, education level, indicators for the year when first worked, number of jobs ever had, tenure with the current employer (in weeks), industry and the number of years working there, occupation and the number of years working in that occupation, whether employed in 1998 and total years of employment. Further characteristics are the form of employment (whether full-time), the share of full-time employment in employment years in 1994–98, total weeks of employment, the number of weeks unemployed and the number of weeks out of the labor force, and whether health problems prevented work. Moreover, several higher-order (squared and cubed) and interaction terms are included to make the propensity score specification more flexible. p-values of the two-sample t-tests in Table 3 in Appendix A reveal that women in our sample differ significantly (at the 5% level) from men in a range of variables. For instance, males have on average more labor market experience, while females have a higher average level of education. Important differences also arise in other factors related to labor market performance (e.g., industry, occupation, employment form, etc.).

Although X includes and even surpasses the set of variables conventionally used in wage decompositions, further potentially important characteristics mediating the effect of gender on wage are not considered. For instance, risk preferences, attitudes towards competition and negotiations, and other socio-psychological factors (see e.g., Bertrand (2011) and Azmat and Petrongolo (2014)), are not available in our data. Their effects thus contribute to the unexplained component.

Potential confounders W related to factors determined at or prior to birth include race, religion, year of birth, birth order, parental place of birth (in the U.S. or abroad), and parental education. We acknowledge that further confounders not available in our data but correlated with G, X, and/or Y likely exist. For instance, see Cobb-Clark (2016) for a review of biological factors, such as sensory functioning (e.g., time-space perceptions), emotions, and levels of sex hormones, potentially linking gender with labor market behavior and outcomes. In particular, some studies relate higher levels of prenatal testosterone to stronger preference for risk (Garbarino, Slonim, and Sydnor (2011)) and sorting into traditionally male-dominated occupations (Manning, Reimers, Baron-Cohen, Wheelwright, and Fink (2010) and Nye and Orel (2015)). Therefore, we do not claim that controlling for W fully tackles endogeneity bias. Nevertheless, we are interested in the sensitivity of decompositions w.r.t. to the inclusion and exclusion of W, even if these variables only comprise a subset of the actual confounders.

Finally, we define the number of children in 1999 younger than 6 and 15 years old, respectively, as instruments Z for selection into our employment indicator S. Such instruments based on the number of children in a household have been widely used as instruments for labor supply in the empirical labor market literature, see for instance Mulligan and Rubinstein (2008). We, however, note that the validity of this approach is not undisputed, as the number of children might be correlated with unobservables also affecting the wage outcome, like relative preference for family and working life. For this reason, Huber and Mellace (2014) provided a method to partially test instrument validity, namely a joint test for the exclusion restriction and additive separability of the unobservable V in the selection equation. They applied them to children-based instruments for female labor supply in four data sets, but found no statistical evidence for the violation of the IV assumptions. As a word of caution, however, their tests cannot detect all possible violations of instrument validity even asymptotically, as they rely on a partial identification approach. Even though concerns about the instruments may therefore remain, it is our aim to verify how sensitive decompositions are across different methods, also w.r.t. modelling selection based on instruments commonly used in the literature. In a robustness check, we consider an indicator for the respondent's mother working for pay back when the respondent was 14 years old as an additional instrument for selection. This, however, yields very similar point estimates based on (10) and (11) as when using the children-based instruments alone, see the discussion below.

4 Empirical results

We decompose the gender wage gap based on the five approaches outlined in Section 2. Table 1 provides the estimated effects (est.) along with standards errors (s.e.) and p-values (p-val) using 999 bootstrap replications. It also shows the shares (% tot.) of the explained and unexplained components in the total gender wage gap. The last two columns (Trimmed obs., %) indicate, respectively, the number and the share of units dropped in the IPW estimations due to a trimming rule that discards observations with extreme propensity scores larger than 0.99 and/or smaller than 0.01. This is done to prevent the assignment of very large weights to specific observations (due to small denominators in IPW) as a consequence of insufficient common support across gender or selection into employment.

Our main specification includes the full list of post-group characteristics (X) presented in Table 3 in Appendix A, as well as several higher-order and interaction terms. The standard Oaxaca-Blinder decomposition (Oaxaca-Bl.) based on (2) and (3) as well as IPW (IPW no W) based on (4) and (5) invoke

Assumption 1 and thus neither control for the potential endogeneity of X nor for selection. Therefore, estimations are conducted in the subsample with $S = 1$. Under Assumption 2, IPW is based on (6) and (7) and includes potential confounders W listed in Table 3 in Appendix A (IPW with W) to tackle endogeneity. Under Assumption 3, IPW based on (8) and (9) uses these covariates to control for both endogeneity and selection (IPW MAR). Finally, under Assumption 4, IPW based on (10) and (11) in addition utilizes a combination of the number of children younger than 6 and 15 years old as instruments (Z) for selection into employment (IPW IV).

Table 1: Gender wage gap decomposition based on NLSY79: main specification

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.083	0.021	0.000	27.9%	0.215	0.024	0.000	72.1%	0	0.0%
IPW no W	0.293	0.019	0.000	0.118	0.030	0.000	40.1%	0.176	0.031	0.000	59.9%	28	0.5%
IPW with W	0.264	0.017	0.000	0.096	0.028	0.001	36.5%	0.168	0.030	0.000	63.5%	28	0.5%
IPW MAR	0.365	0.035	0.000	0.219	0.033	0.000	59.8%	0.147	0.035	0.000	40.2%	90	1.4%
IPW IV	0.141	0.324	0.665	-0.005	0.102	0.964	-3.3%	0.145	0.328	0.658	103.3%	584	8.8%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

When applying the classic Oaxaca-Blinder decomposition, 28% (0.083) of the total gender wage gap of 0.299 is attributed to differences in the included post-group characteristics X, while about 72% (0.215) remains unexplained. All the Oaxaca-Blinder estimates are highly statistically significant.[9] In contrast to the Oaxaca-Blinder decomposition, IPW without W does not impose linearity of Y in X given G but instead requires an estimate of the propensity score $\Pr(G = 1|X)$, which is obtained by logit regression. Figures 5 to 13 and Tables 4 and 5 in Appendix A present, respectively, histograms and summary statistics (minimum, mean, and maximum) of the within-group propensity scores used in our IPW-based estimations.[10] Figure 5 suggests a decent overlap in the distribution of estimates of $\Pr(G = 1|X)$, implying common support in observed characteristics across females and males over most of the support of X. Applying a trimming rule that excludes observations with propensity scores < 0.01, we drop 28 units, or 0.5 %, from the sample. Compared to the Oaxaca-Blinder decomposition, the the explained component is slightly larger and the unexplained component somewhat smaller, while total wage gap remains almost unchanged. For IPW including potential confounders W, Figures 6 and 7 in Appendix A display the histograms of the logit-based estimates of $\Pr(G = 1|W)$ and $\Pr(G = 1|X, W)$ and point to decent common

[9]The regression-based Oaxaca-Blinder estimator does not rely on common support, see the discussion in Section 2, and therefore does not require trimming observations with extreme propensity score values.

[10]Table 7 in Appendix A additionally provides the number and the share of trimmed observations for each propensity score.

support w.r.t. either propensity score. Therefore, (only) the same 28 observations as for IPW are without controls dropped from the sample. Controlling for W leads to moderately smaller estimates of the total wage gap as well as the explained and unexplained components when compared to IPW without controls.

IPW MAR relies on estimating the selection propensity score $\Pr(S = 1|G, X, W)$ to control for the employment decision based on observables, again by logit regression. Figure 10 in Appendix A presents histograms of estimated selection probabilities for individuals who worked less than 1,000 hours in the past calendar year ($S = 0$) and those who worked 1,000 hours or more ($S = 1$). We note that the selection probability is close to zero for a subset of individuals but clearly larger than zero for most of the sample. 90 (1.4%) observations are dropped from estimation, once the additional condition that selection propensity scores must not be smaller than 0.01 is added to the previous trimming rule. The total wage gap (0.365 log points) and the explained component (0.219 log points) are considerably larger than under IPW controlling for W (but ignoring selection). In contrast, the magnitude of the unexplained component (0.147 log points) is slightly smaller, resulting in an overall drop of its share in the total wage gap to 40%. Any estimates discussed so far are statistically significant at the 1% level.

In addition to controlling for observables, our last estimator, IPW IV, uses the number of children under 15 and under 6 years as instruments to control for selection. It requires the estimation of $p(Q) = \Pr(S = 1|Q)$ (with $Q = (G, X, W, Z)$), $\Pr(G = 1|W, p(Q))$, and $\Pr(G = 1|X, W, p(Q))$. Figures 11, 12, and 13 provide the logit estimates of the respective propensity scores. Common support is by and large satisfactory. The trimming rule discards observations with estimates of $\Pr(G = 1|X = x, W = w, p(Q) = p(q)) < 0.01$, of $\Pr(G = 1|W = w, p(Q) = p(q)) > 0.99$, and of$p(q) < 0.01$, all in all 584 cases (8.8%). This needs to be kept in mind when interpreting the results, as trimming generally changes the target population for which the parameters are estimated. The total wage gap drops substantially when compared to previous estimates and amounts to 0.141 log points. The unexplained component is similar in magnitude to the IPW MAR estimate, while the explained part is very close to zero but even negative. However, any of the IPW IV estimates is far from being statistically significant at any conventional level, pointing to a weak instrument problem.

We conduct several sensitivity checks by gradually reducing the set of post-group characteristics X. Table 8 in Appendix A presents the estimates obtained when dropping any higher-order and interaction terms of X, such that the functional forms in the outcome and propensity score specifications become less flexible. While the total wage gap estimates remain largely unchanged, the explained components generally

decline slightly (by about 0.03 log points), and the unexplained components increase, on average, by the same amount. The exception is the IPW IV decomposition, where both the total gap and its explained component somewhat increase, whereas the size and the share of the unexplained component decline. However, all the IPW IV estimates remain statistically insignificant. All in all, these differences are minor, which suggests that our results are rather robust to the exclusion of higher-order and interaction terms of X.

Our next robustness check excludes not only the higher-order and interaction terms, but also all variables in X that reflect developments or histories like years in marriage, years worked in current occupation, etc. We point out that many of these variables are frequently not included in wage decompositions, even though they appear a priori similarly important as characteristics measured at a particular point in time. For instance, one would suspect that not only the current occupation matters for human capital accumulation and the determination of the current wage, but also employment history and tenure in the current occupation. The exclusion of these additional variables generally decreases the explained component and increases the unexplained component, which accounts for 77% to 96% of the total gap across the first four methods. IPW IV yields different and even more extreme estimates, which are, however, at best marginally significant. Table 2 provides the results.

Table 2: Robustness check: parsimonious set of X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.067	0.019	0.000	22.5%	0.231	0.022	0.000	77.5%	0	0.0%
IPW no W	0.298	0.019	0.000	0.026	0.023	0.269	8.6%	0.272	0.026	0.000	91.4%	1	0.0%
IPW with W	0.269	0.017	0.000	0.011	0.023	0.648	3.9%	0.258	0.027	0.000	96.1%	2	0.0%
IPW MAR	0.362	0.032	0.000	0.076	0.025	0.002	20.9%	0.287	0.032	0.000	79.1%	1	0.0%
IPW IV	0.124	0.324	0.703	-0.186	0.102	0.067	-150.6%	0.310	0.328	0.345	250.6%	850	12.8%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

The Oaxaca-Blinder decomposition yields quite stable estimates when compared to the main specification of Table 1. The total gap estimate does not change, while the explained component decreases and the unexplained component increases each by about 0.02 log points, or about 5 percentage points of the total gap. For the IPW estimators not accounting for selection, the explained components decline by about 0.1 log point, now constituting only a small share of the total gap and losing their statistical significance. Over 90% of the total wage gap remains unexplained both for IPW with and without controlling for W. Also for the IPW estimators accounting for selection, the explained components decrease considerably, while the

explained components increase and the total gap is slightly smaller than before. In the case of IPW MAR, the unexplained part now accounts for nearly 80% of the total wage gap. All the IPW MAR estimates are statistically significant at the 1% level. The IPW IV estimator yields rather implausible results. The large unexplained component of 0.31 log points comprises 251% of the total wage gap, due to a negative estimate of the explained component. However, none of these estimates are statistically significant at the 5%.

As a final robustness check for IPW IV, we add an indicator for whether an individual's mother worked for pay when the individual was 14 years old as an additional instrument for selection into paid work. Table 9 in Appendix A shows that the estimates remain unchanged compared to the main specification. Overall, our empirical results suggest that estimates of the gender wage decomposition are dependent on the choice of underlying identification assumptions and, to some extent, the definition of the observed characteristics X. Given the variability of estimates across methods and specifications, we advise to be cautious w.r.t. the use of wage decompositions for policy conclusions, for instance about the magnitude of gender discrimination in the labor market. [11]

5 Conclusion

We assessed the sensitivity of average gender wage gap decompositions in data from the U.S. National Longitudinal Survey of Youth 1979, comparing several decomposition methods and sets of included variables. We first discussed the identification problem from a causal perspective, namely separating the explained component of the wage effect of gender operating through observed characteristics from the unexplained component. Five decomposition techniques were reviewed. Starting with the linear Oaxaca-Blinder decomposition, we gradually relaxed the identifying assumptions regarding functional form, exogeneity of observed characteristics and gender, and selection into employment. Specifically, we considered inverse probability weighting (IPW) as a semiparametric analog of the standard Oaxaca-Blinder decomposition. We also included IPW versions controlling for confounders (of observed characteristics, gender, and the wage outcome) or for both confounders and sample selection into employment, the latter either based on observed variables or instruments. When applying all five estimators to the data, we also considered less and more parsimonious definitions of the observed characteristics and instruments included in the analysis.

[11] The differences between the estimates of the total wage gap are statistically significant at the 5% level across the compared estimators, except the comparison between the Oaxaca-Blinder and IPW without potential confounders W. These results are available upon request.

We found the total wage gap as well as the explained and unexplained components to differ importantly across some of the methods considered. Furthermore, the definition of the observed characteristics related to the explained component mattered: Including only levels of variables rather than both levels and histories generally reduced the explained and increased the unexplained components across the considered estimators. Given our results, the usefulness of wage decompositions that neither account for identification issues like endogeneity and selection into employment nor for histories of observed characteristics appears questionable in terms of policy conclusions, for instance, when aiming at quantifying gender discrimination. Unfortunately, a vast number of empirical applications rely on exactly such kind of decompositions. At the very least, we advise checking the robustness of the results across several decomposition methods and variable specifications to improve upon the status quo of the literature.

References

Angrist, J. D., and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Epiricist's Companion*. Princeton University Press.

Arellano, M., and S. Bonhomme (2010): "Quantile selection models," *unpublished manuscript*.

Azmat, G., and B. Petrongolo (2014): "Gender and the labor market: What have we learned from field and lab experiments?," *Labour Economics*, 30, 32 – 40.

Barsky, R., J. Bound, K. Charles, and J. Lupton (2002): "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," *Journal of the American Statistical Association*, 97, 663–673.

Bertrand, M. (2011): "New Perspectives on Gender," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1543–1590. Elsevier.

Bičakova, A. (2014): "Selection into Labor Force and Gender Unemployment Gaps," *CERGE-EI Working Paper*, 513.

Blau, F., and L. Kahn (2006): "The US gender pay gap in the 1990s: Slowing convergence," *Industrial and Labor Relations Review*, 60, 45–66.

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

Bureau of Labor Statistics, U.S. Department of Labor (2001): "National Longitudinal Survey of Youth 1979 cohort, 1979-2000 (rounds 1-19)," Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2009): "Inference on Counterfactual Distributions," *CeMMAP working paper CWP09/09*.

Cobb-Clark, D. A. (2016): "Biology and Gender in the Labor Market," *IZA DP No. 10386*.

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

Duncan, O. D. (1967): "Discrimination against Negroes," *Annals of the American Academy of Political and Social Science*, 371, 85–103.

Firpo, S., N. M. Fortin, and T. Lemieux (2007): "Decomposing Wage Distributions using Recentered Influence Functions Regressions," *mimeo, University of British Columbia*.

——— (2009): "Unconditional quantile regressions," *Econometrica*, 77, 953–973.

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

Frölich, M. (2007): "Propensity score matching without conditional independence assumptionwith an application to the gender wage gap in the United Kingdom," *Econometrics Journal*, 10, 359–407.

Garbarino, E., R. Slonim, and J. Sydnor (2011): "Digit ratios (2D:4D) as predictors of risky decision making for both sexes," *Journal of Risk and Uncertainty*, 42(1), 1–26.

García, J., P. J. Hernández, and A. López-Nicolás (2001): "How wide is the gap? An investigation of gender wage differences using quantile regression," *Empirical Economics*, 26, 149–167.

Goraus, K., J. Tyrowicz, and L. van der Velde (2015): "Which Gender Wage Gap Estimates to Trust? A Comparative Analysis," *Review of Income and Wealth*, 63, 118–146.

Greiner, D. J., and D. B. Rubin (2011): "Causal Effects of Perceived Immutable Characteristics," *The Review of Economics and Statistics*, 93, 775–785.

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

Heckman, J. J. (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, 475–492.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Huber, M. (2014): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

——— (2015): "Causal pitfalls in the decomposition of wage gaps," *Journal of Business and Economic Statistics*, 33, 179–191.

Huber, M., and G. Mellace (2014): "Testing exclusion restrictions and additive separability in sample selection models," *Empirical Economics*, 47, 75–92.

Huber, M., and A. Solovyeva (2018): "Evaluating direct and indirect effects under sample selection and outcome attrition," *working paper, University of Fribourg*.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 5171.

Juhn, C., K. Murphy, and B. Pierce (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101, 410–442.

Kunze, A. (2008): "Gender wage gap studies: consistency and decomposition," *Empirical Economics*, 35, 63–76.

Lemieux, T. (1998): "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection," *Journal of Labor Economics*, 16, 261–291.

Maasoumi, E., and L. Wang (2016): "The Gender Gap Between Earnings Distributions," *working paper, Emory University*.

Machado, C. (2017): "Unobserved selection heterogeneity and the gender wage gap," *forthcoming in the Journal of Applied Econometrics*.

Machado, J., and J. Mata (2005): "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics*, 20, 445–465.

Manning, J. T., S. Reimers, S. Baron-Cohen, S. Wheelwright, and B. Fink (2010): "Sexually dimorphic traits (digit ratio, body height, systemizingempathizing scores) and gender segregation between occupations: Evidence from the BBC internet study," *Personality and Individual Differences*, 49(5), 511 – 515.

Manski, C. F. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.

Melly, B. (2005): "Decomposition of differences in distribution using quantile regression," *Labour Economics*, 12, 577–590.

Mora, R. (2008): "A nonparametric decomposition of the Mexican American average wage gap," *Journal of Applied Econometrics*, 23, 463–485.

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

Neuman, S., and R. L. Oaxaca (2003): "Gender versus Ethnic Wage Differentials among Professionals: Evidence from Israel," *Annales d'conomie et de Statistique*, (71/72), 267–292.

——— (2004): "Wage Decompositions with Selectivity-Corrected Wage Equations: A Methodological Note," *The Journal of Economic Inequality*, 2, 3–10.

Newey, W. K. (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

Ñopo, H. (2008): "Matching as a Tool to Decompose Wage Gaps," *Review of Economics and Statistics*, 90, 290–299.

Nye, J., and E. Orel (2015): "The influence of prenatal hormones on occupational choice: 2D:4D evidence from Moscow," *Personality and Individual Differences*, 78(Supplement C), 39 – 42.

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

Olivetti, C., and B. Petrongolo (2008): "Unequal pay or unequal employment? A cross-country analysis of gender gaps," *Journal of Labor Economics*, 26, 621–654.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Sloczynski, T. (2013): "Population average gender effects," *IZA Discussion Paper No. 7315*.

Wooldridge, J. (2002): "Inverse Probability Weigthed M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 141–162.

Yamaguchi, K. (2014): "Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An extension of the DiNardo-Fortin-Lemieux method," *RIETI Discussion Paper Series 14-E-061*.

A Appendix

Table 3: Summary statistics and mean differences by gender

Variables	Male($G=1$)	Female($G=0$)	Difference	p-value
Outcome Y (non-logged, refers to selected population with $S=1$)				
Hourly wage	19.370	14.164	5.206	0.000
Mediators X (refer to 1998 unless otherwise is stated)				
Married	0.566	0.568	-0.002	0.882
Years married total since 1979	6.430	7.537	-1.107	0.000
Northeastern region	0.153	0.155	-0.002	0.857
North Central region	0.242	0.237	0.005	0.602
West region	0.206	0.195	0.011	0.244
South region (ref.)	0.399	0.414	-0.015	0.205
Years lived in current region since 1979	14.839	15.246	-0.407	0.000
Resides in SMSA	0.811	0.816	-0.005	0.584
Years lived in SMSA since 1979	13.488	14.201	-0.713	0.000
Less than high school (ref.)	0.129	0.101	0.028	0.000
High school graduate	0.459	0.416	0.043	0.000
Some college	0.208	0.271	-0.063	0.000
College or more	0.204	0.213	-0.009	0.413
First job before 1975	0.065	0.046	0.019	0.001
First job in 1976–79	0.115	0.128	-0.013	0.083
First job after 1979 (ref.)	0.821	0.825	-0.004	0.623
Numer of jobs ever had	10.555	9.239	1.316	0.000
Tenure with current employer (wks.)	276.056	212.662	63.394	0.000
Industry: Primary sector	0.227	0.078	0.149	0.000
Industry: Manufacturing (ref.)	0.140	0.053	0.087	0.000
Industry: Transport	0.115	0.048	0.067	0.000
Industry: Trade	0.134	0.142	-0.008	0.322
Industry: Finance	0.040	0.064	-0.024	0.000
Industry: Services (business, personnel, and entertain.)	0.121	0.124	-0.003	0.768
Industry: Professional services	0.113	0.297	-0.184	0.000
Industry: Public administration	0.054	0.052	0.002	0.751
Years worked in current industry since 1982	3.555	2.622	0.933	0.000
Manager	0.234	0.258	-0.024	0.022
Technical occupation (ref.)	0.039	0.038	0.001	0.907
Occupation in sales	0.067	0.082	-0.015	0.021
Clerical occupation	0.056	0.212	-0.156	0.000
Occupation in service	0.102	0.163	-0.061	0.000
Farmer or laborer	0.276	0.042	0.234	0.000
Operator (machines, transport)	0.170	0.063	0.107	0.000
Years worked in current occupation since 1982	2.180	1.727	0.453	0.000
Employment status: employed	0.877	0.748	0.129	0.000
Number of years employed status since 1979	13.204	11.271	1.933	0.000
Employed full time	0.846	0.599	0.247	0.000
Share of full-time employment 1994-98	0.896	0.658	0.238	0.000

Continued on next page

Table 3 – continued from previous page

Variables	Male($G=1$)	Female($G=0$)	Difference	p-value
Total number of weeks worked since 1979	661.794	560.408	101.386	0.000
Total number of weeks unemployed since 1979	62.343	49.744	12.599	0.000
Total number of weeks out of labor force since 1979	146.118	265.276	-119.158	0.000
Bad health prevents from working	0.045	0.055	-0.010	0.071
Years not working due to bad health since 1979	0.326	0.557	-0.231	0.000
Pre-treatment covariates W				
Hispanic (ref.)	0.193	0.186	0.007	0.488
Black	0.287	0.297	-0.010	0.413
White	0.520	0.517	0.003	0.840
Born in the U.S.	0.935	0.939	-0.004	0.544
No religion	0.045	0.034	0.011	0.031
Protestant	0.501	0.500	0.001	0.957
Catholic (ref.)	0.352	0.352	0.000	0.967
Other religion	0.096	0.112	-0.016	0.036
Mother born in U.S.	0.884	0.896	-0.012	0.102
Mothers educ. <high school (ref.)	0.376	0.421	-0.045	0.000
Mothers educ. high school graduate	0.393	0.369	0.024	0.048
Mothers educ. some college	0.094	0.091	0.003	0.616
Mothers educ. college/more	0.076	0.071	0.005	0.411
Father born in U.S.	0.878	0.884	-0.006	0.410
Fathers educ. <high school (ref.)	0.351	0.366	-0.015	0.201
Fathers educ. high school graduate	0.291	0.297	-0.006	0.560
Fathers educ. some college	0.087	0.076	0.011	0.105
Fathers educ. college/more	0.131	0.117	0.014	0.085
Order of birth	3.195	3.259	-0.064	0.256
Age in 1979	17.501	17.611	-0.110	0.047
Selection indicator S				
Worked 1,000 hrs or more past year	0.867	0.696	0.171	0.000
Instrumental variables Z				
Number of children under 15	1.286	1.209	0.077	0.008
Number of children under 6	0.353	0.295	0.058	0.000
Mother worked at 14	0.543	0.539	0.004	0.718
N of obs.	3,162	3,496	.	.

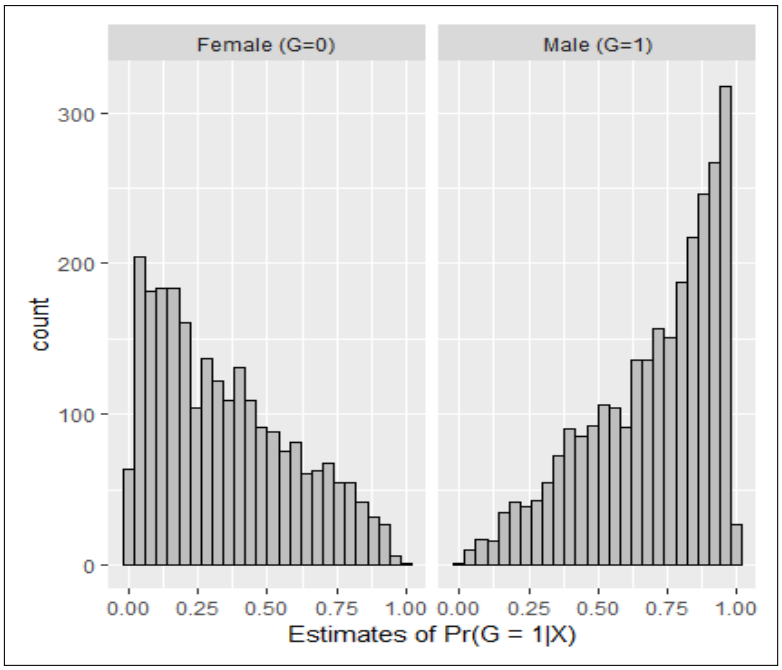


Figure 5: Distribution of the estimated $\Pr(G = 1|X)$ by treatment states in seleted population

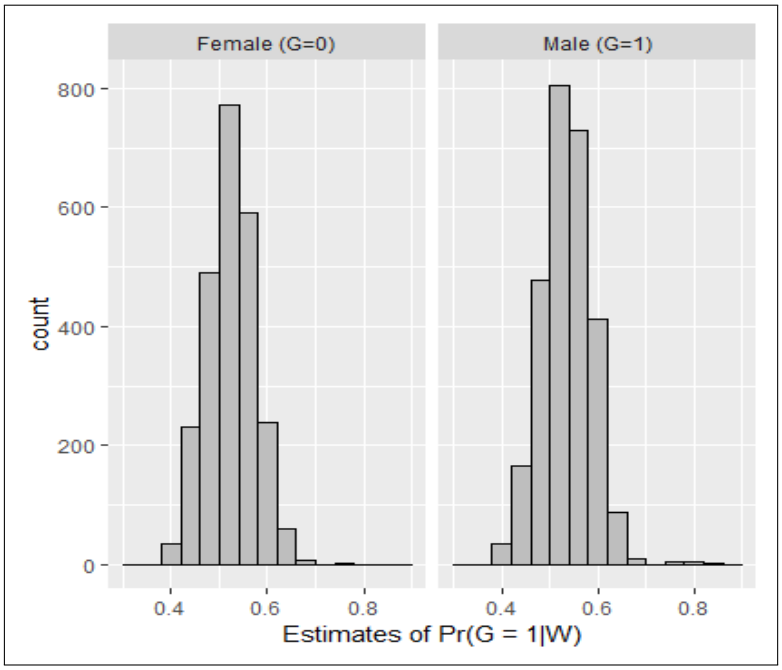


Figure 6: Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in seleted population

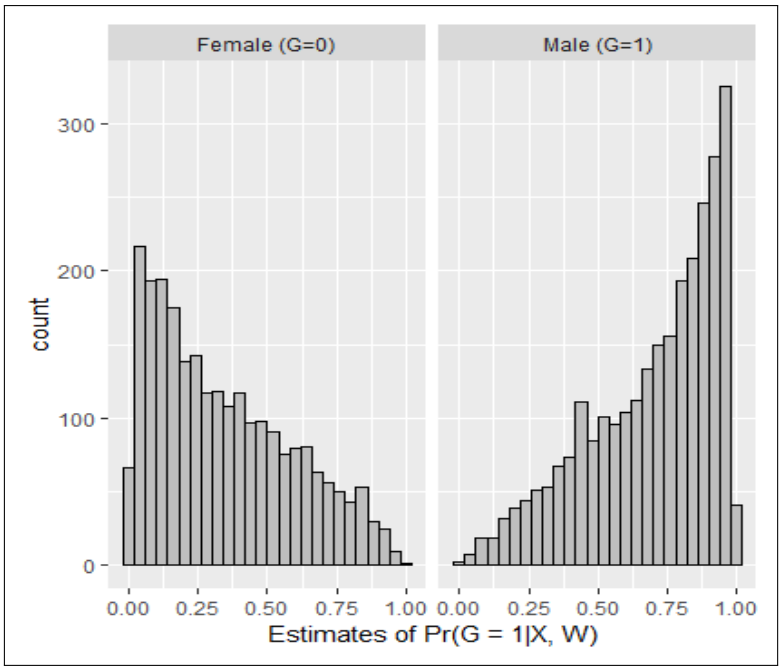


Figure 7: Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in seleted population

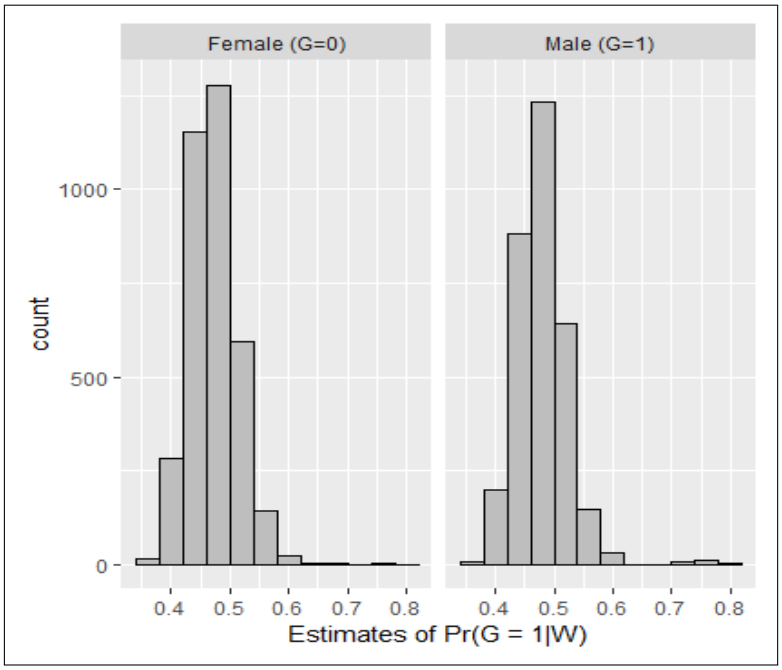


Figure 8: Distribution of the estimated $\Pr(G = 1|W)$ by treatment states in total population

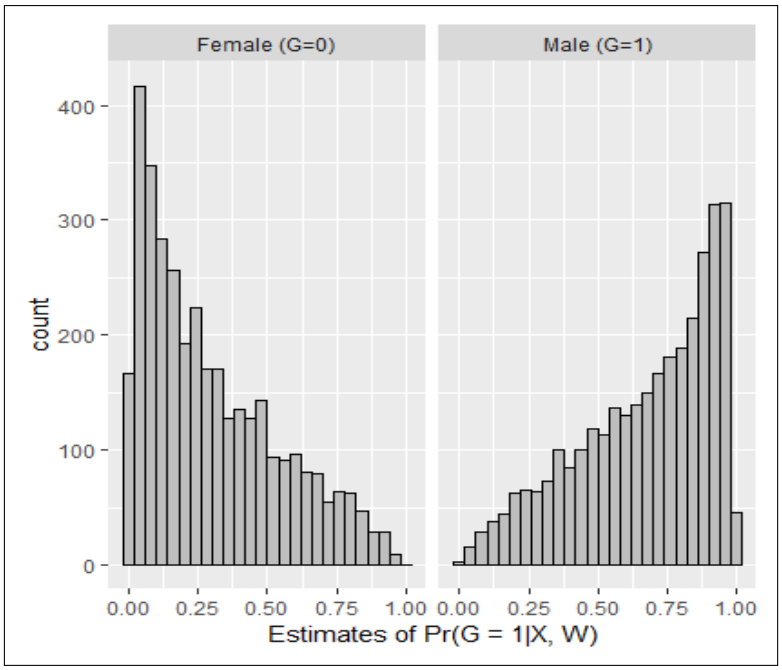


Figure 9: Distribution of the estimated $\Pr(G = 1|X, W)$ by treatment states in total population

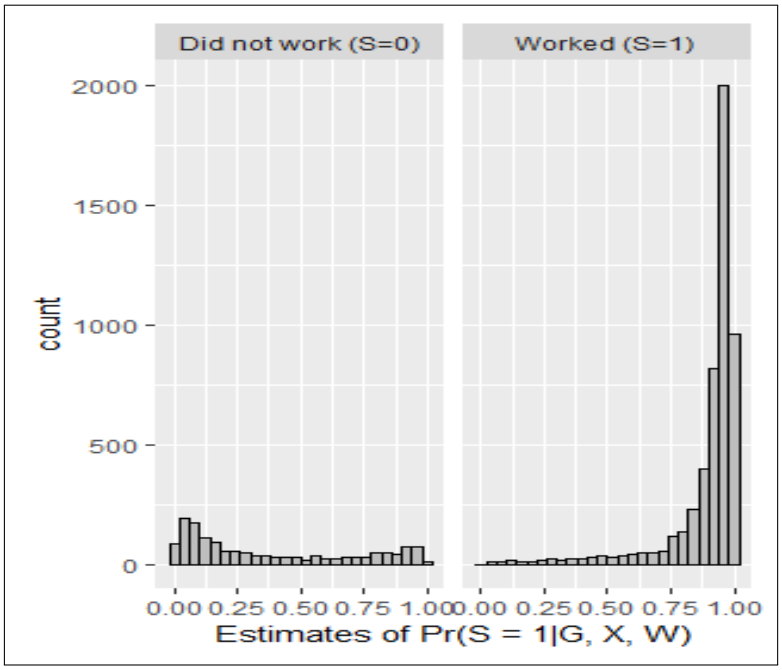


Figure 10: Distribution of the estimated $\Pr(S = 1|G, X, W)$ by selection states

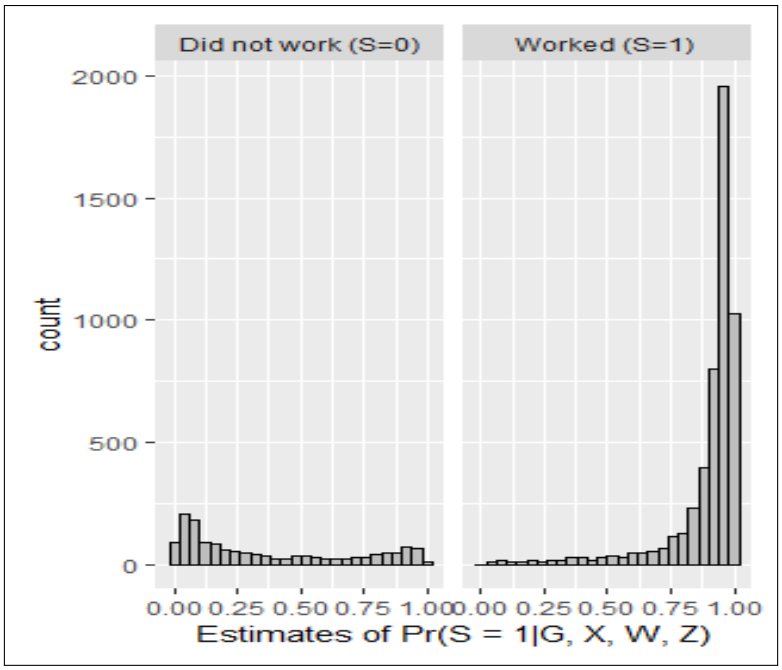


Figure 11: Distribution of the estimated $p(Q) = \Pr(S = 1|G, X, W, Z)$ by selection states

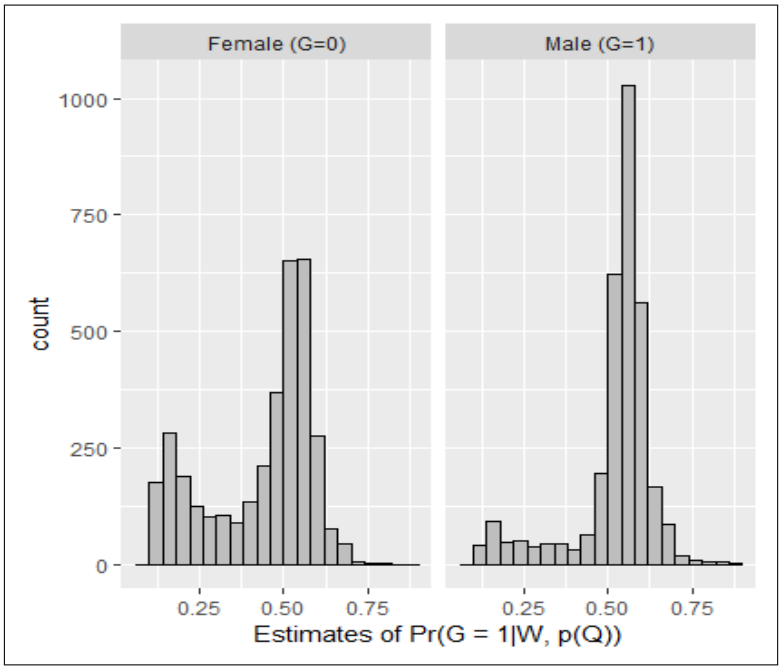


Figure 12: Distribution of the estimated $\Pr(G = 1|W, p(Q))$ by treatment states in total population

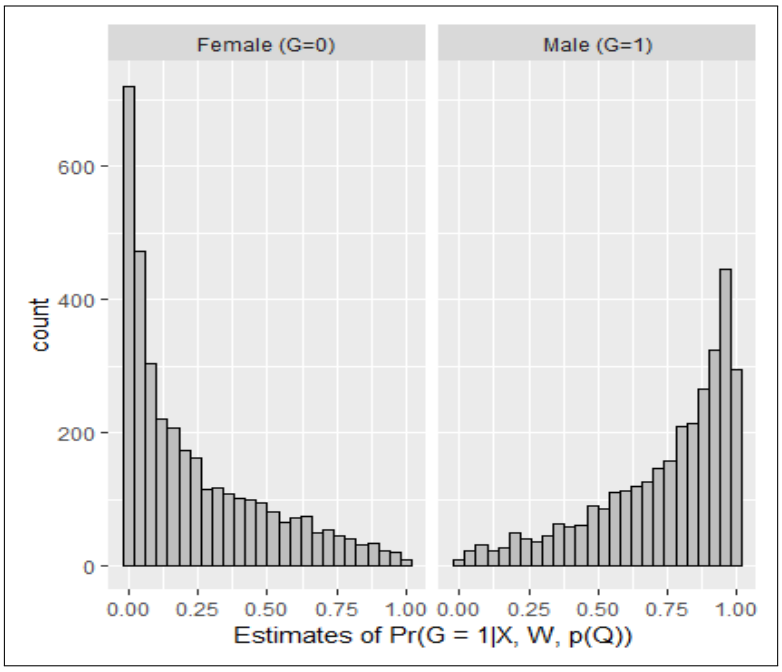


Figure 13: Distribution of the estimated $\Pr(G = 1|X, W, p(Q))$ by treatment states in total population

Table 4: Summary of the estimated treatment propensity scores in selected population

	Min	Mean	Max	Min	Mean	Max
	Female (G=0)			Male (G=1)		
$\Pr(G = 1\|X)$	0.00166	0.34454	0.9819	0.01835	0.6943	0.99047
$\Pr(G = 1\|W)$	0.30751	0.52389	0.8023	0.39349	0.53517	0.87171
$\Pr(G = 1\|X, W)$	0.00133	0.34042	0.9816	0.01574	0.69795	0.99287

Table 5: Summary of the estimated treatment propensity scores in total population

	Min	Mean	Max	Min	Mean	Max
	Female (G=0)			Male (G=1)		
$\Pr(G = 1\|W)$	0.36159	0.47140	0.76295	0.37095	0.47881	0.80260
$\Pr(G = 1\|X, W)$	0.00081	0.29707	0.97202	0.01322	0.67155	0.99619
$\Pr(G = 1\|W, p(Q))$	0.10313	0.43167	0.80670	0.09923	0.52273	0.89403
$\Pr(G = 1\|X, W, p(Q))$	3.22×10^{-7}	0.23804	0.99983	0.00065	0.73682	0.99999

Table 6: Summary of the estimated selection propensity scores in total population

	Min	Mean	Max	Min	Mean	Max
	Did not work (S=0)			Worked (S=1)		
$\Pr(S = 1\|G, X, W)$	0.00327	0.36952	0.99272	0.02076	0.89392	0.99911
$\Pr(S = 1\|G, X, W, Z)$	0.00315	0.36669	0.99386	0.02150	0.89473	0.99909

Table 7: Number of trimmed observations for each propensity score

Trimming condition	obs.	% tot.
Treatment propensity scores in selected population		
$\Pr(G=1\|X)<0.01$	28	0.5
$\Pr(G=1\|W)<0.01$	0	0.0
$\Pr(G=1\|W)>0.99$	0	0.0
$\Pr(G=1\|X,W)<0.01$	28	0.5
Treatment and selection propensity scores in total population		
$\Pr(G=1\|W)<0.01$	0	0.0
$\Pr(G=1\|W)>0.99$	0	0.0
$\Pr(G=1\|X,W)<0.01$	61	0.9
$\Pr(S=1\|G,X,W)<0.01$	29	0.4
$\Pr(S=1\|G,X,W,Z)<0.01$	30	0.4
$\Pr(G=1\|W,p(Q))<0.01$	0	0.0
$\Pr(G=1\|W,p(Q))>0.99$	0	0.0
$\Pr(G=1\|X,W,p(Q))<0.01$	554	8.3

Table 8: Robustness check: no interactions in X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
Oaxaca-Bl.	0.299	0.019	0.000	0.084	0.020	0.000	28.1%	0.215	0.023	0.000	71.9%	0	0.0%
IPW no W	0.295	0.019	0.000	0.093	0.029	0.001	31.7%	0.201	0.030	0.000	68.3%	21	0.4%
IPW with W	0.265	0.017	0.000	0.074	0.028	0.009	27.7%	0.192	0.030	0.000	72.3%	22	0.4%
IPW MAR	0.375	0.034	0.000	0.175	0.033	0.000	46.5%	0.201	0.033	0.000	53.5%	44	0.7%
IPW IV	0.148	0.324	0.649	0.031	0.102	0.758	21.2%	0.116	0.328	0.723	78.8%	673	10.1%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with propensity scores (specific to each estimator) below 0.01 or above 0.99.

Table 9: Mother worked at 14 as an additional IV, full set of X

	Total gap in log wages			Explained (Indirect)				Unexplained (Direct)				Trimmed	
	est.	s.e.	p-val	est.	s.e.	p-val	% tot.	est.	s.e.	p-val	% tot.	obs.	%
IPW IV	0.140	0.156	0.369	-0.005	0.080	0.948	-4%	0.145	0.175	0.408	104%	583	9%

Notes: Standard errors and p-values are estimated based on 999 bootstrap replications. The trimming rule discards observations with $\Pr(G=1|X=x,W=w,p(Q)=p(q))<0.01$, $\Pr(G=1|W=w,p(Q)=p(q))>0.99$, and $p(q)<0.01$.